A componential model for understanding the effects of composite interface metaphors on web-based task performance and learning

Karel Hurts

Department of Psychology, Leiden University
Currently with CogniTech, Inc.
Resediaatraat 8
2313 DG Leiden, The Netherlands
Tel. (31)715226996
Email hurts@euronet.nl

September 2011

Running head: composite interface metaphors
Two experiments describe how a composite metaphor interface (CMI) for performing website tasks can be developed using an iterative and theory-based design approach. The CMI consisted of two semantically unrelated metaphors: the library metaphor and the travel metaphor. A new componential model for predicting the effectiveness of CMIs is described and used for deriving and testing specific hypotheses. Results show that an improved version of the CMI resulted in better declarative knowledge about metaphor-interface mappings than a non-metaphorical control interface (even before the interface was first used), and in less mental effort, as measured by clicking behavior. It also resulted in faster performance of web-based tasks, but only if differences in pre-existing declarative knowledge were taken into account. The discussion addresses i.a. the question how mismatches between interface functions and the metaphor they belong to (confusions) can be measured and used for improving the quality of the CMI.

**Keywords**: metaphor-based learning; human-computer interaction; cognitive aging.
1. Introduction

1.1 Background

For years, metaphors have been used to facilitate user tasks on the World Wide Web. The most popular ones are implicit, or embedded, analogies, and are used as a way to control the complexity of the user interface of a web browser or web site. This is done by designing the actions, underlying concepts, and appearance of the (new) web site or browser on the basis of actions, concepts, and appearances the user is already familiar with: “… metaphors can facilitate active learning by providing clues for abductive and adductive inferences through which learners construct procedural knowledge of the computer” (Carroll, Mack, & Kellogg, 1988).

This type of metaphor is also called interface metaphor. Examples of interface metaphors for the World Wide Web are the book metaphor and the tab (folder) metaphor. Interface metaphors are, in turn, an example of what are sometimes also called “cognitive” metaphors: mechanisms of the human mind allowing us to use what we know about our physical and social experience to provide understanding of other and less familiar subjects (Lakoff & Johnson, 1980).

Despite their popularity, systematic knowledge about how to design good interface metaphors is scarce. Some authors even doubt whether metaphorical interfaces should be designed at all (Halasz & Moran, 1982; Nelson, 1990). Looking at metaphorical interfaces more neutrally, one could say there is disagreement about various issues, such as what the relative strength is of single and composite interface metaphors, what the long-term effects are of using interface metaphors, and whether or not the two single metaphors making up a composite metaphor should be semantically related (see also Hsu, 2005).

In this article, two laboratory experiments are described that show how a composite interface metaphor can be developed on the basis of two component metaphors and
incorporated into a simulated web browser, using a data-based and theory-based design approach. The study also investigated some conditions related to usage context influencing the effectiveness of such an interface for supporting web-based task performance and learning.

1.2 Previous research into composite metaphor interfaces

Composite interface metaphors are composites of two or more metaphors implemented in a single user interface. It has been claimed that composite interface metaphors are to be preferred over integral interface metaphors (i.e., single metaphors describing the entire interface) from a designer point of view. The rationale is that adding new functions to an existing interface is easier when the designer can choose between two or more metaphors for representing functions to users (Neale & Carroll, 1997; Hsu, 2005).

Despite some evidence supporting these claims, the question how well a CMI helps users perform computer-based tasks has not been answered yet unambiguously. Several studies have compared the effectiveness of CMIs to that of single metaphor interfaces from a performance point of view (Hsu, 2005; Hurts & Kooman, 1999; Smilowitz, 1995). Though superior performance for a CMI has been reported (Hsu, 2005; Hurts & Kooman, 1999), these results are not always easy to interpret, because of confounding factors or because of moderating variables.

For example, type of interface was confounded with the number of interaction opportunities afforded by each interface in the Hsu (2005) study. In the same study, the results depended on the level of the user’s expertise. Also, it is unknown whether the superiority of the CMI reported by Hurts and Kooman (1999) was influenced by the fact that the two component metaphors making up this interface were semantically related.
1.3 Overview of article structure

In the next sections, first a componential model for understanding the effectiveness of composite interface metaphors will be introduced. Second, the problem statements of this study will be listed and explained.

Third, Experiment 1 will be described, showing how a metaphorical web interface was developed, composed of two semantically unrelated component metaphors: a library metaphor and a travel metaphor. The interface enabled the performance of a wide range of standard browser tasks using a simulated website. Participants were asked to answer a set of predefined questions about the website. The effectiveness of this interface was assessed against a set of logical and empirical benchmarks, and data were collected on performance and learning that could be used for interface improvement.

Fourth, Experiment 2 will be described, in which an improved version of the CMI and a non-metaphorical version of the same interface were compared empirically. In this experiment, empirical data were also collected on various indices of the participants’ declarative knowledge about metaphor – system mappings. It will be shown how these data can, in principle, be used to improve the effectiveness of the CMI even further.

Finally, the outcomes of the study will be summarized and their implications for theories about, and future research into composite interface metaphors indicated.

2. Componential model

As an aid for deriving specific hypotheses, a macro-level conceptual model was developed for understanding the effectiveness of composite interface metaphors. This model is shown schematically in Figure 1 and is based on the early models developed by Alty, Anderson and others for understanding the effectiveness of single interface metaphors (Alty, Knott,
Anderson, & Smyth, 2000; Hammond & Allinson, 1987; Smyth, Anderson, & Alty, 1995). It is also intended as a pictorial summary of the literature reviewed in the previous section.

For the sake of brevity, the model will be referred to as “the componential model” in order to emphasize the role of a small number of components in explaining a large numbers of ways in which people can interact with metaphorical interfaces. The componential model is also inspired by Tversky’s (1977) model of making similarity judgments and Mac Cormac’s (1985) study on the role of cognition in models about metaphors in linguistics. Tversky’s model emphasizes the importance of both the number of similar elements and the number of dissimilar elements among a set of objects that have to be compared with respect to their overall similarity.

Figure 1 shows the world of the interface functions (system) and the worlds of two interface metaphors as a set of partially overlapping circles. An “interface function” refers to some interface action or task that the interface can perform on the computer or (indirectly) in the real world, irrespective of its physical representation in the interface or of the precise way the task or action should be performed. The various abbreviations and labels mentioned in Figure 1 have the following meaning:

a. S-M+ or mismatch type I: a built-in metaphor suggests non-existing interface functions (i.e., functions not supported by the interface). This is called by Alty et al. (2000) the conceptual baggage of a metaphor.

b. S+M- or mismatch type II: the interface offers access to computer functions that are not suggested by a built-in metaphor; or a built-in metaphor suggests functions that should be activated in a non-metaphorical way. This can be said to indicate low interface coverage by the metaphor.
c. *Matches 1 and 2* (this corresponds to the *quantitative* part of the intersection of the metaphor circles and the system circle): interface functions suggest the presence of metaphorical interface elements (and the other way around) in the intended way (1 and 2 represent the two interface metaphors).

d. *Connectedness 1 and 2* (this corresponds to the *qualitative* part of the intersection of the metaphor circles and the system circle): a built-in metaphor is associated with its intended interface functions in an easy, natural way; it is made up of a coherent set of elements that, nonetheless, each have their own, easily perceptible, role in suggesting interface functions. This can be said to indicate *high connectedness*. Together with matches, this concept can be compared to the notion of *within-metaphor consistency* and to that of a metaphor’s *contextual relevance* (Alty et al., 2000).

e. *Confusions* (only applies to CMIs; this reflects the *non-intended* part of the intersection of the circles labelled Metaphor 1 and Metaphor 2): interface functions suggest the presence of the wrong (not-intended) built-in metaphor, or a built-in metaphor suggests interface functions that, in fact, belong to the other built-in metaphor. Confusions of this type can be said to reflect *low between-metaphor consistency* (Alty et al., 2000).\(^1\)

f. *Overlap* (only applies to CMIs; this reflects the *intended* part of the intersection of the circles labelled Metaphor 1 and Metaphor 2): elements of different metaphors are associated with the same existing interface function(s), or different metaphors actually even share elements. This instantiates what can be called *semantic relatedness* of component metaphors.

Generally speaking, the model of Figure 1 predicts that the higher the combined number of matches, the higher the connectedness, the lower the combined number of mismatches, and

\(^1\) Between-metaphor consistency was defined by Cates (2002) in a slightly different way by equating it with the extent to which auxiliary and underlying metaphors are complementary. In this article, this definition was not adopted, because the componential model does not distinguish between auxiliary and underlying metaphors.
the lower the number of confusions, the more effective the CMI will be. As yet, we do not have specific expectations about the impact of overlap on overall interface effectiveness.

More specifically, for any metaphorical interface to be more effective than a completely non-metaphorical interface, the combined number of mismatches and confusions should be small, compared to the combined number of matches, and the connectedness should be high for both metaphors. Note that in the case of a non-metaphorical interface, the circles labelled Metaphor 1 and Metaphor 2 in Figure 1 would disappear, and interface effectiveness would no longer be dependent on matches, mismatches, confusions, overlap, or connectedness.

Before closing this section, two characteristics of the componential model should be mentioned.

First, mismatches, metaphor confusions, and low connectedness probably can never be completely avoided. However, implicit in the model of Figure 1 is the notion that the negative influence of these factors on overall interface effectiveness can be kept to a minimum through appropriate user training and display design.

Second, many other factors influence the effectiveness of metaphorical interfaces beyond those shown in Figure 1. For example, the precise effects of matches, mismatches, connectedness, and confusions also depend on the physical design of the metaphors, including the interaction techniques used (the “look and feel” of the interface). These factors are not shown in Figure 1, because they were not emphasized in this study. However, both when designing the interface metaphors, and when discussing the results of the experiments, the potential impact of these factors will have to be taken into account.
3. Problem statements

In this section, the problem statements of this study will be described. They were derived from the componential model and from some additional, micro-level models and theories concerning the way people perform cognitive tasks using interactive devices, as will be described below. The following pragmatic starting points constrained the scope of the study.

First, the interface metaphors to be investigated only concerned web browser functions and were composed of two component metaphors: a library metaphor and a travel metaphor. The way these metaphors were designed will be described in section Experiment 1. This was done in order to avoid mismatches type I as much as possible.

Second, the component metaphors were designed to be semantically unrelated. Though it has been claimed that underlying and auxiliary metaphors should be complementary and not be too contrasting (Alty et al., 2000; Cates, 2002), it was felt that this should not be interpreted as a requirement for component metaphors to be always semantically related. Besides, the empirical psychological literature shows that people generally do not have trouble solving problems that can be represented using multiple metaphors, as long as these metaphors do not generate conflicting expectations (see, e.g., Fauconnier & Turner, 2003; Rumelhart & Norman, 1981).

Third, the interfaces to be investigated were (almost) completely described by the component metaphors. This was accomplished by limiting the interface functions as much as possible to those that could be described metaphorically. This was done in order to avoid mismatches type II as much as possible.

The first hypothesis follows naturally from the models and theories reviewed in the previous sections.
Hypothesis 1. (To be addressed in Experiment 1.) A composite metaphor interface to a newly developed, simulated, web browser can be developed, of which the intended system – metaphor mappings can be learned by computer users with various levels of computer experience. Moreover, such an interface allows acceptable performance of web-based tasks after a relatively short learning period.

In Experiment 2, a fine-tuned composite metaphor interface (i.e., an interface improved on the basis of evaluation data obtained in Experiment 1) was compared to a non-metaphorical control interface. Here, it was assumed that the number of matches is large compared to the combined number of mismatches and confusions. In addition, it was assumed that the connectedness of the two interface metaphors would be high:

Hypothesis 2. (To be addressed in Experiment 2.)

A fine-tuned composite metaphor interface allows better web-based task performance than a non-metaphorical control interface.

With regard to the amount of knowledge that could be acquired in Experiment 2, it was expected that, first, the system - metaphor mappings built into a composite metaphor interface would induce more initial declarative knowledge about the meaning of function labels than in a non-metaphorical control condition.

Second, with a composite metaphor interface participants would be able to improve their declarative knowledge quicker over the course of the experiment, because of the motivating effect of minor initial mismatches, as in the case of mismatches type 1 and confusions (see, e.g., Alty et al., 2000; Neale & Carroll, 1997). These initial mismatches would not play a role in the control condition.

Finally, a composite metaphor interface would allow more incidental learning from website content (compared to a non-metaphorical control interface), because learning to
operate the interface would be finished in an earlier phase of the experiment, after which mental resources could be spent on this type of learning. Therefore:

_Hypothesis 3._ (Also to be addressed in Experiment 2.)

_a) With a fine-tuned composite metaphor interface it is clearer to participants what the (declarative) meaning of function labels is than with a non-metaphorical interface._

_b) A fine-tuned composite metaphor interface allows a larger increase of this declarative knowledge, as a result of working with the interface._

_c) More incidental learning of website content will be possible with a fine-tuned composite metaphor interface._

The last problem statement of this study pertains to the identification, probability of occurrence, and performance effects in Experiment 2 of _matches, mismatches, metaphor confusions, and connectedness_, as measured by declarative knowledge tests. First, based on the general literature on cognitive skill acquisition (see, e.g., Anderson, 1982), it was expected that pre-existing declarative knowledge differences would partly mediate the performance benefits attributed by Hypothesis 2 to a composite metaphor interface.

Second, though the two interface metaphors were designed in such a way that the number of metaphor confusions could be kept to a minimum, it was expected that, when a composite metaphor interface was first used, some metaphor confusions could still be observed, as measured by a declarative knowledge test. Because of their richness, initially the two metaphors (library and travel metaphor) probably have several meanings in common, resulting in initial confusions among the intended interface meanings belonging to each of them.

Third, it was expected that the same or a similar declarative knowledge test would indicate relatively high connectedness, when first using a composite metaphor interface (i.e., relatively few confusions among the intended interface meanings within each metaphor). Each of the
two worlds, to which the metaphors refer, usually is experienced in real life as a whole. This will cause relatively few confusions among the intended interface functions belonging to the same metaphor, even before first using the composite metaphor interface.

Fourth, under the same circumstances and using similar tests, it was expected to find that each metaphor offers mental access to many potential interface functions, including some that were not implemented in the tested interface (indicating “conceptual baggage” of the metaphors). In terms of Figure 1, these “errors” correspond to mismatches type I. Therefore:

\textit{Hypothesis 4.} (Also to be addressed in Experiment 2.)

\begin{enumerate}
\item The performance benefit attributed to a composite metaphor interface by Hypothesis 2 can partly be explained in terms of amount of pre-existing declarative knowledge.
\item Declarative knowledge tests will reveal confusions between the two metaphors making up a fine-tuned composite metaphor interface, when this interface is first used.
\item Before first using a fine-tuned composite metaphor interface, declarative knowledge tests will indicate relatively few confusions within metaphors (high connectedness).
\item Before first using a fine-tuned composite metaphor interface, declarative knowledge tests will reveal that metaphors suggest functions that are not present in the interface (“conceptual baggage” of the metaphors).
\end{enumerate}

4. Experiment 1

4.1 Introduction

Experiment 1 aimed at developing a composite metaphor interface (CMI) to a simulated web browser, using existing design guidelines. Usability and learnability were assessed in comparison to a set of logical and empirical benchmarks (Hypothesis 1).

There was only one experimental condition and participants were asked about their amount of general computer and internet experience with the purpose of investigating the
impact of this experience on interface effectiveness in an exploratory way (based on Hsu, 2005).

Interface learnability was tapped using a declarative knowledge test, measuring the amount of declarative knowledge about metaphor-interface mappings. This test was administered both before and after the experiment.

Age differences with respect to web-based task performance and amount of declarative knowledge were also investigated in an exploratory way, based on studies showing the importance of age in understanding the effectiveness of web interfaces in general (Stronge, Rogers, & Fisk, 2006).

At the end of Experiment 1, an analysis was made of the specific interface functions or metaphorical function labels that caused poor web-based task performance or poor performance on the knowledge test. This information was used for the design of an improved interface and/or experimental method, to be used in Experiment 2.

4.2 Method

4.2.1 Participants

Seventeen participants were recruited. Their ages varied from 13 to 77. The 10 youngest participants (until 24 years of age) were all enrolled as full-time university student. The remaining seven participants (from 24 years of age) were, on the average, 50 years of age, and were mainly recruited through an institute for permanent education. Each of them was paid € 6.50 for their participation, which lasted between 40 and 50 minutes. Eighty-one percent of the participants was female.

Participants were required to have normal vision, and not to be (or have been) a student in computer science. Amount of computer and internet experience was measured off-line through a number of questionnaire rating scales.
4.2.2 Website tasks

The website tasks were performed using a simulated website in combination with a simulated, simple web browser. Both website and browser were implemented in Toolbook Instructor 8.0. The website simulated the website of the Computer Science Department at Leiden University as it existed around the year 2000. It consisted of a subset of 67 pages of the real website and was deliberately designed to only contain static webpages which, in turn, only contained text, pictures, (textual) hyperlinks, and a few (pictorial) hotspots. The pages formed a set of sequentially organized pages, divided into several hierarchical levels. Each level of pages was announced at the next higher level of the hierarchy in the form of a menu of topics. Navigation to website pages was also possible using hyperlinks and hotspots (in addition to using browser functions).

There were nine web-based tasks consisting of questions to be answered using the simulated website. The questions required the participants either to look up a piece of information, to save, print, copy, paste, or annotate a piece of text or a page, or to make a page easily accessible for a future visit. The questions were presented on (and the responses were collected through) a set of separate “task pages” that were easy to distinguish visually from the website pages. See Table 1 for a listing of the questions posed in these tasks. Participants could jump back and forth from task pages to website pages as often as they wanted, using a special “go to next question” button on all website pages, and a “go to homepage” button on all task pages.

| Insert Table 1 about here |

4.2.3 Design of the metaphors

The interface metaphors to be incorporated in the simulated web browser were developed using an approach based on the approaches described by Carroll, Mack, and Kellogg (1988), by Alty et al. (2000), and by Cates (2002). These early approaches were combined and
adapted in order to make them suitable for the development of composite interface metaphors. The componential model that was described in the second section provided additional guidance.²

First, sixteen interface functions were defined to be incorporated in the simulated web browser. See Table 2 for a listing of these functions. They were chosen to be rather generic browser interface functions, such as “search”, “copy”, “help”, “save”, and “history”.

Second, two metaphors were chosen for accessing these interface functions. One metaphor was based on the world of books and the library (library metaphor), the other metaphor was based on the world of travel (travel metaphor). Both metaphors are relevant in the context of performing web-based tasks and suggest rather different interface functions. They are also rich metaphors, able to suggest many possible interface functions (Smilowitz, 1995; Neale & Carroll, 1997).

A systematic analysis of properties and operations of both browser and metaphor was conducted, inspired by the three-tier model proposed by Carroll et al. (1988), and by the POPITS model proposed by Cates (2002). From this, it was decided that eight of the sixteen browser functions were best represented by the library metaphor, and the other eight functions were best represented by the travel metaphor, with little apparent ambiguity about the metaphor each function belonged to. For example, the library metaphor emphasizes the informational (books contain retrievable content), material (books can be stored, bookmarked, and written in), hierarchical (books consist of chapters, which, in turn, consist of pages), and sequential (pages are often read in a fixed sequence) nature of the World Wide Web. In contrast, the travel metaphor emphasizes the nature of the World Wide Web as places or

² Though the resulting interface is similar to the interfaces developed by Smilowitz (1995), the design approach of this study is different, because it was based on the componential model of Figure 1. For example, in this study explicit consideration was given to the probability of component metaphors being confused with each other, a consideration that was absent in Smilowitz’ (1995) study.
events to be visited or experienced, to be recorded (in the form of photographs), to travel to in an arbitrary way (aided by shortcuts), and to store in the form of a trip history.

The navigation function “back” was implemented as usual for web browsers, i.e., meaning “go back to the most recently visited page”. Activation of the navigation function “forward” would cause a jump to the first website page (or topic) expanding the currently visited page, if possible. Otherwise, activation would cause a jump to the next topic that belonged to the same hierarchical level of topics, if possible. In all other cases this function could not be activated.

Third, labels and icons were designed for each interface function in such a way that the latter would be suggested as well as possible by the former. See Table 3 for a listing of the metaphorical labels we came up with.

Insert Table 2, Table 3, and Figure 2 about here

Table 3 shows for which ten functions icons were designed in Experiment 1. Figure 2 shows these icons in the form of buttons. Corresponding function labels would become visible as soon as the mouse pointer hovered over or near the button in the form of a so-called “tooltip”. The functions for which no icons were designed were only represented by labels and were activated through options in the main menu of the browser.

Note that the metaphors were not embedded in the physical appearance, or “look and feel”, of the webpages themselves. Rather, they were only part of the browser interface. Hence, the name browser metaphor.

4.2.4 Procedure

First, briefing of the participants took place and an informed consent was filled out. As an introduction, participants were told in general terms that the purpose of the experiment was to “investigate new types of web browser”.

Second, the pretest questions belonging to the declarative knowledge test were answered and a background questionnaire was filled in. On this questionnaire, participants indicated
their experience with computers in general (5-point rating scale), the frequency of them using computers in general (3-point rating scale), and the frequency of them using internet in general (3-point rating scale).

Third, a written instruction was read aloud by the experimenter, followed by an on-line instruction and demonstration of a few common computer concepts (i.e., mouse cursor, menu items, buttons, hyperlinks, and so on). Participants were also asked to perform a simple search task to check their understanding of the computer concepts.

Fourth, the experiment itself was started in which participants were asked to complete the nine website tasks on their own. Instructions emphasized the importance of both accuracy and speed of task performance.

Finally, the posttest questions belonging to the declarative knowledge test were answered. The posttest questions were identical to the pretest questions.

4.2.5 Experimental design

Only the CMI was used in Experiment 1. See Figure 2 for a screenshot of this interface.

Web-based task performance was measured in terms of accuracy and speed of performing all nine website tasks (see Table 1). Accuracy was expressed as the number of questions answered correctly using the correct method. A task was considered partly completed (corresponding to half an accuracy point) if only the question was answered correctly, or if only the correct method was used. Performance speed was expressed in minutes needed to complete all website tasks.

In order to test Hypothesis 1, we used an accuracy benchmark of 4.5 (50%) correctly answered website questions, on the average, or higher. With respect to speed, we used a benchmark of, on the average, 100 seconds per task. This time limit was derived from the study by Hsu (2005), who observed that most participants were able to perform simple search
tasks (comparable to the ones used in the present study), using a metaphorical interface, within 100 seconds.

**Acquisition of declarative knowledge** was measured by means of the declarative knowledge test. This test consisted of nine multiple-choice questions about the mappings from function label to intended interface function, and the other way around. A participant’s knowledge acquisition was expressed as that person’s pre-post difference in the accuracy of answering these questions. The same questions were used for pretest and posttest. Each question was assigned a 1 for “correctly answered” or a 0 for “incorrectly answered”. For two example questions, see Appendix 1.

In order to test Hypothesis 1, we required that the CMI allow a significant increase in declarative knowledge accuracy.

### 4.3 Results

Table 4 contains the descriptive statistics for the performance indices (accuracy and time of completing web-based tasks), the declarative knowledge test scores, and a few background variables. It can be seen that most participants indicated to have 2 - 5 years of computer experience. Closer analysis of the data reveals that about 36% of all participants reported to have at most 2 years of computer experience (similar to the “novices” studied by Hsu, 2005). On the other hand, about 23% of the participants claimed to have more than 5 years of computer experience (similar to the “experts” studied by Hsu, 2005).

4.3.1 *Pre-existing familiarity with metaphorical function labels*

In order to find out whether how successful the metaphorical function labels had been designed, the declarative knowledge scores were analyzed that were obtained before the start of the experiment. It turned out that the average accuracy (5.41) was significantly higher than what would be expected on the basis of chance, \( t(16) = 8.93, p < 0.001 \).
This is taken this as evidence that the participants initially recognized the meaning of the metaphorical function labels, based on their previous experience with the metaphors.

4.3.2 Usability and learning results (Hypothesis 1)

All website tasks listed in Table 1 could be completed without external help or interruptions. However, the answers given to website question 8 could not be measured reliably, due to problems in the software responsible for recording these answers. Therefore, accuracy scores were only computed for the eight remaining questions.

Table 4 shows that the average accuracy of performing website tasks was about 63% (5 out of 8 questions answered correctly). The average speed of performing all nine tasks was about 2.5 minutes per task.

With respect to the acquisition of declarative knowledge, the results show that the average increase in the accuracy of this knowledge was 1.24 points (average pre-post difference). This increase was significantly different from 0, $t(16) = 3.26$, $p < 0.01$.

4.3.3 Effects of age and computer experience (exploratory questions)

For each participant a single computer experience score was computed by summing the scores of the three rating scales pertaining to computer experience. This overall score ranged from 3 (least computer experience) to 11 (most computer experience).

The correlation with performance (not shown in Table 4) was marginally significant for time ($r = -0.46$, $p < 0.10$) and significant for accuracy ($r = 0.65$, $p < 0.05$) in the expected direction: more experienced people performed better. The correlation with increase in declarative knowledge accuracy was positive, though not significantly so, $p > 0.10$. These results mimic those of previous studies into the effect of general computer expertise on the effectiveness of a metaphorical computer interface (Hsu, 2005).

With respect to age, analyses revealed that, compared to younger participants, older participants:
1. Performed worse on web-based tasks. The correlation with age was statistically significant, $p < 0.05$, for time-on-task ($r = 0.54$), and marginally significant, $p < 0.10$, for performance accuracy, $r = -0.46$.

2. Had lower declarative knowledge accuracy, but this was only true for the posttest scores ($r = -0.45$, $p < 0.10$).

In addition, the number of questions answered correctly on the declarative knowledge test was positively correlated with the number of website tasks completed (marginally significant correlation, $p < 0.10$), but this was only true for the posttest.

In summary, web-based task performance was age-dependent in the direction predicted by Stronge, Rogers, and Fisk (2006): older participants performed worse. Note, however, that this effect could partly be explained in terms of declarative knowledge accuracy, which at the end of the experiment was lower for older participants.

It should also be noted that in this experiment age was correlated with (self-rated) general computer experience ($r = -0.71$, $p < 0.01$): older participants were less experienced.

### 4.4 Conclusions and redesign of the composite metaphor interface

Though the average accuracy of performing website tasks seemed acceptable (63% correct), the average speed of performing these tasks seemed slow (2.5 min per task). Participants were able to improve their declarative knowledge accuracy significantly, but the size of the improvement (1.24 accuracy points, on average) was not large in an absolute sense. In addition, we cannot exclude the possibility that participants improved their performance on the declarative knowledge test merely because of repeated exposure to the same questions (sensitization). Therefore, we concluded that Hypothesis 1 was only partly confirmed and that the CMI of Experiment 1 was in need of revision.

The need to improve the CMI was all the more apparent when one takes the fact into account that the novice and older participants of Experiment 1 performed worse than the
younger participants and those who considered themselves experienced computer users. Therefore, in an attempt to design a better interface, the metaphors and corresponding interface functions were simplified as follows.

First, an analysis was made of the difficulty-values of the nine pretest and nine posttest questions (each value indicating the fraction of participants that answered the corresponding question correctly). The lower the value, the smaller the number of participants that understood the intended meaning of the metaphorical function labels. The lowest scoring questions (difficulty-values lower than 0.30) dealt with function labels Search shelves again (search again) and Make reservation (select). Therefore, it was decided to remove the corresponding functions from the interface.

Second, a similar analysis was performed for the nine website tasks. This showed that all tasks were relatively easy to perform (had difficulty-values of 0.5 or higher), except task 9 which had a difficulty-value of only 0.28 (see Table 1). This task dealt with function label Borrow book (print). Therefore, the corresponding function was removed from the interface as well.

Third, the functions indicated by the labels Make shortcut (make link) and Stop travelling (stop) were removed from the interface, because, in hindsight, they did not add much functionality to the interface.

Fourth, of the remaining 11 interface functions, the navigation functions First, Back, and Forward were moved to a separate navigation panel, using the familiar icons (such as ► and ◄) that are also found on many CD- and video-players. Because these icons are so familiar to present-day users, it was assumed that using them would not introduce a new interface metaphor.

Finally, the label Search shelves (search) was changed into Travel to, turning it from a library metaphor function label into a travel metaphor function label. Also, the label Tour
Guide (help) was changed into Help desk, turning it from a travel metaphor function label into a library metaphor function label. This was done because Travel to seems more compatible with the function of browsing on the world-wide web than Search shelves. Similarly, Help desk seems more compatible with the behaviour of active information inquiry than Tour Guide. Table 5 lists the set of metaphorical function labels that were used in Experiment 2.

Insert Table 5 about here

5. Experiment 2

5.1 Introduction

The Experiment 1 outcomes beg the question how well the redesigned CMI performs (see Table 5 for a listing of the new metaphorical function labels) and how it compares to a completely non-metaphorical interface with respect to usability and learning. This was investigated in Experiment 2, which partly replicated Experiment 1 ((Hypotheses 2, 3a and 3b).

In order to take the findings of Experiment 1 into account regarding the effects of age and level of experience, Experiment 2 used a more homogeneous sample, consisting only of younger participants with a limited amount of computer and internet experience.

Moreover, several measures were taken in Experiment 2 in order to:

1. Investigate the probability of mismatches type I (Hypothesis 4d) using a free association task. In this task, participants were asked to state how they thought the two metaphors could be used for denoting interface functions. These data were also collected to complement the declarative knowledge test data that were used for testing Hypothesis 4b and c.

2. Investigate the confusions made by participants within and between metaphors, as measured by the declarative knowledge test (Hypothesis 4b and c).
3. Measure incidental learning of website content using a recognition task (Hypothesis 3c).

Finally, the question was addressed whether performance differences between the two types of interface can be explained in terms of differences in amount of pre-existing declarative knowledge (Hypothesis 4a).

5.2 Method

Below, only the differences with the method section of Experiment 1 are mentioned.

5.2.1 Participants

Each of 21 participants was assigned at random to one of two experimental conditions. Most of them were enrolled as full-time university student and they had ages ranging from 18 to 26 years. They were required to have at most 2 years of (self-assessed) computer and internet experience.

5.2.2 Equipment and website tasks

Only six website tasks were used.

5.2.3 Procedure

Association task: before the start of the website tasks, but after the declarative knowledge test was filled in for the first time, all participants were asked to write down as many properties, events, or objects as possible, when thinking about the phrases “library world” and “world of travel”. In addition, they were asked to write down how each property, event, or object, they came up with, could be used to denote interface functions.

Declarative knowledge test: now there were only eight multiple-choice questions (instead of nine) in this test. The questions were the same for both conditions, but for the control condition, the questions referred to the neutral function labels used in the interface of that condition.
Recognition task: incidental learning of website content was measured by means of a recognition task, performed at the end of the experiment. In this task, participants were to indicate from a list of ten topics those (s)he remembered having seen in the website. Only six of these topics had actually been covered by the website, but participants were not told this. Performance on the recognition task was expressed as the sum of the number of correctly recognized topics and the number of correctly rejected (not recognized) topics (the maximum score being 10).

5.2.4 Experimental design

Composite metaphor condition. This condition contained 10 participants. The four library metaphor functions and the four travel metaphor functions, listed in Table 5, were combined into a single interface. Only the actions enabled by the navigation panel were denoted in a non-metaphorical way. All functions were implemented through buttons, each of which was identified by means of an icon and a label (see Figure 3 for a screenshot).

Control condition. The control condition contained 11 participants. All eight interface functions were implemented using a button only carrying a non-metaphorical, neutral label without icon. To this end, it was decided to use the names of a random group of letters taken from the Greek alphabet. See Table 5 for a list of these labels in the column labelled “Control”.

In a previous version of this experiment, the labels appearing in the first column of Table 2 were used for the control condition. However, it turned out that these Internet Explorer (IE)-like labels were very familiar to most participants, due to the popularity of the IE-browser. Therefore, it was decided that these non-metaphorical labels would create a source of bias in the comparison with the composite metaphor condition, where less familiar (though metaphorical) function labels were used. By using the names of Greek letters, an attempt was made to avoid such a bias.
Independent and dependent variables. The independent variable type of metaphor/interface had two levels: composite metaphor condition or control condition. Performance speed was now expressed as the total number of seconds needed to perform all six website tasks.

5.3 Results

5.3.1 Web-based task performance (Hypothesis 2)

Table 6 shows the mean performance scores observed for the two conditions. It can be seen that the participants in the composite metaphor condition needed (on the average) almost 100 seconds less time for completing the website tasks than in the control condition (averages 662 sec versus 758 sec, respectively). At the same time, they performed these tasks slightly less accurately (4.35 versus 4.59 tasks performed correctly, respectively). However, neither of these differences was statistically significant, \( p > 0.10 \). (Normality of the time distributions was assumed for each condition, because the skewness and curtosis values of the distributions were not significantly different from 0.)

5.3.2 Learning (Hypothesis 3)

Table 6 also shows that the participants in the composite metaphor condition possessed more declarative knowledge about label – interface mappings (average score 6.1) than in the control condition (average score 3.5), both before the start of the experiment and at the end of the experiment, \( p < 0.05 \). This difference was statistically significant, \( F(1, 19) = 23.82, \ p < 0.001 \). The average pre-post difference (increase of 4.7) was statistically different from 0, \( F(1,19) = 27.74, \ p < 0.001 \), indicating that participants improved their knowledge as a result.
of having worked with the web interface. However, the pre-post differences for the two conditions did not differ significantly, $p > 0.10$ (non-significant interaction), indicating that the participants in the control condition improved as much as in the composite metaphor condition.

With respect to the incidental learning of website content, participants obtained, on the average, a score of 7.4 (out of 10), the scores being somewhat higher for the control condition. The scores differed significantly from 5 (i.e., chance level of responding), $p < 0.001$, e.g., $t(9) = 6.74$ for the composite metaphor condition. However, the two conditions did not differ statistically in this respect, $p > 0.10$.

5.3.3 Performance benefit explained in terms of pre-existing declarative knowledge

(Hypothesis 4a)

As was seen, there was no evidence for a performance benefit for the composite metaphor condition in this experiment. Apparently, the above mentioned advantage in pre-existing declarative knowledge accuracy for this condition did not translate into a corresponding performance benefit. Closer inspection of the data shows that this state of affairs can be attributed to the lack of an overall correlation between pre-existing declarative knowledge accuracy on the one hand, and the two performance indices on the other ($r = -0.21$ for time and $r = 0.19$ for accuracy).

However, further analyses of the performance time data revealed that the correlations with pre-existing declarative knowledge accuracy were rather different for the two conditions (see Figure 4). Whereas the correlation was moderately negative for the control condition ($r = -0.45$), which could be expected, it turned out to be moderately positive for the composite metaphor condition ($r = 0.50$), which was not expected. In other words, the more accurate the pre-existing declarative knowledge, the slower the performance on the website tasks. The difference in correlation coefficients was statistically significant, $F(1,21) = 4.49$, $p < 0.05$ (for
details about how this difference was tested, see description of covariate 2 below). (A similar difference in correlations was not observed for the accuracy data.)

In summary, this study’s finding that performance time was not significantly different for the two conditions may be partly due to this positive correlation with pre-existing declarative knowledge accuracy. In order to show this effect in more detail, the performance time data of Experiment 2 were re-analyzed using analysis of covariance. As covariates served the following two variables:

1. Pre-existing declarative knowledge accuracy.
2. The interaction between group and pre-existing declarative knowledge accuracy. This interaction was represented by means of a dummy variable, created by multiplying the covariate 1 values by 1 for the composite metaphor condition and by 2 for the control condition.

Results showed that the inclusion of these two covariates increased the size of the effect of condition in the direction expected by Hypothesis 2: the corrected mean for the composite metaphor condition went down to 337.57 sec, and the corrected mean for the control condition went up to 1053.21 sec. This effect was statistically significant, $F(1,21) = 4.59, p < 0.05$.

5.3.4 Analysis of declarative knowledge confusions between and within metaphors

(Hypothesis 4b and c)

In order to test the hypotheses about connectedness and between-metaphor confusions, the errors made on the declarative knowledge test (composite metaphor condition only) were divided into four groups:
1. Errors amounting to incorrect associations between *library* metaphor labels in the question and *travel* metaphor interface functions in the answer alternatives.

2. Errors amounting to incorrect associations between *travel* metaphor labels in the question and *library* metaphor interface functions in the answer alternatives.

3. Errors amounting to incorrect associations between *library* metaphor labels in the question and *library* metaphor interface functions in the answer alternatives.

4. Errors amounting to incorrect associations between *travel* metaphor labels in the question and *travel* metaphor interface functions in the answer alternatives.

Errors belonging to groups 1 and 2 were added and counted as between-metaphor confusions in declarative knowledge. Errors belonging to groups 3 and 4 were added and counted as within-metaphor confusions in declarative knowledge. In order to make the scores for between-metaphor confusions and within-metaphor confusions comparable, they were expressed as deviation scores, indicating the degree to which they were larger or smaller than the number of confusions that could be expected on the basis of chance (guessing) in each category.

Table 7 shows the mean scores for each type of confusion, for each learning phase (before or after the experiment), and for each condition. The scores for the control condition were only used as reference values. They were computed by classifying an error as a between-metaphor confusion if the corresponding answer in the composite metaphor condition would be classified as such. The same was true for within-metaphor confusions. However, strictly speaking, the distinction between between-metaphor and within-metaphor confusions does not make sense for the control condition.

Insert Table 7 about here

From Table 7 we conclude the following:
First, focussing on the composite metaphor condition, the average number of between-metaphor confusions was larger than the number of within-metaphor confusions, especially before the start of the experiment (0.05 and -2.85 for between-metaphor and within-metaphor confusions, respectively). The overall difference between the two types of error was statistically significant, $F(1,9) = 30.49$, $p < 0.001$. However, at the end of the experiment, the difference was reduced to only 1 point: -2.05 and -3.05 for between-metaphor and within-metaphor confusions, respectively, $F(1,9) = 4.87$, $p < 0.10$.

Second, again focussing on the composite metaphor condition, the improvement in the average score for between-metaphor confusions (reflected by decreasing values in Table 7) was larger than that for within-metaphor confusions. Statistical testing showed that this difference was significant, $F(1,9) = 13.05$, $p < 0.01$ (significant interaction).

Third, statistical analysis showed that the interaction between type of condition and time of measurement (before or after the experiment) was statistically significant, $F(1,19) = 6.80$, $p < 0.05$, for the between-metaphor confusions. This indicates that participants in the composite metaphor condition were better able to unlearn incorrect associations between labels and functions. However, a similar interaction was not observed for the within-metaphor confusions, $p > 0.10$.

5.3.5 Association task (Hypothesis 4b, c, and d)

The responses to the questions in the association task were analyzed only qualitatively and globally. The results can be summarized as follows.

First, participants had less difficulty thinking of interface functions when confronted with the phrase “world of travel” than with the phrase “library world”. Also, participants mentioned more navigation functions in response to the world of travel. Apparently, they experienced the travel metaphor as richer than the library metaphor.
Second, many interface functions were mentioned by the participants that were not part of the interfaces in this experiment, for example: “filtering search results”, “set security level”, and “printing” in response to library world; and “set colour”, “route planning”, “set language”, “email”, and “checklist” in response to world of travel. In terms of Figure 1, these “errors” correspond to mismatches type I and indicate that the metaphors possess conceptual baggage (Alty et al., 2000).

Third, some metaphorical function labels that were used in this experiment, were not mentioned by the participants. Examples: “Show picture”, “Archive”, and “Post-it”. In other words, these are concepts that do not come to people’s minds easily, when they think about the library world or the world of travel. They indicate low connectedness of the metaphors (or low within-metaphor consistency).

Fourth, sometimes interface functions were mentioned in response to the “wrong” metaphor (between-metaphor confusions). These indicate low between-metaphor consistency. The following examples were found: functions “go to previously visited pages”, “help” or “information” were mentioned a few times in response to the library world (whereas they were designed to be denoted by the travel metaphor). Similarly, “favourites” was mentioned a few times in response to the world of travel (though designed to be part of the library metaphor).

5.4 Conclusions

5.4.1 Comparison with Experiment 1: retesting Hypothesis 1

The participants in the composite metaphor condition completed, on the average, about 2% more website tasks successfully than in Experiment 1. In addition, they performed these tasks in 22% less time, and were able to improve their declarative knowledge about label – interface mappings beyond that observed in Experiment 1 (increase of 4.7 versus 1.6 accuracy points, respectively). (Performance scores were corrected for the number of website tasks performed; in Experiment 1, the average scores obtained by the younger participants were taken
as reference values.) As was true for Experiment 1, participants may have performed better on the declarative knowledge posttest merely because of repeated exposure to the same questions. However, because the increase in declarative knowledge was substantially larger in Experiment 2, repeated exposure cannot be the only reason for the amount of declarative learning that took place in this experiment.

Closer inspection of the performance difficulty values for the composite metaphor condition revealed that none of the website tasks was performed incorrectly by more than 40% of the participants. Similarly, at the end of the experiment all difficulty values for the declarative knowledge questions were higher than 0.70. These values exceed those observed in Experiment 1.

Finally, participants were able to acquire a significant amount of knowledge about website content through incidental learning.

It is concluded that the attempt to make the CMI more usable and more learnable than in Experiment 1 was successful: Hypothesis 1 was reconfirmed in Experiment 2.

5.4.2 Comparing metaphorical with non-metaphorical interface

The declarative knowledge test data of this experiment showed that the metaphorical function labels of the CMI suggested their intended meaning better, both before and after the experiment (Hypothesis 3a confirmed). Moreover, covariance analysis showed that the two groups of this experiment differed significantly (and in the expected direction) in the speed of performing web-based tasks, after group differences in pre-existing declarative knowledge accuracy had been taken into account, but not in the accuracy of web-based task performance (Hypothesis 2 partially confirmed).

Meanwhile, it must be concluded that the group difference in declarative knowledge accuracy did not translate into a corresponding performance difference (Hypothesis 4a not confirmed). Neither did the increase in declarative knowledge accuracy or the incidental
learning of website content differ significantly for the two groups (Hypotheses 3b and c not confirmed).

5.4.3 Between-metaphor confusions, within-metaphor confusions, and mismatches
The declarative knowledge test data also showed that there were relatively many between-metaphor confusions, at least initially. However, the ability to discriminate the two metaphors (in a declarative sense) improved over time. This indicates that this ability may benefit from operational experience, i.e., from performing website tasks (Hypothesis 4b confirmed in a relative sense). Within-metaphor confusions occurred less often and did not decrease as much over time. This indicates that Hypothesis 4c was confirmed, despite some evidence for low metaphor connectedness in the association task data.

The latter data also revealed several cases of between-metaphor confusion (Hypothesis 4b reconfirmed in an absolute sense). In addition, they revealed cases of metaphor mismatches type I (Hypothesis 4d, about metaphor richness, confirmed).

5.4.4 Other findings
An analysis of mouse clicks used during the performance of website tasks showed that the participants in the control condition needed about twice as many mouse clicks as in the composite metaphor condition, $F(1, 19) = 5.36, p < 0.05$ (values not shown in tables). This suggests that these participants were exploring the web interface more actively. This was probably done in an attempt to compensate for the handicap of having to figure out the meaning of the neutral function labels through a process of trial-and-error.

Further discussion of possible explanations and implications of these results will follow in the next section, Conclusions and Discussion.
6. Conclusions and discussion

6.1 Summary of study objectives

In this study, a composite metaphor interface (CMI) was developed using an accepted design process consisting of logical and empirical benchmarks. The interface was built around two semantically unrelated component metaphors, and offered access to standard functions of a simple, simulated web browser. A simulated, but realistic website served as test bed. Half of the interface functions to which participants had access were described by the library metaphor, the other half was described by the travel metaphor. In a second experiment, an improved version of the interface was compared to a non-metaphorical control interface consisting of neutral function labels. Several hypotheses were tested in this study, most of them derived from a componential model for understanding the effectiveness of CMIs (Figure 1).

6.2 Findings of Experiment 2

The results of Experiment 2 confirmed that the attempt to improve the usability and learnability of the CMI designed in Experiment 1 had succeeded: website tasks were performed better and faster, and the improvement over the course of the experiment in declarative knowledge about metaphor – interface mappings was larger. Moreover, a significant amount of incidental learning took place over the course of the experiment. This shows that the detailed investigation of performance errors conducted at the end of Experiment 1 had paid off.

As expected, the CMI also resulted in significantly higher declarative knowledge accuracy, both before and after using the interface for the first time, compared to a control interface using neutral, non-metaphorical function labels. Comparing the two types of interface with respect to web-based task performance, only a partial confirmation of the expectations was found. In terms of performance accuracy, no difference between the two
types of interface was observed. However, in terms of performance speed the expected benefit for the CMI was observed, but only after the results had been corrected for pre-existing declarative knowledge accuracy: web-based tasks were performed faster with the CMI.

Apparently, the declarative knowledge benefit, that was observed for the composite metaphor interface, did not translate (as was expected) into a corresponding performance benefit for that interface. Indeed, for performance speed, an opposite effect was observed: the difference between groups in declarative knowledge accuracy caused an existing difference in speed to disappear (become insignificant). In other words, the speed difference would have remained significant if all participants had possessed an equal amount of declarative knowledge.

The absence of an effect of type of interface on performance accuracy in Experiment 2 may be explained in several ways. Only the three most likely explanations are mentioned below. First, the duration of the experiment may have been too short for the declarative knowledge advantage of the CMI to translate into a performance benefit.

Second, and closely related to the issue of duration, perhaps participants did not really need metaphors in order to complete the website tasks used in this study. The web interfaces they were to work with were relatively simple and the metaphors were not embedded in the webpages themselves. A brief episode of paired-associates learning may have been all that participants needed in order to become accurate performers of website tasks, regardless of the type of interface.

Finally, it may be that the behaviour of the experimental web interfaces conflicted with participants’ previous computer experience and their existing mental models about the operation of web interfaces in general. This previous experience may have interfered with the way participants approached performance of the website tasks. Though in Experiment 2 the amount of previous computer and internet experience was deliberately kept to a minimum and
(through random assignment of participants to conditions) controlled to be equal for both conditions, it cannot be excluded that participants in the composite metaphor condition were particularly vulnerable to interference caused by previous experience. This latter hypothesis receives some empirical support from the counterintuitive correlation between pre-existing declarative knowledge accuracy and performance speed that was observed for the composite metaphor condition (see Figure 4).

The absence of a *learning* benefit for the CMI (i.e., both types of interface allowed an equal amount of learning), as measured by the declarative knowledge test, may represent a ceiling effect: at the end of the experiment, the declarative knowledge test scores were almost at their maximum value for the composite metaphor condition (see Table 6). With respect to incidental learning of website content, it may be that the absence of a learning benefit in this regard was partly due to the fact that in this study the participants were not prepared to be tested on their knowledge of website content. This may have inhibited their motivation to learn in this way, explaining the lack of a difference between the two conditions.

Despite the absence of a learning benefit for the CMI, declarative knowledge improved over the course of the experiment for both types of interface. This indicates that operational experience is beneficial for learning interfaces in a declarative sense. In the composite metaphor condition, confusions occurred especially between metaphors and especially at the beginning of the experiment. Apparently, there is something about the combination of two component metaphors in a single interface that users have to get used to, more so than they have to get used to the use of metaphors per se.

The association task data provided evidence for both metaphors being semantically rich, at least to a certain extent, the travel metaphor being, on the average, richer than the library metaphor. From the componential model of Figure 1, it follows that metaphor richness (or the occurrence of mismatches type I) may be a partial explanation for the absence of a perform-
ance benefit for the CMI. However, paradoxically, other research suggests that rich metaphors may also be beneficial, especially when first using a metaphorical interface (Alty et al., 2000). This paradox presents an interesting issue to be addressed by future research.

More generally, mismatches of any type cannot be completely avoided in metaphorical interfaces. Neale and Carroll (1997) suggest that this disadvantage be turned into an advantage and that an appropriate interface design encourage its users to focus on (and, hence, learn) aspects of interface or computer functionality that go beyond the world of a single metaphor, for example, by using composite metaphors. It may be that the sharp improvement in the discrimination between metaphors that was observed in Experiment 2 actually demonstrates what Neale and Carroll were referring to. However, more research is needed to test the generality and robustness of these findings.

6.3 Remaining findings

The Experiment 1 finding that younger and more experienced participants performed better on the website tasks confirms the findings of previous studies (e.g., Stronge, Rogers, & Fisk, 2006; Hsu, 2005) that showed the importance of age and experience in determining the effectiveness of metaphorical interfaces. Specifically, in this study the effect of age could partly be explained in terms of older participants being less experienced and less able to improve their (declarative) metaphorical knowledge over the course of the experiment. Future research is needed to separate the effects of age and experience that turned out to be confounded in this study: for example, it is likely that future seniors will be more technology savvy and, therefore, will be better performers of website tasks.
6.4 Implications of findings for theories and future research

How well does the model of Figure 1 stand in the face of the findings of this study?

First, the reader should be reminded that the model of Figure 1 was only used as a framework for deriving more specific hypotheses. These hypotheses also depended on other (micro-level) theories and models. Therefore, the findings of this study do not allow a direct test of the quality of the componential model in its entirety.

Second, the model can be of help in explaining the experimental findings, even if they were not expected. For example, the fact that no performance accuracy benefit was observed for the CMI (compared to the non-metaphorical control interface) can also be explained in terms of the negative performance effects associated with metaphor mismatches type I and between-metaphor confusions.

Third, the redesign of the CMI of Experiment 1 was partly based on the declarative knowledge data collected in that experiment, and, therefore, partly on data regarding metaphor connectedness and between-metaphor confusions.

Fourth, the association task data collected in Experiment 2 pertaining to metaphor mismatches type I and between-metaphor confusions can be used for improving the design of the CMI even further. The same is true for some of the declarative knowledge data collected in that experiment.

Finally, in future experiments it is worth investigating the effect of making the two component metaphors of this study more distinguishable, for example, by using visual means to separate or group the buttons and function labels belonging to each metaphor. In terms of the componential model, this might help to increase between-metaphor discrimination.
Acknowledgements

Nelleke van Wouwe and Ronald Vos are acknowledged for their help with collecting the data reported in this article in partial fulfillment of their master’s degree in Cognitive Psychology at Leiden University.
References


Appendix 1. Examples of multiple-choice questions, measuring the learning of declarative knowledge

Note: The first question is about a library metaphor function, the second one about a travel metaphor function. Only used in conditions in which metaphors are used for accessing interface functions.

1. In order to store pages which contain interesting topics for future visits, I would use the following function.
   A. ‘Bookmark’
   B. ‘Search shelves’
   C. ‘Post-it’
   D. ‘Archive’

2. What do you think you can do with function ‘Make photograph’?
   A. Make a copy of a picture, webpage (part) or document.
   B. Copy a particular idea from another website for one’s own use.
   C. Re-use a piece of text.
   D. Merge two pictures, pieces of text, or webpages.
About the authors

Dr Karel Hurts obtained a Ph.D. degree in Experimental Psychology/Human Factors (Texas Tech University, 1985). He worked as research associate at the University of Groningen, and as assistant professor in Cognitive Psychology at Leiden University, The Netherlands.

Currently, he is adjunct professor in Psychology at Webster University, Leiden, The Netherlands. He also owns CogniTech, a firm for conducting applied cognitive services in the Netherlands.
Figure captions

Figure 1. Componential model showing various generic factors determining the overall effectiveness of a composite metaphor interface.

Figure 2. Screenshot of website and browser interface in composite metaphor condition (Experiment 1).
Note: All buttons carry Dutch labels as follows: “Home” (first button), “Make photograph”, (seventh button), “Show picture” (eighth button), “Travel log” (ninth button), and “Make shortcut” (tenth button) for travel metaphor. The remaining labels belong to the library metaphor: “Previous book” (second button), “Next book” (third button), ”Bookmark” (fourth button), “Post-it” (fifth button), and “Search shelves” (sixth button). The top-right white button contains the Dutch translation of “Go to next question”.

Figure 3. Screenshot of website and browser interface in composite metaphor condition (Experiment 2).
Note: Translations of (Dutch) labels appearing beneath each icon in top row are as follows (from left to right): “Travel to”, “Travel log”, “Make photograph”, “Show picture”, “Post-it”, “Bookmark”, “Archive”, and “Help desk”.

Figure 4. Scatterplots showing relationships holding in Experiment 2 between pre-existing declarative knowledge accuracy and time-on-task, for each of the two types of interface (composite metaphor or control).
Note: Pre-existing declarative knowledge accuracy expressed as number of questions answered correctly (out of 8) on declarative knowledge test before beginning of experiment. Time-on-task measures number of seconds needed to complete all web-based tasks.
Figure 1
Universiteit Leiden

De opleiding informatica

- Informatica student
- Deelhouding
- Aangepast programma
- Wegwijzer informatica
- Kaart
- Universiteit Leiden
- Agenda
- Curriculum
- Roosters
- Vakken
- Examen
- LIACS (medewerkers)
De opleiding informatica

- Informatica studeren
- Deeltijdopleiding
- Aangepast programma
- Waguizer informatica
- Kaart
- Universiteit Leiden

- Agenda
- Curriculum
- Boekjes
- Vakken
- Bijval
- LIACS (medewerkers)

- SOEL
- Opleidingscursussen
- Studieadviesing

Figure 3
Composite Interface Metaphors

Figure 4

Composite Metaphor

Control

Pre-existing declarative knowledge accuracy

Time [sec]
Table 1. List of website tasks used in Experiment 1.

<table>
<thead>
<tr>
<th>For each of these tasks, always use the menu items and buttons, if possible.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Make the page containing the roster for computer science as a part-time study easily accessible for a next visit.</td>
</tr>
<tr>
<td>2. How many pages are devoted to LIACS?</td>
</tr>
<tr>
<td>3. Store the page “Study expenses” as a file on the computer.</td>
</tr>
<tr>
<td>4. What is the title of the page preceding the page mentioning the word “Theology”?</td>
</tr>
<tr>
<td>5. Look up the page “Multimedia”. What pages or text will be found if this information is looked up again?</td>
</tr>
<tr>
<td>6. Attach a personal note to the page containing the city map of Leiden.</td>
</tr>
<tr>
<td>7. Duplicate a piece of text from the page “Educational Committee Computer Science” and place this in the response window of that page.</td>
</tr>
<tr>
<td>8. Look up the third page you have visited until now.</td>
</tr>
<tr>
<td>9. Make a paper copy of the first page. Immediately afterwards, cancel the command.</td>
</tr>
</tbody>
</table>
Table 2. Definitions for each of the sixteen interface functions (Experiment 1).

<table>
<thead>
<tr>
<th>General Function Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search again</td>
<td>Make a new search for the previously searched topic or page.</td>
</tr>
<tr>
<td>Search</td>
<td>Search for a topic or page entered by the user in a search box.</td>
</tr>
<tr>
<td>Save</td>
<td>Save the current page as a file on the computer.</td>
</tr>
<tr>
<td>Favourites</td>
<td>Add this page to the list of pages made easily accessible.</td>
</tr>
<tr>
<td>Forward</td>
<td>Move to the next page.</td>
</tr>
<tr>
<td>Print</td>
<td>Print the current page.</td>
</tr>
<tr>
<td>Annotate</td>
<td>Add a comment to the current page for later use.</td>
</tr>
<tr>
<td>Back</td>
<td>Return to the last-visited page.</td>
</tr>
<tr>
<td>First</td>
<td>Go to the first page.</td>
</tr>
<tr>
<td>Select</td>
<td>Select a text fragment, picture, or document.</td>
</tr>
<tr>
<td>Copy</td>
<td>Make a duplicate of the selected text or picture.</td>
</tr>
<tr>
<td>Paste</td>
<td>Move the duplicated text or picture to a user-defined place.</td>
</tr>
<tr>
<td>History</td>
<td>Show the history of webpages visited before. These pages appear in a list as hyperlinks so that pages can immediately be visited.</td>
</tr>
<tr>
<td>Make link</td>
<td>Make current page easily accessible from outside the browser.</td>
</tr>
<tr>
<td>Stop</td>
<td>Stop loading of page currently being loaded.</td>
</tr>
<tr>
<td>Help</td>
<td>Request on-line help about a user-defined topic.</td>
</tr>
</tbody>
</table>
Table 3. Interface function labels and corresponding general function names (Experiment 1).

Note: Function names printed in italics were activated through menu items containing labels only. The other function names were activated through buttons (icons) and described by tooltips (labels) if the mouse pointer hovered over or near the button.

<table>
<thead>
<tr>
<th>Library metaphor</th>
<th>General function name</th>
<th>Travel metaphor</th>
<th>General function name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-it</td>
<td>Annotate</td>
<td>Tour guide</td>
<td>Help</td>
</tr>
<tr>
<td>Previous book</td>
<td>Back</td>
<td>Make photograph</td>
<td>Copy</td>
</tr>
<tr>
<td>Next book</td>
<td>Forward</td>
<td>Home</td>
<td>First</td>
</tr>
<tr>
<td>Search shelves</td>
<td>Search again</td>
<td>Travel log</td>
<td>History</td>
</tr>
<tr>
<td>Archive</td>
<td>Save</td>
<td>Make reservation</td>
<td>Select</td>
</tr>
<tr>
<td>Borrow book</td>
<td>Print</td>
<td>Stop travelling</td>
<td>Stop</td>
</tr>
</tbody>
</table>
Table 4. Descriptive statistics for dependent and some background variables of Experiment 1 (composite metaphor interface only).

Note: Entries in each but last column represent mean, standard deviation, and number of observations (participants), respectively. Accuracy expressed as number (out of 8) of web-based tasks (questions) completed successfully. Time-on-task expressed as number of minutes needed to complete 9 web-based tasks. Declarative knowledge expressed as number of questions answered correctly (out of 9) on declarative knowledge test. Amount of Computer Experience: self-rated using one of five categories representing increasing amounts of experience. Age expressed in years.

<table>
<thead>
<tr>
<th>Age</th>
<th>Time-on-task</th>
<th>Accuracy</th>
<th>Declarative knowledge before exp.</th>
<th>Declarative knowledge after exp.</th>
<th>Amount of Computer Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.53</td>
<td>22.48</td>
<td>5.22</td>
<td>5.41</td>
<td>6.65</td>
<td>Modal and median rating category: 2-5 years of computer experience. (17 participants)</td>
</tr>
<tr>
<td>17.99</td>
<td>9.04</td>
<td>1.72</td>
<td>1.46</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>16</td>
<td>17</td>
<td>17</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. Interface function labels and corresponding general function names (Experiment 2), broken down by type of metaphor and type of condition.

<table>
<thead>
<tr>
<th>Component metaphor</th>
<th>Composite Metaphor Condition – Function Label</th>
<th>Control Condition - Function Label</th>
<th>General function name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library</td>
<td>Post-it Alpha</td>
<td></td>
<td>Annotate</td>
</tr>
<tr>
<td></td>
<td>Archive Sigma</td>
<td></td>
<td>Save</td>
</tr>
<tr>
<td></td>
<td>Help desk Kappa</td>
<td></td>
<td>Help</td>
</tr>
<tr>
<td></td>
<td>Bookmark Epsilon</td>
<td></td>
<td>Favourites</td>
</tr>
<tr>
<td>Travel</td>
<td>Make photograph Gamma</td>
<td></td>
<td>Copy</td>
</tr>
<tr>
<td></td>
<td>Show picture Delta</td>
<td></td>
<td>Paste</td>
</tr>
<tr>
<td></td>
<td>Travel to Labda</td>
<td></td>
<td>Search</td>
</tr>
<tr>
<td></td>
<td>Travel log Omega</td>
<td></td>
<td>History</td>
</tr>
</tbody>
</table>
Table 6. Descriptive statistics for dependent variables of Experiment 2, broken down by type of interface.
Note: Entries in each cell represent mean, standard deviation, and number of observations, respectively. Control condition uses neutral labels for denoting interface functions. Time-on-task expressed as number of seconds needed to complete all web-based tasks. Accuracy expressed as total number of web-based tasks (questions) completed correctly (out of 6). Declarative knowledge expressed as number of questions answered correctly (out of 8) on declarative knowledge test.

<table>
<thead>
<tr>
<th>Type of interface</th>
<th>Time-on-task</th>
<th>Accuracy</th>
<th>Declarative knowledge before exp.</th>
<th>Declarative knowledge after exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>758.00</td>
<td>4.59</td>
<td>3.00</td>
<td>4.36</td>
</tr>
<tr>
<td></td>
<td>263.48</td>
<td>1.43</td>
<td>1.10</td>
<td>1.86</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Composite metaphor</td>
<td>662.30</td>
<td>4.35</td>
<td>5.00</td>
<td>7.20</td>
</tr>
<tr>
<td></td>
<td>175.77</td>
<td>1.55</td>
<td>1.25</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
Table 7. Between- and within metaphor confusions on declarative knowledge test, broken down by learning phase and type of interface (Experiment 2).

Note: Entries in each cell represent mean, standard deviation, and number of observations, respectively. All means expressed as deviation from the number of confusions that could be expected on the basis of chance (guessing) in each category (positive values indicate more errors, negative values indicate fewer errors than expected on the basis of chance).

<table>
<thead>
<tr>
<th>Type of interface</th>
<th>Within metaphor, before exp.</th>
<th>Within metaphor, after exp.</th>
<th>Between metaphor, before exp.</th>
<th>Between metaphor, after exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>-0.20</td>
<td>-1.02</td>
<td>-0.61</td>
<td>-1.07</td>
</tr>
<tr>
<td></td>
<td>1.21</td>
<td>1.27</td>
<td>1.12</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Composite metaphor</td>
<td>-2.85</td>
<td>-3.05</td>
<td>0.05</td>
<td>-2.05</td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>1.25</td>
<td>0.82</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>