MASTER

Depth and breadth on the road
information hierarchies for mobile use

Geven, A.W.

Award date:
2005

DEPTH AND BREADTH ON THE ROAD: INFORMATION HIERARCHIES FOR MOBILE USE

A.W. Geven

Arjan W. Geven
December 2005

Final Thesis for the Technology and Society study, with specialization Human Technology Interaction and ICT

First Supervisor
prof. dr. D.G. Bouwhuis
Technische Universiteit Eindhoven
Department TM

Second Supervisor
dr. A.I. Cristea
Technische Universiteit Eindhoven
Department W&I

Company Mentors
prof. dr. M. Tscheligi
Mag. R. Sefelin
CURE – Center for Usability Research and Engineering, Vienna

Summary

The optimal way to structure information in hierarchies has occupied researchers for at least two decades. This study presents a literature overview of this research on the structuring of information followed by an experiment that deals with the special case of information architecture for mobile devices. Previous research focused on experimental testing of several structures on desktop computers to find out which structure performs best. Some of the existing literature also provides theoretical models for search times in such structures. The factors that are used in these models are user response time, computer response time, the time necessary to read the amount of items on a page before a decision is made, and the total amount of items per level and the number of levels. These models predict the average time a user needs to find an item on the deepest level of the hierarchy. In normal computer / internet scenarios, this means that the optimal amount of items lays around 7 to 13 items per level depending on the exact model that is used. Unfortunately, these models predict only search time, and not other factors that are important for the general usability of an information hierarchy.

Besides search time, errors are a frequently used measure for the quality of an information structure. Errors that are made during navigation are often a result of sub optimal labeling of the items in the hierarchy. A user may choose an item because it looks promising, just to find out that it contains something else than expected. Therefore it is important to choose consistent labels for the items. The second reason for errors is that users can get lost in a navigation structure, especially if it has many levels of navigation and users do not know where they are anymore. This pleads for information structures with only few levels with many items per level as well. If a user makes an error, than users will have to perform less backtracking in such a broad hierarchy then if they would be already on a deep level of a narrower hierarchy. Unfortunately, users' personal opinions were not reported in previous research on search time and errors. This results in very objective information on the one hand, but ultimately does not give any information about what the users themselves prefer, which is at least as important.

Recently, mobile devices added new challenges to the research on information structuring. The limited screen size, navigation methods and data transfer rates make the search for the optimal information structure even more complex, and alter the prerequisites for the optimal information hierarchy that were described above. Most importantly, long lists of items are harder to navigate through because of the limited screen size and the advantage of large breadth over large depth that was found in desktop research is reduced for mobile devices.

In this study an experiment was done to investigate the usability of four different information hierarchies. The four hierarchies ranged from very narrow (4^6, four items per levels, six levels deep) to very broad (64^2), with two hierarchies in between (8^4 and 16^3), were tested on three mobile devices with different properties. The goal of the experiment was to test how the demands on an information hierarchy change under the special circumstances that mobile devices introduce. Fifteen users were asked to test the hierarchies with the three devices. The hierarchies were (loosely) based on the Austrian yellow pages, which were adapted to fit exactly into the four hierarchies with varying depth and breadth. The participants were asked to look up a number of items on the deepest level of each hierarchy, where time, errors, satisfaction, number of key presses, certainty of choice and perceived complexity were

measured. The users returned on a second and third occasion to see whether the preferences of the user for a certain hierarchy would change with increasing expertise.

In terms of search time and error, no significant differences were found between the four hierarchies, although the performance on the different devices was significantly different: the device with the smallest screen performed the worst and the device with the largest screen performed the best. In terms of satisfaction however, it was clear that the participants of the experiment preferred the narrow hierarchies (4^6 and 8^4) to the broader hierarchies (16^3 and 64^2), which almost none of the participants preferred. The narrowest hierarchy was also perceived as less complex than it actually was, which was not the case for the other hierarchies; the broadest hierarchy was perceived as more complex than it actually was which speaks in favor of the narrowest hierarchy. Users also need the least key presses in this hierarchy to reach the target item, which increases rapidly as the hierarchy becomes broader.

Although the users gained more experience in using the four hierarchies while they used them on three occasions of one hour over three weeks, their preference did not change significantly. After three weeks, they still preferred the narrower hierarchies to the broader ones.

All in all, the optimal information hierarchy for mobile users would be one in which the amount of items per page would not cross the eight-item limit, and preferably would have between four and eight items per page.

Acknowledgements

The research described in this study was performed at the Center for Usability Research and Engineering (CURE), in Vienna. When I came to Vienna at the end of February 2005, I did not really have an idea what to expect from this graduation assignment. The country was new to me, the capital great, but very large. The language was strange, partly because of the interesting Austrian dialect! However, nine months later, the country is not new anymore, I have got to know its capital and I even learned to speak its language.

Where I had to speak English with my company mentor in the beginning, we switched to German after a while. Accordingly, I performed the experiment in German as well, with mostly Viennese students. Therefore, I would like to thank the fifteen students that took part in the experiment for their time and help. Of course, I owe a big thank you to all the people at CURE who assisted me during my stay in Vienna, and especially Reinhard Sefelin, my day-to-day mentor, who helped me with good discussions about the study from the beginning until the end, and Manfred Tscheligi, for offering me the chance to perform my research at CURE. Furthermore, I'd like to express my gratitude towards Don Bouwhuis and Alexandra Cristea – my first and second supervisors at the TU/e – for assisting me throughout the duration of this project and helping me in writing this final thesis.

Finally, I absolutely have to thank you, Nora. I've had a great time in Vienna, and it would not have been possible without you.

Isn't it great to study abroad?

Thanks everybody.

Arjan Geven

Table of Contents

Table of Figures and Tables

FIGURES

TABLES

1 Introduction

On the desktop, large screens and multiple input possibilities like keyboard and mouse assist the user in his or her information search. When we turn to mobile devices, we do not have this wealth of supporting technology. Although the screen size of wireless devices has grown over time and input devices have certainly improved, they will never be the same as their desktop counterparts; bulkiness and mobility just do not go very well together. A compromise has to be found between this increasing mobility and a reduction in usability. The information structure has to be adapted to compensate for the new context in which it is used. The question is: exactly how should it be adapted? We have two decades of research for optimal forms of hierarchies, but hardly any of these mention the size of the display. In this report, these previous findings are related to the new context of mobile use.

Mobile devices give users the technology to have access to the Internet from anywhere. Previous research shows that users have problems using such a device though (Buchanan *et al.*, 2001). To reach a usable solution, it is important to investigate the complex interaction between the user and the technology. This study presents a relevant contribution to the field, by searching for ways to optimally structure information for mobile users, a field that has so far hardly been investigated.

1.1 GOAL

CURE is a research institute for usability research in general with a special focus on the usability of mobile devices. This goal of this study is to deepen the existing the knowledge on how web pages and applications can be structured for the specific context of mobile use. This is done by searching the existing body of literature for general solutions of the problem of structuring information, and adding information to this existing knowledge in the form of an experiment in which the formulated hypotheses are tested.

1.2 MAIN QUESTION

In 2003, CURE performed a small experiment in this area, where users had to navigate through a number of different structures to find out which one of these structures showed the best fit to the mobile context of use. It turned out that users have different preferences for information structures when using devices with different properties (Giller *et al.* 2003). The efficiency in using the structures also differed. Because this was a relatively limited experiment with few participants, the results were also limited. Now, CURE wanted to perform a study on a larger scale in which the influence of information architecture is investigated further. This leads to the main question of this study as formulated below.

The main question that is covered in this report is: “How can the usability of an information structure of an application or website be optimized for the special situation on mobile devices?”

The question formulated above has been refined below to allow for more fine-grained analysis and is therefore split into four sub questions:

1. How can an information structure help mobile users with information navigation?

2. How do the specific properties of mobile devices (social and physical context, transfer speeds) influence the usability of an information architecture?
3. How do screen size and input method influence users' preference for a certain information architecture?
4. What is the influence of increasing expertise on this preference?

Not all aspects of information navigation can be included in a laboratory study, because such a study is by nature very limited. These aspects (questions 1 and 2) are therefore presented solely in the literature review. This review of literature is meant to position the experiment in the broader perspective of information retrieval.

Since there is hardly any literature that covers the topics of questions 3 and 4, these questions are the subject of a controlled laboratory study, as presented in the chapters after the literature review.

2 Literature Review

In this chapter, a review of related literature on information navigation in the mobile context is presented. The global layout of the usability problem is presented, with the three most relevant aspects: the user and his context, the system, and the interface between the user and the system, as also displayed in Figure 1 below.

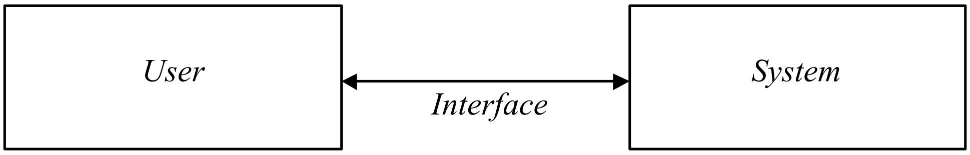
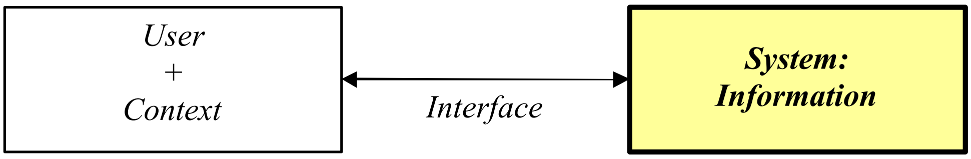


Figure 1: Model for User System Interaction

The goal of the user is to gain information. In this case, this information is contained in the system (the internet), which is described first. The focus is then moved to the interface that the user needs to reach the system. The system is a given property, so that the interface is the part where usability can be gained. The interface is the part where the information that is contained in the system is translated to a form that the user understands. Here, decisions are made how to present the information to the user, and the user can access the information using the interface. How can information be presented via this interface in a way that makes sense to the user, especially in his or her personal (mobile) context? Finally, the user also has a role in accessing the information. The user has a certain expertise, might be using a specific device, in a special situation. These are user issues that have to be taken into account when designing for the user.

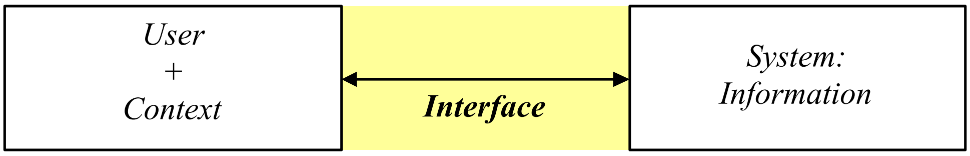
2.1 THE SYSTEM



In this case, we use the internet as ‘the system’. The internet is full of (relatively) unstructured information in the form of websites about all kinds of subjects, and this (lack of) structure is where this report is about. The information on the internet is generally loosely organized. To give users some way to access the information, they are offered navigational tools such as menus, hyperlinks and search engines. Without these means of navigation, a user will not find the information and the information is wasted. Therefore, such structuring of information is very important. By now, search engines are used to index the internet and make information more accessible by means of keyword search. Google for example claims to have indexed over eight billion pages last year (Searchenginewatch, 2004). So, information is available. But it is not always and equally accessible. The system is only a very small part in this review. The information is a ‘given’ that can be accessed by the user by means of an interface, but no more than that. That is why the attention to this part of the system is limited.

To be able to access this wealth of information, we need an interface. This interface can be a search engine; it can also be the provision of structure to information.

2.2 The Interface



To aid the user in her search for information, information should be structured. This structuring of information is an important factor in the usability of a design, since users get lost easily in websites. This structure of information is seen as the interface between the content (the information) and the user. A good architecture provides users with ways to find the information they are looking for easily, and prevents them from getting lost. This is mainly done by organizing the information in a logical way for the user to understand, choosing the right terms for labels, and supplying a navigation system (Rosenfeld & Morville, 2002). In this chapter, the literature is reviewed that deals specifically with the ordering of information in hierarchies.

To provide structure to information, professionals can develop organization schemes and taxonomies where items are grouped together by topic, task, audience, or sorted alphabetically, chronologically, geographically, or a combination of these. When the amount of items is large, items are usually further structured into navigation menus or hierarchies in order for users to find items efficiently. This is a very precise task, since users have very different ideas of what to find behind a certain label, not to mention the fact that there are usually dozens of keywords that can describe a certain concept, keywords of which only one can be chosen to represent the concept in the form of a label. Chances that the word chosen actually matches the keyword in the head of the user is minimal (keyword overlap was found to be approximately 18% in general domains, Rosenfeld & Morville, 2002). This labeling of keywords is a domain of its own, one that has occupied e.g. librarians for decades, and is too much off-topic to discuss here thoroughly.

Of particular interest in this study is how information can be organized optimally in the form of hierarchies: navigation hierarchies. Imagine a tree-like structure, where one node leads to multiple child-nodes. These child-nodes have their respective child-nodes, and so on. At the end of this tree is the information, organized in the leaves. This basically is a navigation hierarchy. An example is the organization in folders and subfolders in the Windows File System, where one can expand each folder, e.g. C:, to see the contents of that node, e.g. Program Files, Windows, My Documents. Each of these folders contains files: the nodes with information. The hierarchical file system is a way of organizing the information in the computer. It helps users to find the information they need by providing a logical structure. This is a form of a navigation hierarchy.

These navigation hierarchies can be narrow, with just a few options on each level, or broad with many options per levels and fewer levels (Figure 2). The number of option per level is referred to as ‘breadth’ and the number of levels in a hierarchy is referred to as ‘depth’. The balance of depth and breadth of the hierarchy is particularly important for the usability of an information structure (e.g. Snowberry, Parkinson & Sisson, 1983). In a narrow hierarchy, a user will have to navigate through many levels before reaching the destination node. Along the way, the user will have to make a lot of decisions with the possibility of getting lost, not even counting the time cost of waiting for each consecutive page to load. In a broad hierarchy, a user has more options at his disposal and thus less navigational load, but this also comes at the cost of possible information overload. Several researchers have investigated the optimal balance of items in hierarchies and menus in the last two decades. These are discussed more or less chronologically.

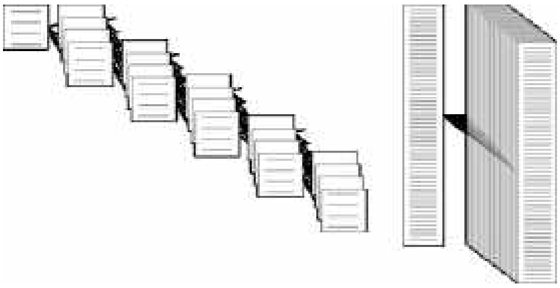


Figure 2: a deep hierarchy (4 items per level, 6 levels deep) and a broad hierarchy (64 items per level, two levels deep)

2.2.1 EXPERIMENTING WITH VARIOUS DEPTHS AND BREADTHS

In 1981, Miller found that when depth increased, so did response time to select a desired item. In his experiment, participants had to find words in four different hierarchies, each with a total of 64 items (2^6, 4^3, 8^2 and 64^1). The two conditions in the middle were the optimal ones in terms of efficiency, showing a clear tradeoff between depth and breadth. Miller's conclusion is then that a low-depth hierarchy with items on a few levels is better than a narrow hierarchy. He argued that short-term memory was a limiting factor, since it is difficult to remember exactly in what branches certain items are located. In this experiment, the items were randomly ordered, which means that scanning all options was needed to find the target item.

Snowberry *et al.* (1983) performed a similar experiment with the same four conditions, but additionally ordered the items under subheadings in the broadest condition. It was then found that the optimal condition was the broadest and ordered one, and search time and accuracy both degraded as depth increased. Snowberry *et al.* changed the experiment quite significantly by ordering items on the first level into categories. They practically created a single level that has the advantages of two levels, where users do not have to select a heading before seeing its underlying items. This means that search time does not increase dramatically because users do not have to read all items displayed, and the number of errors is reduced because users have more overview of what is happening. It is quite interesting to see what the difference between the ordered and unordered 64-item condition in terms of required navigation time are. This is depicted in Figure 3. Also, error rates dropped significantly from the narrowest condition to the broadest, ordered, condition, from 34% in the narrowest condition to 4% in the single level condition.

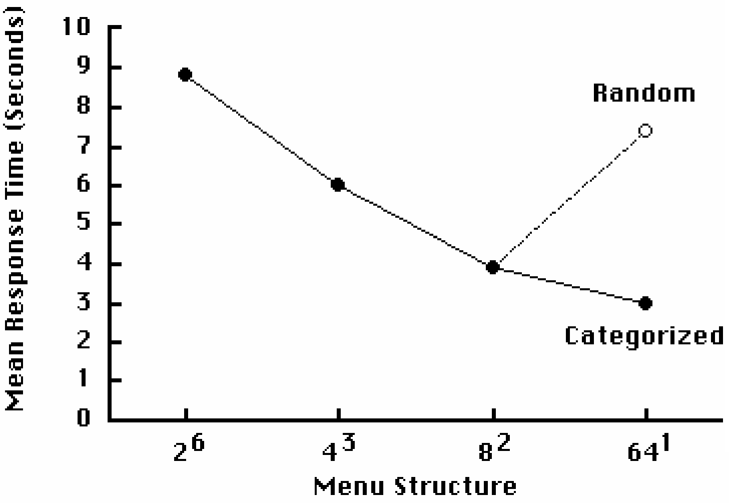


Figure 3: Response times per structure by Snowberry, Parkinson and Sisson (1983), adapted from Norman (1991)

Kiger (1984) also investigated optimal hierarchy design, but with differently shaped hierarchies. He used five different categories; the broadest categories being two levels deep and laid out as 8x8, 4x16 and 16x4. The other two hierarchies were 2^6 and 4^3. Kiger found that both search time and number of errors were lower in the two-level hierarchies than in the narrower hierarchies. More specifically, the best results were obtained when a (relatively broad) 16-item choice was followed by a 4-item choice. Subjective ratings of participants also showed that they favored hierarchies with the least depth as well. The most favorable according to the participants was the 8x8 hierarchy.

Landauer and Nachbar (1985) used four different hierarchies to display 4096 items. The hierarchies were 2^{12}, 4^6, 8^4 and 16^3. The items were either integers from 1 to 4096 or alphabetically ordered words. On each level of the hierarchy, a more precise range was given in which the target item was to be found. The distinction between categories was very clear because of this special labeling and errors were not permitted by the system (e.g. the system would refuse an erroneous selection and wait until the right option was chosen). This made the task somewhat different from the other experiments described. The time to reach the target item in the experiment is predicted to be shorter in the broad condition because users have to make a decision for every selection that has to be made. As the hierarchy gets broader, fewer selections have to be made, resulting in shorter times. This turned out to be true in the experiment, where finding the target item in the narrowest hierarchy took twice as long as in the broadest hierarchy. The main limitation of this experiment can be found in the very clear, non-overlapping categories. Users can skim quickly over items and the time needed to scan all items is thus rather low. When the processing of each option takes longer, a different trade-off might be shown.

Tullis (1985) seconds the claim that a broad hierarchy would be better than a narrow hierarchy. Based on an experiment, he argued that it is difficult for users to predict which items are located in which categories, especially in a narrow hierarchy. In a broad hierarchy, users know better which items would fall where. Additionally, if a user makes an error at the beginning of a hierarchy, the results can be more disastrous in a narrow hierarchy, since the user will have to perform more backtracking to correct the error. The aim of Tullis' study was to find out if either a broad or a narrow hierarchy would be better to organize functions in an operating system. The participants used either a narrow hierarchy that consisted of a list of 15 items and a deeper underlying structure (up to four levels deep); or a broad hierarchy that consisted of 45 items (two levels deep). The time it took users to perform tasks did not differ significantly between the two conditions, but users had to take fewer steps and made fewer errors in the broad hierarchy. Tullis' conclusion was that a broad hierarchy is better, while it makes it easier for users to predict which path they have to take.

Wallace, Anderson and Schneiderman (1987) also looked at two different menu structures (three levels and six levels). It was demonstrated that the broad hierarchy was faster and resulted in fewer errors than the narrow hierarchy. Zaphiris (2000) replicated the study by Kiger using a browser interface, and also found that broader hierarchies are faster and appreciated more than narrow hierarchies. The access times of items at the end of the hierarchy are proportional to the depth of the tree structure, which means broad hierarchies are preferred. Jacko and Salvendy (1996) found that with an increase of depth, subjects' perceived complexity of a menu also increased. This as well favors broad hierarchies over narrow ones. Overall, broad hierarchies reduce memory load on the user, require less navigation, which results in less disorientation.

Although most researchers found that broad hierarchies were easier for users to navigate than narrow menus, Landauer and Nachbar (1985) argued that it might be possible that narrow hierarchies are favored more in special cases where categorical decisions about each alternative is independent and slow. Paap & Roske-Hofstrand (1986) elaborated on this claim, and argue that narrow hierarchies do have the advantage that users are better protected against choosing unlikely responses, especially when categories are complex or ambiguous. They call this advantage "funneling", because a user can zoom in on a certain aspect, while avoiding similar choices in other parts of the hierarchy. Some evidence exists on this subject, for example, Dray *et al.* (1981) found that novice users navigated faster using a narrow hierarchy compared to a broad hierarchy. Khan and Locatis (1998) also investigated the effects of information presentation. They examined the effect of link density and presentation format as on users' search efficiency and accuracy. Link density was defined as number of links per page, and link presentation varied between links presented in a list and links presented inside a paragraph of text. They found that fewer items per page (low link density) presented in a list have the best effects mainly in terms of search accuracy and search task prioritization. These data indicated that fewer links per display and representing links in a list instead of embedding the links inside a paragraph of text reduced cognitive load and reduced information-processing demands on users.

2.2.2 PREDICTING OPTIMAL BREADTH / DEPTH

Another way to estimate the optimal breadth and depth of a hierarchy is by using a mathematical model. Such models can never predict a situation 100% accurately, but they can give an educated guess of the number of alternatives per level that is reasonable and can be evaluated in validation research.

The linear model

Lee and MacGregor (1985) presented a simple linear quantitative model for predicting the search time that users need to find items in a hierarchy with a certain breadth and depth. The formula for this prediction takes into account the time needed to read options, the time needed to click on an option and the time a computer needs to react on this click. This results in the following formula:

$$T = d * (b*t + k + c) \quad (1)$$

With d is depth, b is breadth, t is processing time per option, k is the “key press time” and c is the computer response time. For this model two important assumptions have to be made. The first is that human and computer response time are equal across all levels of the hierarchy. Variations in this behavior cannot be taken into account, so these response times are assumed to be invariable. The second assumption is that users read all items in the list before they select the option of their choice. In other words, it is assumed that users will search through the items in a form of *exhaustive search*: all items are processed, and are processed exactly once. Whether users really use an exhaustive search during the examination of a level can be doubted, it is more likely that they will terminate the search as soon as they found a good candidate option and select it without evaluating all options. Because not all options will also be evaluated in such a *self-terminating search*, this will result in faster selection and can also be modeled. A self-terminating search ends when the user finds an item of interest. The idea of two types of search strategy was developed by Sternberg (1969), who found that users used two types of searching through short-term memory: exhaustive search and self-terminating search. On average, this is the option halfway down the list. This means that the formula has the factor b in the formula has to be replaced by (b+1)/2. The formula then looks like this:

$$T = d * (((b+1)/2)*t + k + c) \quad (2)$$

Although there are now two formulas, one for exhaustive search and one for self-terminating search, there are a number of other options, like redundant search, where some options are evaluated more than once, or a combination of these types of search. Therefore, a more general way of expressing the formula would be:

$$T = d * (E(n)*t + k + c) \quad (3)$$

This is the most general form of the formula, where the type of search is not directly included and can be any function E(n). If an exhaustive search is expected, E(n) = b, if a self-terminating search is expected, E(n)= (b+1)/2. When the user can restrict the scope of search because of ordering or more experience with a certain list of items, the formula can also be adapted to this situation, by replacing the number 2 with a general scope-factor f (Paap & Roske-Hofstrand, 1986). This leads to the function E(n) = (b+1)/f, where f is “an index of the degree to which the scope of the search can be restricted”. A higher restriction in the scope means that more items are skipped, and thus the search can be completed faster without reading all options. In a normal self-terminating search, approximately half of all items will be read. When users gain more experience with the system, and might not need to look at one third of the items: the value for f for this user would be 3. The optimal number of items per level for this user would be higher than for a user with an f-value of 2. If f equals 1, it is practically an exhaustive search. Is f smaller than 1, then items are reviewed more than once and it is a redundant search.

Since depth and breadth are exponentially related with each other ($N = b^d$), the factor d can be replaced in the formula by (log N / log b). This leads to the formula:

$$T = ((\log N) / (\log b)) *(E(n)*t + k + c) \quad (4)$$

If values for k and c are estimated, an optimal b can be found. Norman (1991) shows a graph of average search times for different breadths, for users with different reading speeds (see Figure 4). The values that were chosen for k and c are 1.0s and 0.5s, respectively. For users with a higher speed of reading the items, the theoretical optimal number of items per level shifts a bit towards higher numbers, but generally it can be said that the optimum lies at 4 to 7 items per level. However, these numbers might be overly restrictive. As mentioned before, the average amount of options that are evaluated might be less than considered here, since users are able to selectively skip items very well because they are ordered. Since providing no order at all does not have any advantage, items are usually ordered either alphabetically or semantically. Both ways of ordering increase the value of f for users and shift the optimal number of items to the right. Optimal breadth increases from 13 to 37 as f increases from 2 to 10 (Paap and Roske-Hofstrand, 1986).

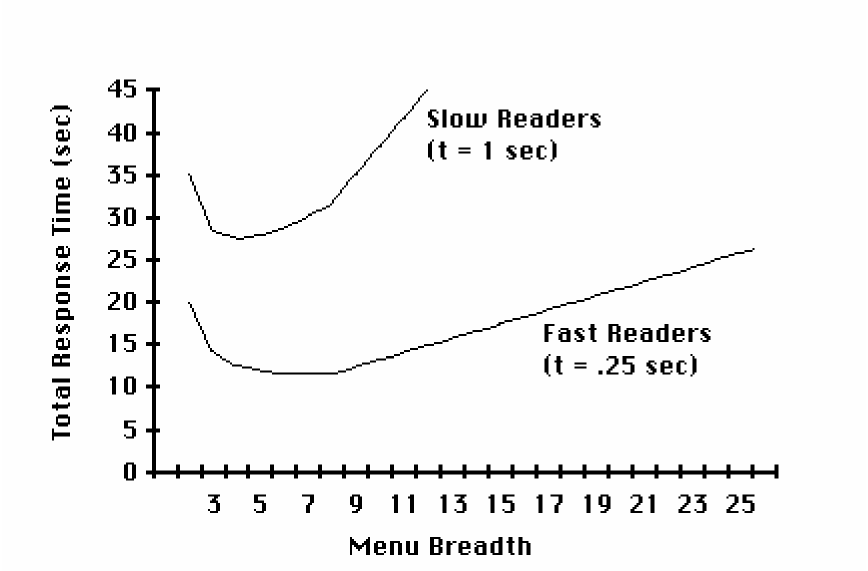


Figure 4: Total search time as a function of the number of options on one level (Norman, 1991).

Log-linear model

The above discussion focuses on a linear model for search times. But what if these times are not linear? Norman argues that the decision time and key-press time are not constant. Instead, they also depend on the number of alternatives, and he suggests a log-model based on the experiment from Landauer and Nachbar (1985). The log-model takes into account that the ‘user-component’ is larger when there are more options to consider, based on the Hick-Hyman-Law (Hyman, 1953): decision time $t_d = c + k*\log_2(n_i)$, where n_i is the number of (equally likely) options at level i. Combined with Fitts’ law that says that a user needs more time to hit a smaller target or a target that is further away compared to a bigger one or one that is closer by: movement time $t_m = c + k*\log_2(d/w)$ (Fitts, 1954). Since the distance to a target is constant, and the size (w) of the target varied according to 1/n in Landauer and Nachbar’s experiment, they could replace the factor $\log_2(d/w)$ with the factor $\log_2(n)$. This leads them to the following formula for predicting user the total selection time on one level:

$$T_{\text{one level}} = t_d + t_m = 2(c + k*\log_2(n)) \quad (5)$$

This is the formula for selecting an item on one level. The number of items on a level is equal to $\log_n(N)$, where n is the number of items per level and N the total number of items. The total search time is then:

$$T = \log_n(N)^* \, 2(c + k^*\log_2(n)), \qquad (6)$$

which can be rewritten as

$$T = 2(k^*\log_2(N) + c^*\log_n(N)) \qquad (7)$$

This matched the results from the experiment of Landauer and Nachbar very well, where indeed a logarithmic decrease was found when the number of options per page increased. So, with more options on one page, T decreased. Unfortunately their experiment is somewhat limited. In the 2^{12} condition each target took up half the screen, in the 46 condition each target took up one quarter of the screen and so on. As described above, they used varying hierarchies to show ranges of integers or words and asked users to search the target item as fast as possible. This combination makes it likely that decision time and movement time in their particular situation can be estimated quite well with the laws of Hick-Hymann and Fitts, but it could be hard to apply this to the real world. Not many menu systems are really designed for such visual search. Instead, search consists of a combination of abstraction power and educated guessing (in which category will my target item be located?). This severely limits the applicability of the experiment by Landauer and Nachbar, although their theoretical notions are very interesting.

Criterion-based search model

Above, the two models were described (formulas 4 and 7) that predict user performance when they search for an item in a hierarchy. Both are based on a combination of user and computer reaction time, and a expression of how many items are evaluated before one is selected. This expression for number of evaluations could be logarithmic or linear, in the form of a self-terminating or exhaustive search (or redundant even). Which one of these functions for E(n) is ‘the right one’ is difficult (or even impossible) to say. Paap and Cooke (1997) describe an additional model based on MacGregor, Lee and Lam’s investigation (1986) on how users search and select options in a menu: a combination of search techniques called the criterion-based decision model, which has been verified in further experiments by Pierce, Parkinson and Sisson (1992) and Pierce, Sisson and Parkinson (1992). The criterion-based decision model incorporates as the name implies two criteria, which are implicitly set by every user. When reading a list of options, each item is evaluated using these two criteria. The first is the low or L criterion, which is used to decide whether an option is worth it to be even taken into account in the judgment. For example, if a user is looking for a new sink for the bathroom, he or she may decide that the category “electronics” will not contain his target item, so this category is immediately discarded based on the L-criterion. It does not occupy a spot in his short-term memory anymore, because it is of no interest. As the user reads on, he or she might come across a category called “furniture”. This category could contain the desired sink for the bathroom, but the user is not sure yet. This means that the L-criterion has been used to decide to keep the item in short-term memory. After the L-criterion has been ‘passed’, the high or H criterion is consulted. This criterion is used to decide whether the current option is the top pick in the list. If the user thinks that the current category most definitely will contain the target item, he or she will not look further down the list but select the current category right away. No further evaluations are made; the current category is the one. Say that the category "furniture" is a candidate category for the user, but did not make it past the H-criterion. It means that although the option is likely to contain the item, it is not likely enough for the user to just go for it. Instead, it is saved in memory and more options are evaluated. When the user later finds a category that does pass the H-criterion for selection (e.g. "bathroom"), the "furniture" category will be discarded and the "bathroom" category will be chosen. If the user does not find a candidate that passes the H-criterion, only then will the items that did pass the L-criterion but were not that good that they passed the H-criterion (like "furniture") be evaluated again. The user will then select the candidate that is most likely for him or her to contain the target item, "sink". For a graphical depiction of this decision process, see Figure 5 below.

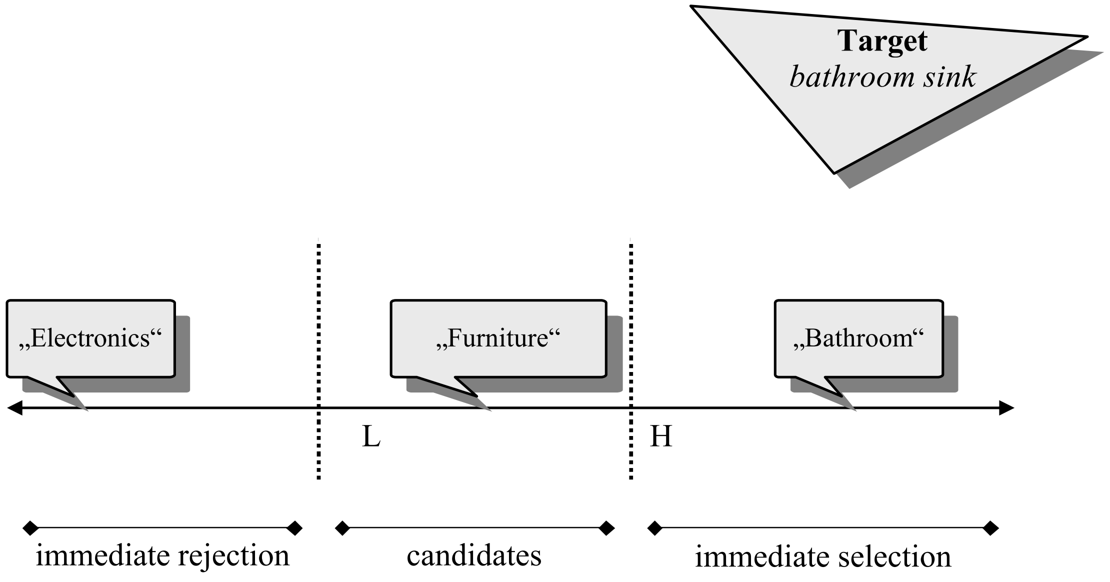


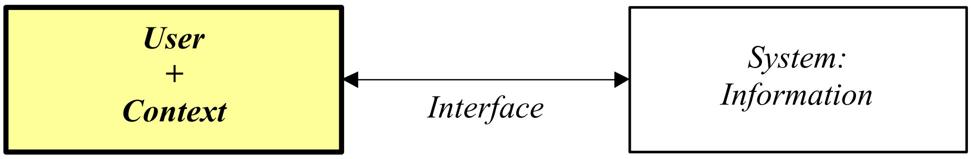
Figure 5: Selecting a fitting category for "bathroom sink" under the criterion-based decision model with a Low and a High criterion. Not all category labels pass the H criterion for immediate selection.

The criterion-based decision model gives a more detailed view of the decision process that guides a user into choosing the right option and predicts how many options the user will evaluate before selecting one. The values of the L- and H-criteria are adjusted by the user based on properties of the system like the number of options on one level. The L-criterion is always set higher than the a-priori chances (simple guessing chances). In a list with 4 categories, this means that a category only passes the L-criterion when the user-estimated chance that the target item can be found in this category is higher than 25%. Each item has an a priori chance of .25, and the L-criterion might lie e.g. at 0.40, depending on the knowledge of the hierarchy and the specific topic. The H-criterion for immediate selection always lies higher than the L-criterion. In the case of 4 options, it may lie even so high that the chances are high that all 4 items are read before a selection is made: an exhaustive search is performed. In a hierarchy with 16 options, H is lowered to keep the amount of reading effort low, but L is actually made stricter to keep the number of candidates at a reasonable level. The two criteria draw nearer to each other, and the effect is that the search approaches a self-terminating search. The first likely candidate is selected without evaluating the rest of the options. Unfortunately, the fact that the criteria shift with different numbers of options already shows that there are no fixed values for the two criteria. This makes it very difficult to estimate (not to say impossible) to estimate values for L and H. Additionally, the wording of the category labels, the level at which the user is currently working, personal aspects of users and their contexts and so on may all influence these criteria. If the user has more knowledge of the system, this increases the values for L and H to higher values (the user is quite sure of what he or she is looking for). And what happens if the user makes an error? In the models discussed before, effects of errors are completely left out; they are not part of those models. In the criterion-based decision model, it is possible to include some of these effects. If the user made an error, H might be set higher to avoid accepting an option too fast. When the first choice was wrong, L might be set lower to include more alternatives. However, real mathematical incorporation of such effects has not been tried.

The criterion-based model seems to be the most realistic theoretical model of the three models that have been discussed here (the linear model, the log linear model and the criterion-based model). It gives a good reason *why* users do not evaluate all options and is able whether users will use a certain search strategy, which makes it a very interesting and promising model for predicting search. Unfortunately, it does not help directly in predicting the total search time, because a formal version of this model does not exist. Its value therefore lies in that it describes the thought process, even if this is not formalized in the form of a formula.

So far, all models trying to estimate E(n) have been similarly limited, which makes it practically impossible to determine the number of options evaluated exactly. In the current models, only few of named factors have been taken into account.

2.3 The User



The third part of the model as presented above is the user, including the user’s context. The user is the person that wants to make use of the system; the person that wants to access the information that is contained in the system. How users deal with information systems depends on the goals they set for themselves and the tasks they want to perform. This is described in this part of the literature review. Besides these tasks and goals, other factors influence the way the user can and will achieve these goals. This is the context of the user, consisting of the user’s identity, the technology available and the location where the user wants to achieve his or her goals.

Several more or less comparable models exist of how users behave based in tasks they perform and the goals they want to achieve. Users generally show information behavior that fits in one of the four categories with different levels of directedness: passive attention, passive search, active search and ongoing search (Wilson, 1997). Choo, Detlor and Turnbull (2000) relate information seeking to modes of organizational scanning, where also four distinct modes can be distinguished: undirected viewing, conditioned viewing, informal search and formal search. Undirected viewing can be thought of as general browsing, or page flipping, much like zapping on TV. In conditioned viewing, the general browsing has narrowed down to some selected topics. During informal search, a user looks in an unstructured way for information regarding a specific topic, and in formal search the user makes a planned effort to obtain specific information. Rosenfeld and Morville (2002) differentiate between three groups of users in a fishing metaphor, where users are looking for 'the perfect catch' (a single right answer), 'lobster trapping' (information about a subject where it is not obvious whether there is single right answer) or 'indiscriminate drift netting' (exhaustive search to find as much information as possible).

Catledge and Pitkow (1995) divide people over three types of browsers: the serendipitous browser, the general-purpose browser and the searcher. The 'serendipitous browser' is the most undirected browser, almost randomly accessing pages if it looks somehow interesting. The 'searcher' is the most directed browser, navigating through long paths to find specific information to solve a certain problem. The general-purpose browser is somewhere in between the former two. Catledge and Pitkow came to these three types by investigating how people browse the internet using a modified version of Xmosaic that recorded the users' browsing activities. Users were classified based on these characteristics. Websites should be tailored to each of these three kinds of browsers to support their specific way of looking for information. Basically, this means that specific information should be available by means of a search or a navigational hierarchy for more directed searching by 'searchers' who knows what they are looking for, but also that general information should be available for the 'serendipitous browser' who is not directly searching for detailed information to solve a problem. This can be linked to the models of search that were described in chapter 2.2.2. A user who knows exactly what he or she is looking for will find the information required by means of a kind of systematic search. The chances of finding the target are 100% (given that the information exists). A user who does not know his or her target will look randomly, the chance that the target is found will increase with time, but it does not ever fully reach 100%. Therefore, a randomly browsing or serendipitous user will be aided by additional information and not directly by search queries or hierarchical structures.

These are three comparable notions of different user tasks and goals. Not all users are just browsing the internet and might stumble upon something interesting, but also not everybody has a very clear idea of what they are looking for. All of the above mentioned researchers indicated three or four different goals and associated tasks. For each of these tasks, the demands on the interface are also different. Although the system stays the same, users with different goals are best supported with different navigational means. For example the serendipitous browsers from Catledge and Pitkow (showing passive attention according to Wilson or undirected viewing behavior according to Choo, Detlor and Turnbull) do not need a "search" field. They are not looking for something in particular, but want general information instead. When they read a text, they can be aided by links to related information on a subject to continue their serendipitous browsing activity. These browsers might also be interested most in 'entertaining' websites that bring extra pleasure to the browsing activity. On the other hand, directed searchers (showing active search or formal search behavior) know exactly what they are looking for and do not need as much related information; instead they usually look for a search function on a website or use a general search engine to find what they are looking for. The other types of browsing behavior are somewhere in between these extremes; each with their own set of goals and respective requirements.

Information navigation is not a one-way process. Saracevic argues that there is an interaction between the user and the information retrieval system on more than one level (Saracevic, 1996). He presented a basic model of the interaction process in information retrieval where interaction occurs in several levels (or strata), like the cognitive and the situational level, as well as on the query processing and the information input level. Information retrieval is basically an interactive process, during which the knowledge of the user changes which in turn changes the search tactics of the user and his criteria of assessing relevance (Vakkari, 1999).

Pirolli and Card (1995) suggested that users search for information on the internet much like animals in the wild who search for food. Like the foraging of these animals, they coined the information search 'information foraging'. Animals follow the scent of their prey to find it, users follow the scent of links to find information, the 'information scent'. A link gives a user a clue about what lies behind it. If this scent is strong enough, a user will follow the link. If the scent is not strong enough, the user will go back or try another link. Navigation becomes then an expected-cost/expected-rewards trade-off. Pirolli and Card also note that the cost/rewards trade-off is influenced by the situation of the user. The careful labeling of items is very important so that users go through with the navigation structure until they reached the target item, and also still expect the target to be where it actually is. As Rosenfeld and Morville (2002) note, labeling is very important because it is the only means of communication that is possible between the website or application and the user. If users do not understand the labels, they will not have any other way to access the information, and will thus not reach his goal. Rosenfeld and Morville later state "designing effective labels is perhaps the most difficult aspect of information architecture" (p. 92), because language is very ambiguous with all its synonyms and homonyms and labels are never perfect. Additionally, every user interprets labels differently, so that a designer can never be sure that the user understands what was meant. They also provide some guidelines for label design: narrowing scope whenever possible, develop consistent labeling systems and not just labels, and be consistent in style, presentation, syntax, granularity, comprehensiveness and audience. This all increases the information scent of the website or application, so that the expected rewards of the target item will be much higher than the expected rewards of any other item in the hierarchy, and helps reduce ambiguity between the items, making it easier for the user to find what he or she is looking for.

2.4 THE USER'S CONTEXT

For desktop computing, the notion of context is not that important, as the context in which it is used is fairly constant. For mobile devices, context is continually changing.

Context can be used to aid the interaction between the user and the device. Before this can be done, however, it is important to understand what context is and exactly how context can be used. Many researchers have defined context differently. Some use the word to describe the personal characteristics or state of the person who uses the device (e.g. experience, emotional state), others to the situation, the location, or the cultural setting in which the device is used. Dey and Abowd (2000) give the following definition of context:

> "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."

In the literature on context, the term 'context' is often used to describe only the *location* of the user. The above definition is wider and allows for a more general notion of the term, including many situational aspects besides location that can characterize the situation of the user. This situation of the user can be subdivided into a number of categories that each have specific influences on how the user interacts with the system.

First of all, there is the user *identity*, or who is using the system. Each user has specific demands to the system and has a specific way of dealing with an information system. Each user has a mental model of the system, and these models are generally incomplete and contain errors. Additionally, each user has specific cognitive, affective and physical features that can be taken into account. Already mentioned before is the *location* of the user. This is currently a hot topic in usability research, since more and more devices are, or will be equipped with some form of location awareness. The presence of GPS, RFID and similar technologies enable devices to sense where they are, and with that, also where 'their' user is. This allows adapting content to needs of users related to where they currently are. These needs that are related to a certain location can also be related to certain user *activities*. If a user is driving a car through a city, he or she needs to visually focus on the road and navigational software should not demand too much visual attention from the driver. Finally, the *technology that is available* to the user for performing the task influences the way he or she can deal with the task. A slow appliance might not be the right means to communicate real-time high-bandwidth information to the user, but if it is the only device currently available, some adjustments might be made in the way this information is presented in such a way that the user can more effectively deal with the information.

The five dimensions described above, *identity*, *location and activity*, *time* and *technology*, together define the user's context. This description of context is definitely wider than the dimension 'location' alone that is often used. The dimension *activity* is embedded in the location and is therefore not discussed separately; the same goes for the dimension time. The main three dimensions are discussed in more detail in the following paragraphs 2.4.1 about technology, 2.4.2 about the user's identity and 2.4.3 about the location.

2.4.1 CONTEXT DIMENSION: TECHNOLOGY

Information tasks are complicated, including navigating through information hierarchies. As shown above, users do not always perform well on such tasks, depending on a large number of factors. An additional factor in the user's context is the use of a certain technology or appliance, especially mobile devices, since users are able to take these devices to different contextual settings, where desktop computers are fixed to one specific situation: on the desk. Although mobile devices are portable, mobile phones and PDAs are relatively miniaturized in their hardware for the sake of portability, compared to their desktop counterparts. The display is very small, input is at least complicated, storage is limited, battery life is short, and data transfer rates are low (Nielsen, 2000; Buchanan *et al.*, 2001). Although the screen size of wireless devices has grown over time and input devices have certainly improved, they will never be the same as their desktop counterparts; size and mobility just do not go very well together. Although these devices have the great advantage that they are portable, the physical properties severely limit the usability, resulting in sub optimal performance by its users.

Screen size of mobile devices

Screen size is the most obvious limitation in the context of mobile devices. Because the screen size of mobile devices is small, users have to hold more information in their short term memory which can hamper information processing, for example when users want to compare information between pages (Albers & Kim, 2000).

Experimental studies have confirmed that screen size has an adverse effect on usability. For example, users read slower on small screens than they do on large screens. In an experiment by Duchnicky and Kolers (1983), it was found that users read text up to 25% slower when reading it on small-width display with a width of 1/3 of the control display with regular desktop width. Display height did not have such a negative effect. Swierenga (1990) conducted a study on the effects of display size to reading comprehension. She found that although performance is not equal among conditions, real problems only occurred when only one line of text was shown. Resiel and Shneiderman did find such an effect of display height (1987). In an experiment, they found that text reading on a 22-line display was up to 15% slower than on a 60-line display (Resiel and Shneiderman, 1987). To compare, a contemporary mobile phone typically can display around five to ten lines, a PDA fifteen to twenty. The effect seems to be stronger when longer texts have to be read.

Not only reading performance degrades with the use of a small screen display; information retrieval also suffers from the use of small screens. Jones *et al.* (1999) performed an experiment where users had to find information on the internet using two types of screens. They report that users of the small screen answered half as many questions correctly as the large screen group. Moreover, 80% of users of small screens indicated that they felt screen size impacted on their ability to complete the tasks, compared to 40% for large screen users. The amount of hyperlinks clicked while performing the tasks was also recorded, but this did not vary significantly between the two groups. Jones *et al.* expected that small screen users would initiate additional browsing efforts to see what a website is about, or “rapidly, randomly ‘hunting’ around the site”, but this behavior did not occur. They did find that path lengths for small screen users were shorter. These users turned faster to the search option in a website, because browsing by clicking links took too much effort. Small screen users also scrolled much more in order find the information they needed compared to the large screen users. Jones *et al.* expected that users would scroll forwards and backwards upon entering a website to gain an overview of the site, but this was not the case.

Jones, Buchanan and Thimbleby (2002) found that screen size has a serious adverse effect on information search and retrieval. Users were asked to complete tasks using a small (WAP), medium (PDA) and large (desktop) sized screen. The task they were given was to search certain pieces of information using adapted versions of the Google-website. The WAP-interface performed poorest, with users being almost 60% less successful than with the conventional screen. Users with the PDA interface performed better, but still 14% below users with the conventional screen. A non-significant trend could be noted in the time it took users to complete the tasks, in that the smaller the screen the longer it took users to complete the task. Users’ satisfaction also decreased with smaller screen sizes and users saw screen size as the most important issue that adversely affected their performance. The main reason why users failed to complete their tasks was that they tended to get lost in navigating through the sites they selected from the search results.

Another adverse effect of screen size is the increased amount of scrolling that is necessary. Marsden, Cherry and Haefele (2002) found that users often do not scroll down on a page. When asked why they did not scroll down, they answered that they simply had not seen the scrollbar and did not know that more information was available. This has serious implications for the design of a hierarchy. If users do not scroll down, options on the first few lines of a display will be selected faster than options further down in a list, even if such an option is not correct. In conditions where labels are not very clear, this is particularly true. Options that are not shown on the first screen (e.g. due to scrolling or paging) will be found more slowly than options that are shown immediately. Scrolling and/or paging will also earlier result in errors because users try to select an option from the visible options, instead of scrolling down to the end or looking on more than one page. This is a complication that might lead to the preference of deeper hierarchies on smaller screens, to prevent users to scroll unnecessarily.

There are several ways to deal with the limitations of small screens. One way is to just send a page to the device and let the user deal with it. Obviously, this does not give the desired results as it does not adapt at all to the situation of the mobile user. It gives the user a website that is designed for a desktop display, so users will have to scroll both vertically and horizontally to see the whole website. The disadvantages of this approach have been shown by Jones *et al.* (1999).

Much software has been produced to accommodate mobile users by rearranging text flow and reducing the size of images or omitting them completely, also called Web Clipping Applications (Wobbrock et al, 2002). Examples of such software are Plucker (www.plcker.org) and Skweezer (www.greenlightwireless.net), which both rearrange text flow to prevent users from having to scroll horizontally. However, the generated small screen website might still not be very usable depending on the original website layout and the amount of vertical scrolling needed increases fast. A better solution (but not fully automated) is to entirely reproduce content and present it in a special format for mobile users. This is already done commercially and has the advantage that content is designed especially for use on mobile devices, but the disadvantage is that content has to be restructured by hand. This means that only selected content will be available from a small number of providers and often at a certain fee (MacKay & Watters, 2003).

Advancements in automating the adaptation of content has been made by Buyukkokten *et al.* (2000, 2001a, 2001b), who worked on a method to automatically summarize sentences in a website. This enables users to get a quick overview of the contents of a website without having to scroll. When a user wants to read more about a certain part, that part of the summary can be clicked and it then expands so the user can read more. Buyukkokten *et al.* also studied the usability of this summarization technique, which showed a 45% gain in browsing speed and a reduction of 42% in required pen movements. Björk and Redström (1999) combined this technique with flip zooming so that users can see information in context, even when not all text fits the screen. Others have investigated ways to increase screen usage by using transparent buttons to gain more space (Kamba et al, 1996); zooming out and in on the website to gain overview (Wobbrock et al, 2002); use fisheye views (Bederson et al, 2004) or use sounds to reduce the dependency on visual items (Brewster, 2002).

The option under investigation in this research is to investigate the effectiveness of different hierarchies for small devices. Fewer options can be shown in one screen, so a deeper hierarchy with fewer options per page might be preferred. In an experiment by Giller *et al.* (2003) it also turned out that users seemed to prefer to see all items at a glance, with faster response times for broad hierarchies but a subjective preference for deeper ones. The main question is then whether users prefer to scroll down to see the remaining options, or to choose an option from the ones available and move deeper into the hierarchy. This question will return in the experiment.

Input modalities

Text input varies with different input modalities. Usually, these input modalities are directly coupled to certain device classes. A desktop user usually has a keyboard and a mouse at his disposal. A PDA user usually has a stylus with a touch screen and possibly a small keyboard, which is sometimes only shown on the screen itself. Mobile phone users are the most limited with only a numeric keypad, on which alphabetic input is laborious and error-prone. Each specific input device has its disadvantages, in text input and in selection. These are discussed below.

Osborne Rao (2000) studied the usability of three PDAs with different input devices. Users were asked to use these devices to perform some everyday tasks like entering an address or editing an appointment. Using a keyboard to input text is relatively fast compared to using a stylus, because handwriting recognition for stylus-based input was slow and error-prone. The input was significantly faster using a keyboard than using of a stylus. The combination of both, as featured in some devices, was preferred by users, mainly because they are similar to the combination of the mouse and keyboard on a desktop computer. Rao suggests that devices should use a keyboard for text entry, but a stylus for pointing and selecting.

Text input will not be used in the present experiment, which means that the presence of a keyboard will not be much of an advantage for the device under investigation. A stylus, however, would serve as a very useful selection device, compared to the numpad-based selection on most mobile phones. MacKenzie, Sellen and Buxton (1991) showed that pointing using a stylus was very similar to using a mouse for pointing tasks. The stylus in their experiment was used on a tablet instead of directly on-screen, which may have caused some small bias in favor of the mouse.

Mizobuchi, Mori, Ren and Michiaki (2002) compared the usability of stylus pointing and selection on a mobile device to pointing and selection using the cursor keys on the device with varying target sizes. They used a total of 36 options arranged in a square. In one condition the buttons all had a 2mm radius, in the other extreme the radius was 5mm. The white space between the buttons was 1.5mm, so the centers of the buttons moved further away from each other as the button size increased. The dependent variables were the time needed to go from selecting the first target to selecting the second target and the number of errors made. Significant differences were found in both time and error by input device as well as in an interaction with target size. It turned out that less time was needed to complete a task with the stylus than with the keys (approximately 0,8s and 1,4s, respectively). As the target size decreased, however, stylus input became slower. With decreasing size, users also made more errors in selecting the right button. With target sizes of 5mm, the number of errors was the same for both conditions, but for all smaller targets the number of errors became higher for the stylus and stayed approximately the same for the keys.

Mizobuchi *et al.* also found an effect for different number of targets (16, 36 and 64). For larger amounts of targets, the difference in time between navigating with the cursor keys and with the stylus became larger. As the path length increased, task times for both input devices increased, but much more for key input than for stylus input. Overall, stylus-based input was preferred for pointing and selecting items on-screen. It is faster, as long as the buttons are not too small. This effect is even larger when more items are shown on a screen. In terms of accuracy, a stylus was only preferred when the target size was at least 5mm. With smaller targets, the precision of cursor keys was higher than that of a stylus.

Although users will not be able to choose the type of input device, it is important to take the specific abilities and disabilities of specific input methods into account.

2.4.2 CONTEXT DIMENSION: IDENTITY

Wang, Hawk and Tenopir (2000) investigated the several levels of interaction in a study to find out how users search for factual information on the Web and what personal characteristics influenced this search. They divided the user dimension into four factors: *cognitive*, *affective, physical* and *situational* factors. The information retrieval process of users is supposed, among others, to be influenced by their mental model and cognitive style (cognitive), feelings and tendencies of the user (affective), as well as by particular motor skills and users' control of input devices (physical) and the particular task at hand (situational).

Personal characteristics: Cognitive factors

In the cognitive domain, the cognitive style of a user can influence the way a user navigates on the web. Cognitive style refers to the way an individual works in cognitive activities like perception and problem solving. Styles have been categorized in a number of (binary) ways: holist or serialist style, field-independent or field dependent style and verbalist or imagery style (Wilson, 1999).

Holists are people who adopt a more global approach to solving a problem. They first try to get an overview and then fill in the details, as compared to serialists, who tend to examine one thing at a time (Pask, 1976). Ford, Wilson, Foster, Ellis and Spink (2002) found that these cognitive styles also have an effect on search behavior. As expected, holists displayed more exploratory behavior to gain insight in the general structure before diving into the details, and holists report greater valuing of serendipitous browsing. Ford (1985) designed a test for measuring whether people are holists/serialists: The Study Process Questionnaire, which was reliability-tested by Clarke (1993), which are also used in assessing learning styles of students.

Another categorization of cognitive style is between field-independence and field-dependence. Highly field dependent individuals have more difficulty to transfer knowledge from one place to the next. This would make navigation through many screens more complicated for field-dependent individuals than for field-independent individuals. Wang *et al.* found this in their experiment, where a significant relation was found between time spent on a website and field-dependence. Palmquist and Kim (2000) also found that field-dependent (novice) searchers took longer and traversed more nodes in locating relevant information than field-independent novices. However, for more experienced users, they did not find significant results. There exists a standard test to measure field-dependence-independence, the Embedded Figures Test (EFT), where a user is presented with a picture and after that a second picture containing the first picture somewhere. The user is asked to locate the first picture in the second picture where response time is measured. Field-independent persons find the target faster than field-dependent persons.

The third categorization of cognitive style is that of verbalizers and imagers. Verbalizers are individuals who tend to think in words; imagers tend to think more in images. An information presentation that matches verbalizers or imagers aids them in task performance (Ford, Miller & Moss, 2005). On the internet, information retrieval performance using keyword search tends to be better for verbalizers. Imagers however tend to have fewer problems with disorientation and information overload compared to verbalizers. Cognitive style can be measured with the *Cognitive Styles Analysis* from Riding and Cheema (1991). The user receives a series of statements about the relationship between two words and asks the user whether that statement is right or not. Half the statements are about words (conceptual relations) and half of them about visual relationships. Verbalizers will respond faster on the first kind of statements, imagers on the second kind.

The second factor within the cognitive domain is the mental model a user builds up during navigation. A mental model is a person's internal psychological representation of how something works in the real world. It is important that a user's mental model of the system is in line with the system itself (Norman, 1988). However, mental models can be incomplete, incorrect and contradictory and vary over time. This also applies to mental models of the internet, information seeking and navigation. Wang *et al.* found that not all users had a correct model of how search engines, the browser and the computer worked. The users' expectations of what the system would do were not always right, which resulted in errors in navigation. For example, when a web directory and a search engine were presented to users on the same screen, some users assumed that limiting a search in a category was available. Their conclusion was that "the web is a difficult environment for developing correct user mental models due to heterogeneous objects, poor interfaces, and diverse Web organization". Albers and Kim (2000) mention that using an incorrect mental model makes it difficult for people to comprehend the situation and to interpret relations between items of information correctly.

Once a certain mental model is activated (whether correct or not), people will base their decisions on this model. In the case of an incorrect or incomplete model, this can result in performing correct actions in the wrong situation (Wickens, 1999). It is therefore important that users activate a correct model from the start of the interaction. Since the web is a complex environment for navigation, a correct mental model is important to have an idea where the user is on the web (Albers & Kim, 2000).

Personal characteristics: expertise

Experience is a cognitive factor that shapes the expectations of users. Experience within the subject domain has already been mentioned. The more users know about a subject the more they will be certain about the relevance of items (Vakkari, 1999). Experience in information retrieval itself can also have an effect on search behavior. Wang *et al.* however did not find a relation between experience in searching the web and the time needed to complete a search task. Others did find significant differences between users with different levels of web or computer experience. For example, Hölscher and Strube (2000) performed an experiment in which previous web and domain experience was related to search results. Successful search was related to knowledge in both areas. In another study, it was found that users who were "system experts" on using search engines performed better than novices in search information using search engines. Expertise on browsing (as opposed to searching) did not help users to perform better on browsing tasks. (Lazander, Biemans & Wopereis, 2000).

There is a large degree of variability in search performance among the participants in hierarchy experiments. Large individual differences between the participants might account for these performance differences. Norman (1991) mentions two types of individual differences, inherent differences that are not related to the task (perceptual ability, cognitive processing) and acquired differences in knowledge of problem solving strategies or knowledge of the subject. Of particular interest are acquired differences, because the information hierarchy might be adapted for people with different levels of knowledge of the system or the subject.

Expert users handle problems differently than novice users (Norman, 1991). This is also the reason why experts who have knowledge of the system perform better in navigation through hierarchies than novices. Experts are more certain of their choices, because they already have an idea of the endpoints of his selections. Experts typically search forwards rather than backwards and can plan out a search path in advance (Norman, 1991). Not only system knowledge experts but also domain knowledge experts perform better on hierarchies. Hollands and Merikle (1987) performed an experiment where items were organized into semantic categories, alphabetical categories or random categories. Experts were significantly better in the semantic categories, which provided a similar categorization to the knowledge in their heads. For novice users, a keyword directory or alphabetically ordered categories improve performance (Lee *et al.*, 1986; Hollands & Merikle, 1987). For system experts, it does not matter so much whether hierarchies are broad or deep; they will manage to find what they are looking for because of their greater ability to discriminate between concepts and categories according to Jacko, Salvendy and Koubek (1995). For novices, it would be better to use a deep hierarchy because it can funnel them in the right direction by which they avoid making mistakes that could happen easily in a broader hierarchy, especially when items are similarly named (Paap & Cooke, 1997).

Novices can learn to use a navigation hierarchy simply by working with it a lot, becoming experts by trial and error. In the end, the benefits of any ordering gradually disappear as users already know where items are located (Card, 1982) and ordering is not as important anymore. In this situation, a navigational system in which users can select an item fast (a broad hierarchy) will perform better than a system where users have to go through several screens before being able to select an item. Paap and Cooke (1997) mention a number of experiments in which performance differences did not fade away completely after a number of tasks. These studies used more complex tasks for their subjects, and may be more representative for normal hierarchy tasks. The organization of a hierarchy may aid a novice in building a conceptual model of the system, depending on whether it is logically organized or not.

The (total) search time is an important measure for the quality of a system; for experts, this search time can be reduced by providing aids that are – although they require more time to learn – faster to use. In hierarchy design, one can think of a broad hierarchy in a similar way. It might require some more time to get used to, but in the end, They use a hierarchy a lot, and do therefore want short ways to the target item. Broad hierarchies are the better choice for them. For novices, clarity is most important. Because the uncertainty in a deep navigation hierarchy is higher than in a broad hierarchy, a broad hierarchy is also the better choice for novices, although their performance will be slower than the performance of experts.

Personal characteristics: Affective factors

In the affective domain, two factors are distinguished (Spielberger, 1983): long-term tendencies, which can influence the way of navigation and short-term feelings, which can influence navigation, but may change with positive or negative experiences during navigation. Both factors can be measured with the State Trait Anxiety Inventory (Spielberger, 1983). Wang *et al.* found a correlation between times needed to answer a question and short-term feelings and a difference in short-term feelings before and after searches. Positive feelings improved following actions, where negative feelings lowered performance on following actions.

Personal characteristics: Physical factors

The physical part of navigating consists of sensorimotor skills, such as the control of input devices or the ability to read the screen. In normal navigation on the internet, this is hardly an issue, were keyboard and mouse and a normal desktop monitor are used. When navigation takes place on a mobile device, however, this can complicate matters quite a lot. A small screen can seriously affect navigation possibilities, and the input methods that are possible on mobile phones for example are not very inviting for long sessions of use. The specifics of the physical device are discussed later on.

Personal characteristics: Situational factors

The particular task at hand has a great influence on how a user searches. For each task a user executes it with a certain level of preparedness, a motivation to reach the goal, while the demands of the task can vary. Especially, if users are not very motivated to reach the goal, their information search will end quickly if they do not find the information fast. Unfortunately, experimental research on situational factors that influence search behavior and information seeking is lacking (Järvelin & Ingwersen, 2004).

2.4.3 USER'S CONTEXT: LOCATION

An important part of the user's context is taken up by the location and situation in which he or she wants to use the system, which can be thought of as the physical and social context, respectively. Think about a user who sits in the bus and wants to access the information contained in the system. The bus is shaking over the bumps in the road, the passengers around our user are all talking loudly to overcome the noise of the motor and all our user wants is some peace to be able to look up the wanted information. Of course, the current surroundings of the user are not exactly helping to concentrate on the task at hand. Is it possible to design an interface that can help this particular user? There are many similar situations and locations where the conditions for task performance are not optimal. Especially users of mobile devices will usually not sit at home when they want to access the system, hence the portability of the device. But this comes with the disadvantage of changing locations and situations: sitting in a train, walking on the street, drinking a coffee in the local bar with some friends.

There are many very different *physical* contexts in which an information-need may arise; an information-need that can be filled by using a mobile device to look up some piece of information. Advanced sensors and geographical computing enable devices to automatically find out the location of the user. This can also be used to present users with information that is relevant to the location of the user. Based on the current location, the user can be served with local cinema information, shopping information, travel guides, events or historical information. Context-sensitive (location-based) information is very useful for this kind of information, and actually such services are already available and in use. The focus lays on changing the *content* that is presented to the user, which is 'the system' as described in paragraph 2.1. However, the present study is focused on adapting 'the interface' or *how* information should be presented, not on *what* information is presented. The physical context and how it may influence preferences for certain structuring of information is an interesting topic. However, as already hardly any literature exists about preferences and structuring of information for mobile devices, narrowing that down further to "in certain physical contexts" brings the body of existing literature on the topic practically down to zero. However, it might be possible to make an educated guess about the subject. One can argue that in conditions where the constraints on the user are high, like in the bus example described above, the mental workload is also high, which means that task performance generally declines (Cox & Griffith, 1995). In such conditions, it is important to keep the mental workload as low as possible. This turns the small factor physical context into the much broader factors 'workload' and stress. Since keeping workload low is always important and not just in stress-situations, the best information structure to use in such a situation is exactly the same hierarchy as the one that is preferred in other situations.

The changing *social* context means that one is able to use the device alone, but also when one is amidst other people. These other people could be friends with whom one wants to share the information on the mobile device, or strangers from whom one wants to keep his privacy. Such changing social contexts have a particular influence on the design of such devices. A digital camera is equipped with a screen that allows a wide viewing angle; a mobile phone on the other hand has a rather limited viewing angle. In the area of information structures however, it is predicted that their effect on hierarchy design is rather limited. The social context might determine the specific task a user wants to perform in a situation, and adapting to a certain task may seem relevant. However, adapting the interface to a certain social situation, without knowing what task the user wants to accomplish, seems (besides probable technical difficulties) rather useless.

2.5 Conclusions of the Literature Review

This literature review covered the three topics that make up the model of Figure 6. The system, the interface and the user each have an influence on the usability. The system itself is considered to be something that cannot be changed, the information is there. The user can also not be changed, but the specific properties of the user are very important in finding and designing an interface that helps the user to reach his or her goals.

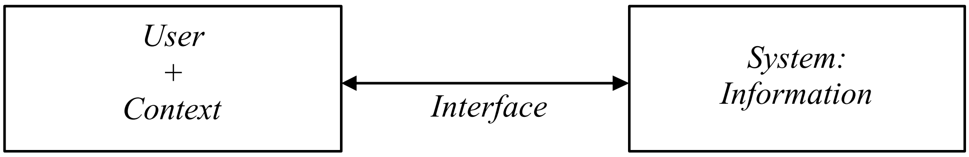


Figure 6: Model of User Information Interaction by means of an interface.

Based on the user's task and the goals he or she wants to achieve, the way he or she searches for information also varies. Serendipitous browsers are not searching directly for a certain piece of information, but are browsing with the intent to find 'something' that interests them. General-purpose browsers have more or less an idea of what they are looking for, and the searcher knows exactly what he or she is looking for. An information hierarchy is a good way to present all three kinds of information users with a structure through which they can access the information that lies beneath it. It provides meaningful information for the serendipitous browser who can decide whether or not to access more information about a certain topic, and on the other hand, it can guide a directed searcher immediately to the target of the search.

Not only do the users have different goals they want to achieve, every person has a different personality and cognitive style, which determines their preferred way of searching information. Some users are more visually oriented (imagers), others think more in conceptual models (verbalizers); some people first try to gain an understanding of the overall structure before diving in (holists), others only gain an idea of the structure by going through it (serialists); some people are better able to see links between concepts (field-independent) than others that have trouble transferring knowledge from one place to the next (field-dependent users). These cognitive styles influence the way that people interact with the internet, and how they understand the structure of a web site. Holists have it easier on a large web site, because they do not get lost so easy when navigating through many options, which is much more difficult for serialists. The cognitive style therefore influences the accessibility of the information contained in the system. Any information hierarchy is of more use to holists to serialists, because they also think in structures, which should also helps verbalizers more than imagers. Additionally, field-dependent users might have trouble when the navigation system is designed in such a way that many screens have to be accessed before reaching the information. Field-independent users will have fewer problems with any information structure. Assisting each and every user the same way is virtually impossible, but it is important to note that various cognitive styles exist and they place different demands on a system.

Besides the personality of the user, which is part of the user context, also the location of the user has been discussed. This location, also called physical and social context, are argued not to have a significant influence on preference for a certain information structure, do place limitations on the design of the device itself. For example, the device should make it possible to hide privacy sensitive data, or to allow users to easily share information with friends; allow operation under sub optimal lighting conditions, etc. However, this will not have an influence on the optimal interface to access the information system.

The first sub question that was formulated in paragraph 1.2 was "How can an information structure help (mobile) users with information navigation?" As was noted above, the information structure is the interface between the user and the information itself, and as such influences the navigation a lot. Based on theoretical predictions, a hierarchical information structure with about 7 to 13 items per level would be optimal to guide users the fastest to the item they are looking for. Experiments confirm that users can find information faster when an information hierarchy is broad and does not have many levels than when the structure is narrow and has many levels. Additionally, users who have more expertise place different demands on a system than novice users. When searching for information, expert users can find information faster than novice users, and make fewer errors on the way to the information, because they know more or less where it will be located. Therefore, information structures can be made that fit better to experts or to novice users, where expert users prefer a broader structure with fewer levels than novice users.

The third factor that was discussed in the user's context was the device that is used. Based on the literature described above, the conclusion of what kind of information structure to use on mobile devices would be simply to use many categories with as few levels as possible. Unfortunately, these findings do not transfer well to smaller screen sizes. The advantages of a broad information hierarchy are reduced drastically when not all options can be displayed on one screen and users have to scroll to see all options. Not all users want to scroll, or even know that they can scroll, which means that options below the cutoff point will not be seen nor clicked. This can be solved by either dividing the options over multiple pages with all its negative effects, or by reducing the number of categories and increasing the depth of the hierarchy, which also has negative effects on user performance. So, three options are left, each with its disadvantages.

The question that follows from these points, is: Is a broad hierarchy combined with a lot of scrolling or paging better than a small hierarchy without scrolling? Giller *et al.* (2003) claim that a narrow hierarchy with many levels is better for smaller screens. Unfortunately, their experiment was relatively small and the results were not all significant. Therefore, the next chapter describes an experiment in which a number of information hierarchies (tree-like structures) are tested on mobile devices. This sheds more light on the preferences of mobile users on information navigation.

3 Experiment

In the previous chapter, the existing literature on the topic of usability of information hierarchies has been discussed. It turned out that a vast body of literature exists on the topic, where recommendations varied between using 4 to 8 and 64 items per level, and some suggested even more would be possible. However, the properties of mobile devices might bring along additional limitations to the amount of items that can be displayed on the screen. That is exactly what this experiment is about.

The experiment tries to answer the question how we can change such information hierarchies to fit the changed context of a user when switching from a desktop computer to a mobile device. Therefore, four different hierarchies are tested in this experiment to see which performed best. An additional goal of the experiment is to find out whether different mobile devices with other technical properties will also show different results, due to other input possibilities or a larger screen. That is why the experiment is performed with three different devices to find out what the optimal hierarchy for each small screen device is. Finally, the experiment is repeated with the same subjects to test for any expertise effects.

The experiment setup with its hypotheses, design and procedure are discussed in this chapter.

3.1 Hypotheses

The experiment is based on seven hypotheses. The first two hypotheses deal with the traditional measures of usability (time, errors and satisfaction). The next four deal with indirect measures for usability: number of key presses, certainty of choice and perceived complexity. The last hypothesis deals with the effects of increasing expertise on the preferences of the user.

3.1.1 Traditional Usability measures

Literature shows that a broad and shallow hierarchy results in higher user satisfaction on desktop screens (Snowberry et al, 1983; Kiger, 1984; Landauer & Nachbar, 1985; Tullis, 1985; Wallace et al, 1987; Zaphiris, 2000). It is expected that this does not hold for mobile devices, because of the physical limitations of the device. Its small screen size will hinder navigation through lists (Giller et al, 2003; Resiel & Schneiderman, 1987). Because of the small screen, users typically have to scroll a lot when the list is longer. Scrolling could be seen as a form of horizontal navigation, where selecting an item and going one level deeper can be seen as vertical navigation. The advantage of this vertical navigation is that it actually narrows down the options to a relevant category, whereas horizontal navigation keeps all irrelevant items in the list. The smaller the screen, the more a user has to scroll through such irrelevant items before finding a relevant one. Therefore, it is expected that users of mobile devices have a preference for the narrower hierarchies. This is reflected in H1.

The larger the screen size of the mobile device, the more it resembles a desktop screen. This implies that it can be expected that the effect that is described in H1 will be seen less pronounced in large devices and more pronounced in smaller devices, in other words, it is expected that an interaction effect occurs between device and hierarchy. We could speak of a two-way effect if the small device would give better results (time, errors, and satisfaction) with the narrow hierarchy *and* if the large device would give better results with a broad hierarchy (see Figure 7). In such a case, the results of a test cannot be predicted alone by the choice of a certain hierarchy, and neither by the choice of a certain device alone. Only the combination of the device and the hierarchy together would yield a reliable prediction, because their combination results in a specific "interaction effect" that does not occur when the two factors are considered in isolation.

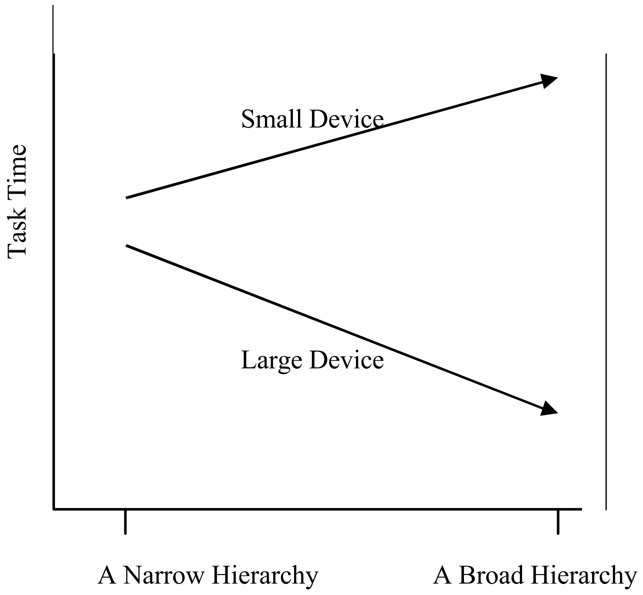


Figure 7: Expected interaction effect. It is expected that the task time of the small device will be lower using a narrower hierarchy and higher using a broader hierarchy. However, with a large device, this effect is expected to be reversed; a broader hierarchy will result in lower task times than a narrower hierarchy on the large device. The combination of these two expectations is the expected "interaction effect". Whether the lines actually cross each other or not is not known. The experiment will have to give an answer to that.

In this experiment, it is expected that such an interaction effect does occur (not only for time, but also for errors and satisfaction). The reason for this is that a large device resembles a traditional desktop screen more than a small device does. The small screen hinders a small device more, which is why narrower hierarchies will yield better results on the small device and larger hierarchies on the large device. This is shown in H2.

> **H1**: On mobile devices, the usability of narrow hierarchies will be higher in terms of (a) time to complete tasks, (b) errors and (c) satisfaction compared to broader hierarchies.
>
> **H2**: An interaction effect is expected between device and hierarchy. For smaller mobile devices, the optimal hierarchy in terms of (a) time, (b) errors and (c) satisfaction will be narrower than for larger mobile devices.

3.1.2 NUMBER OF KEYPRESSES

A number of additional measures were used to try and explain the effects as predicted by the above hypotheses. First, when users have to search and scroll down before finding the item they need, the number of keystrokes rapidly increases because of moving down (and possibly up again) through a list, which is measured as the number of key presses on a mobile device. This number of key presses cannot be the only measure for usability: user typically do not care if they need three or four clicks before reaching a goal (Krug, 2000). However, it does increase users' frustration levels if they need to find something and they need to press more keys than necessary before finally reaching their goal. This frustration level can lead to giving up and abandoning the system altogether, which is definitely not the intended goal of an interface. Therefore, it is important to keep the number of key presses low.

The increase in amount of necessary key presses is especially large in the case of a broad hierarchy, where the target item can be located far down on a page. In narrower hierarchies, the user will need fewer keystrokes, which results in better performance. Imagine a hierarchy with 64 items per page, with two levels. The average number of items that has to be reviewed or scrolled through is 32.5 items (assuming optimal, self-terminating, search) per level, or 65 items in total. In a hierarchy with 4 items per page, with six levels, the average number of items reviewed would be 2.5 per page, or 15 items in total. This is not even a quarter of the amount in the 64-items-case. These are, of course, theoretical values. The actual number of key presses also depends on the errors users make and the type of search they use. If a user changes the type of search between hierarchies, different numbers will be found than expected based on simple scrolling as mentioned above.

Add to that the fact that methods for input are troublesome, and suddenly the number of keys or buttons that have to be pressed can influence the usability significantly. This is represented by hypotheses H3 and H4. H3 presents the hypothesis that the number of key presses is related to the type of *hierarchy* used.

Hypothesis H4 focuses on the *device* used. The prediction is that small devices will introduce more key presses than large devices. This stems from the fact that the smaller the screen, the fewer items fit at once on the screen. A small screen user can see fewer options on the screen, so he or she will need more navigation to see all the items that are available in one list. It is then logical that in order to see all items, a user of a small device needs more key presses than a user of a large device.

> **H3**: The number of key presses in a narrow hierarchy will be lower than the number of key presses in a broad hierarchy.
>
> **H4**: The number of key presses on a small device will be higher than the number of key presses on a large device.

3.1.3 CERTAINTY OF CHOICE AND PERCEIVED COMPLEXITY

Two additional factors were measured in the experiment: certainty of choice and perceived complexity. Certainty of choice is an issue in that it is expected to increase the preference for broad hierarchies. In a broad hierarchy, users are presented with a long list of items; they have more options to consider. The chances that one of these options closely resembles their internal representation of the goal item are higher, and thus their certainty of choosing the right item increases. This pleads for broader hierarchies. Additionally, there is generally no uncertainty for items on the last level. These items either match the target item with 100% certainty or they do not match the target item with 100% certainty (Norman, 1991). This means that the last level of a hierarchy can effectively be skipped in choice certainty considerations. Removing this last level has a pronounced effect on the overall certainty of a hierarchy. Norman and Chin (1988) give a mathematical explanation based on information theory. Each level in a hierarchy has a number of information bits. A list of 2 items contains 1 bit of information (a binary choice), a list of 4 items contains 2 bits of information, a list of 8 items contains 3 bits of information, etc. Each level in a hierarchy also has an amount of uncertainty, equal to the information bits. In a six-level hierarchy with 4 options per page, the total number of information bits can be calculated by adding all information bits for all levels: 2+2+2+2+2+2. This is almost equal to the total amount of uncertainty of such a hierarchy, except that although the last level introduces two bits of extra information, it does not introduce extra uncertainty (as explained above). This means that the total amount of uncertainty is two bits less, 2+2+2+2+2+0 = 10 bits. For a number of hierarchies with each 4096 items in total (or 12 bits of information), the amount of uncertainty is given in the last column of Table 1, which shows that the more items per page, the less uncertainty such a hierarchy contains.

Table 1: Information bits and theoretical uncertainty

Items per page	Levels	Information	Uncertainty
2	12	1 bits * 12 = 12	1 bits * 11 = 11
4	6	2 bits * 6 = 12	2 bits * 5 = 10
8	4	3 bits * 4 = 12	3 bits * 3 = 9
16	3	4 bits * 3 = 12	4 bits * 2 = 8
64	2	6 bits * 2 = 12	5 bits * 1 = 5

Presenting the items in a broader hierarchy has the additional advantage that users have to traverse through fewer levels to reach the target item (which could result in disorientation and its associated uncertainty). This obviously does not correspond well with the previous hypotheses that predict that narrower hierarchies will perform better. Whether this factor will be large enough to reduce other effects or to negate them completely (or even invert them), is the question. This uncertainty could actually be one of the reasons why the literature generally points to preferences of broad hierarchies. The effect of (un)certainty of choice is represented by hypothesis H5.

> **H5**: A broader hierarchy leads to more certainty in choice than a narrower hierarchy, independent of the device used.

In relation to uncertainty that is introduced when users have to traverse through pages before they can find out if they are on the right path is the complexity of the hierarchy. Any way of displaying many items across multiple pages introduces some form of complexity, as users want to and have to remember the navigational paths they follow in an attempt not to get lost. This is increasingly difficult for users to do as more pages and possible paths are introduced and the memory burden increases. However, complexity as such is not a real issue. A very complex system may be presented to the user one step at a time and may be perceived by the user as not complex at all. That is why perceived complexity is a better measure for expected usability than complexity itself (e.g. in terms of number of page transitions). Research does show the perceived complexity of the navigation system is higher in narrow hierarchies than in broad hierarchies (Jacko and Salvendy, 1996). This contradicts the first hypotheses that predicted that users will prefer the narrow hierarchies. The experiment will have to show if this effect is large enough to alter the preferences. Along the lines of Jacko and Salvendy, it is expected that the perceived complexity for narrower hierarchies will be higher than for broader hierarchies in this experiment. This is represented by hypotheses H6.

> **H6**: The perceived complexity is expected to be higher for narrow navigation hierarchies and lower for broad hierarchies.

3.1.4 EXPERTISE

Experience in using an information hierarchy gives users the ability to oversee more options at once without being overwhelmed by the amount of choices. Expert users thus should be able to perform better with all hierarchies, because they already know the endpoints of their selections and alter the way they search information based on this knowledge (Norman, 1991). It can be said that with regard to the criterion-based model described in the previous chapter, an expert places the L and H criteria closer together, as each option is either rejected immediately or accepted immediately, and his or search pattern then resembles a self-terminating search. In case of any ordering logic, this process can be even sped up because the user can jump immediately to the area of the target item without reviewing every single item on his or her way. As the user's response time declines, the response time of the system becomes more and more important in determining the fastest hierarchy. In the case of a slow system fewer transitions means better performance which implies a broad hierarchy will be better in terms of both time and preference.

On the other hand, a narrower hierarchy is expected to assist novice users more, in which they are 'funneled' to the goal item without having to consider too many alternatives. It is expected that their search will not directly approximate a self-terminating search but rather involve a more thorough evaluation of each item before placing it at a certain point in the L-H-criterion-model. More items will fall in between immediate rejection and immediate acceptance and thus more items will have to be evaluated. The computer response time is then not as important as the total human response time, which in turn favors a narrower hierarchy with fewer items per page. This is represented by hypothesis H7.

> **H7**: When participants' expertise increases, their preference for a certain hierarchy will change from a narrow to a broader hierarchy (independent of device).

These are the seven hypotheses that serve as the basis for this experiment. As stated before, the hypotheses are not in perfect agreement with each other. For example, it is expected that a broad hierarchy will have more positive result concerning certainty of choice (H5) and perceived complexity (H6), but requires more key presses (H3) and is expected to perform worse in the traditional usability measures (H1). Although these hypotheses seem to contradict each other, they are to be seen as *complementary* to each other in the sense that together they present a better picture of what happens exactly. Even if the effect of one factor might be so large that it reduces the effect of one or more other factors, it is important to know exactly which factors influence the preference for a certain hierarchy and what influence other factors exactly have.

3.2 PARTICIPANTS

Fifteen people took part in the experiment as novice users. The average age of the participants was 24.3, the youngest participant being 17, the oldest participant 39.Of those fifteen people, fourteen came back a second and third time to participate as intermediate and expert users, based on the experience they gained during the tasks they performed as novices. Participants received a monetary compensation of € 30, - (around $35) for their time and cooperation.

3.3 MATERIALS

3.3.1 LOCATION

All user tests took place at the labs at CURE, where video and audio equipment was available for the necessary recordings. Recorded were video and audio data of the users' actions and comments for further analysis. Additionally, a small camera was mounted on top of the devices so that the experimenter was able to see and record the same thing on the screen as the user could see. This camera did not interfere with the mobility of the device, so that users were still able to move the device around freely.

3.3.2 HIERARCHIES

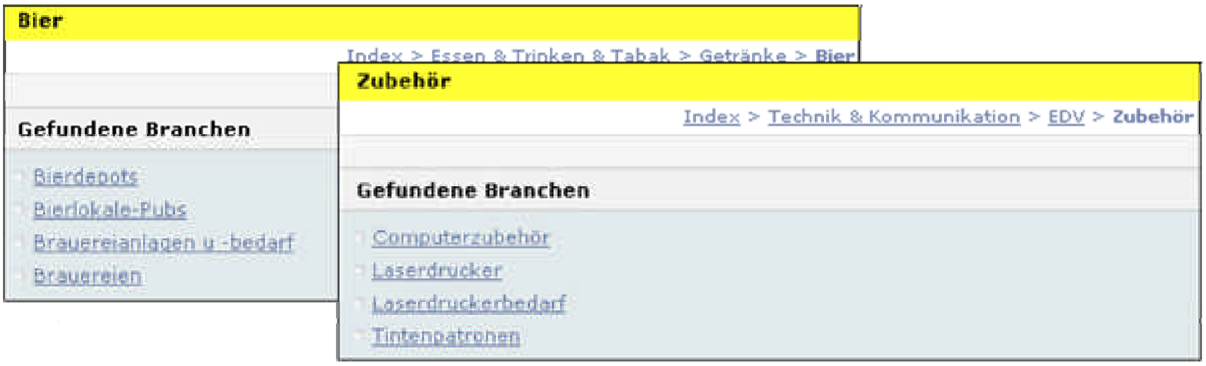
The hierarchies that are chosen for the experiment are the four hierarchies (Very Narrow, Narrow, Broad and Very Broad) that are depicted in Table 2. Each hierarchy has a total of 4096 items on the deepest level of the hierarchy. The main reason for this choice is that participants are not able to learn the paths to the target item by heart. Instead, they have to make conscious decisions about each category and subcategory. This makes differences between hierarchies clearer. The goal of the experiment is not to test the reaction time on a preprogrammed set of paths, but to simulate a real-world situation where a user does not know exactly where he or she is going to end up. This means a large number of items on the lowest level of the hierarchy has specific advantages.

Table 2: Number of items per level and number of levels in the four hierarchies that were used

Hierarchy	Items per level	Number of levels
Very narrow	4 items per level	6 levels
Narrow	8 items per level	4 levels
Broad	16 items per level	3 levels
Very broad	64 items per level	2 levels

When using just 64 items in a hierarchy like Miller (1981) did, users will need only a few trials before they know exactly where each item is located and how to get there and uncertainty is practically zero. Actually, Miller allowed users to study the hierarchies quite thoroughly before he recorded their trials to avoid any learning effects. This also pre-adjusts users' search criteria and patterns. Users are practically already system experts, and their preferences have likely changed accordingly. Additionally, errors made in Miller's "known" environment are most probably slips or unintentionally selecting the wrong item and realizing the slip almost immediately. Such errors are very different from intentionally selecting an item that turns out to be wrong, which can be attributed to an incomplete mental model. All of these are reasons to use a large number of items instead of a small number, as a number of earlier experimenters did. Such a large hierarchy also brings one problem along though, which is the need to construct it.

The hierarchies that were used in the experiment were loosely based on the Austrian yellow pages categorization (Herold.at); see Figure 8 for an example of what these pages look like. Unfortunately, the categories and subcategories of these yellow pages are not symmetrical so they had to be reorganized into a symmetrical hierarchy with equal amounts of options on all levels, which is quite a tedious work for 4096 items.



Bier

Index > Essen & Trinken & Tabak > Getränke > Bier

Gefundene Branchen

- Bierdepots
- Bierlokale-Pubs
- Brauereianlagen u -bedarf
- Brauereien

Zubehör

Index > Technik & Kommunikation > EDV > Zubehör

Gefundene Branchen

- Computerzubehör
- Laserdrucker
- Laserdruckerbedarf
- Tintenpatronen

Figure 8: Examples of the original Herold categorization

The reorganization of the items was done based on the scheme used by Kiger (1984). Starting with designing the 'Very Narrow'-hierarchy with 4 options per level, the 'Broad'-hierarchy with 16 options per level can be constructed by removing the first, third and fifth levels. The second level in the Very Narrow-hierarchy contains 16 options, which matches the first level of the Broad hierarchy perfectly. Of course, the labels of these hierarchies do have to be adjusted for the loss of context information due to removing category headings; usually just slight adjustments are necessary. After that, the 'Very Broad'-hierarchy can be constructed in a similar way, removing all levels from the very narrow hierarchy except the third and sixth (64 options and 4096 options, respectively), again adjusting the labels for the loss of context. Only the 'Narrow'-hierarchy was not that straightforward to construct from the others, because 8 is not a power of 4. This hierarchy was also based on the 'Very Narrow'-hierarchy by selectively joining two subcategories hierarchy and renaming their headings to match their new content, effectively removing two levels.

Each of the four newly constructed hierarchies now has a specific number of items per level (4, 8, 16 and 64) and all have a total number of 4096 items on the lowest level. The first level of all hierarchies is displayed in Figure 9. The complete hierarchies can be found at http://wap.cure-vienna.org/hierarchy.wml as WAP pages.

very narrow

Bau, Wohnen & Umwelt
Finanzen & Recht

narrow

Arbeit & Büro
Banken & Finanzen

broad

Auto & Mobiles
Banken & Finanzen

very broad

Abfallwirtschaft
Arbeit & Beruf

respectively. The number of tasks that users have to perform increases for each added device. Therefore, the experiment was not performed with devices from all device classes, but only with the three most commonly used ones.

3.4 PROCEDURE

Over a period of three weeks the participants were asked to come to the research lab three times, to participate in the experiment. The experiment is set up as a within-subject experiment, which means that each participant performs tasks with all four hierarchies on each of the three devices: a total of 12 different conditions per appointment. This was particularly chosen to be able to filter out individual differences between conditions. In a between-subjects design, more participants would be necessary to reach a similar degree of certainty, at the cost of possible ordering effects if no care is taken in the exact experiment design.

To make sure no ordering effects occurred, the ordering of the conditions was different for each participant. Some participants start with the small, some with the medium and some with the large device. Participant one first uses the Small (Siemens) device with the very narrow-, then the narrow-, then the broad-, and finally the very broad-hierarchy, and repeat this with the medium (Nokia) and the large (iPaq) device; participant two starts with the medium device, but with the same ordering of the hierarchies; participant three starts with the large device, again with the same ordering. Participant four starts again with the first device, but with a rotated ordering of the hierarchies, and so on until participants fifteen. Additionally, the order of the hierarchies was rotated among participants. This makes sure that the design is balanced (as opposed to really randomizing the order). Learning effects still occur, but they do not systematically occur on one place more than on another. This equal spread allows for statistical analysis afterwards.

The experiment starts with a briefing, to inform the participants of the experiment procedure, how long the experiment will take, what information is be recorded, which devices are used and how the participants can use them. The participants do not receive explicit information on the hypotheses or goals of the experiment to avoid bias in their behavior or answers to questions. The participants also receive ethical information, in the form of a video-consent form, the knowledge that they are free to leave the experiment whenever they want and they are told explicitly that the experiment is about testing the product, and not about testing its users.

They are then asked to try out the devices, to get used to the particular three devices and to avoid bias from unfamiliarity with the way the devices work. Each participant performs one example task with each device with the hierarchy they are using first. This is by no means sufficient time to really get to know the contents of the hierarchies, which also is not the intention of this practice. Users are allowed to get used to the devices and how the hierarchies look, but not to learn the contents of the hierarchies. They should not be able to learn certain action-sequences for performing each task, but instead develop a cognitive model of how the hierarchy is organized and use this model. After these tryouts, the experiment starts.

The participants perform three tasks with each hierarchy and each device, consisting of locating a product on the deepest level of the hierarchy. The exact formulation of the target is given in the task, in the context of a sentence. For example, a task is "For the upcoming winter, I am looking for some new *ice-skates*", where the word in italic print is the target item, which can be found in the narrow hierarchy under “free time & sports > sporting equipment > winter equipment > ice-skates” or similar categories and subcategories in the other hierarchies. The whole experiment takes place in German: the explanations, questionnaire, the tasks and hierarchy wordings are all in German. They have been translated here for ease of understanding. The original tasks are included in Appendix.

The number of errors is recorded together with the time needed to complete the task and the number of key presses they need to reach the goal. If participants cannot find the item within 90 seconds, they are given a hint where to find it. If they still do not reach the target item after 150 seconds, the participants are told where the target item is located and the task is terminated. After task completion, the participants are asked to rate the specific hierarchy and to give an estimate of the number of levels the hierarchy has. After completion of each task, participants are also asked to recall as many headings on the way to the target as they can, as an added measure for complexity.

In an additional task for each device and each hierarchy, users are asked to rate the certainty of their choice for each category or subcategory until they found the target, immediately after selecting such a category. This is done in a separate task because giving a rating each time a selection is made severely interrupts the normal task flow. In a situation where a user is brought out of the task to think and answer a question and go back to the tasks, time measurements will not represent times that are measured in other situations. Error measurement in such situation will also not be representative and therefore omitted.

The participants participate in the experiment on three different appointments (spread over three weeks). Because of learning during the tasks, participants have more experience with the information hierarchy at each consecutive appointment. This way, it is possible to test whether such experience has an effect on the preferred type of information hierarchy. The participants are not supposed to learn a certain action-sequence for a particular task, but rather develop a general understanding of the different hierarchies. For this reason, the tasks that participants performed as novice, intermediate and expert users are not equal. Although this means that it is not possible to compare task times and errors between appointments (because of varying task difficulty), it is possible to get a better picture of how user preferences change over time.

After completion of all tasks the participants are thanked for their cooperation and receive a monetary compensation for their time.

4 Results of the experiment

The experiment was performed in the first three weeks of August 2005. The data have been analyzed with SPSS and are reported below.

4.1 TRADITIONAL USABILITY MEASURES

The first two hypotheses dealt with three traditional usability measures: time, errors and satisfaction. It was generally expected that as the hierarchy becomes broader, the usability performance of the hierarchy becomes worse (H1). This effect was expected to be the strongest for the small device and the weakest for the large device (H2). The three independent variables time, errors and satisfaction are discussed below.

4.1.1 TIME TO FIND THE TARGET ITEM

Time was recorded for each task from the start signal until the moment the participant indicated that he or she reached the target item, with a maximum of 150 seconds. Predictions for these times were given in H1a and H2a. The results of the analyses of time are represented in Table 3 and are graphically depicted in Figure 11, for each device and hierarchy.

Table 3: Average search times per hierarchy and device, with their respective standard deviations

	Small		Medium		Large	
	Time	**SD**	**Time**	**SD**	**Time**	**SD**
Very Narrow	66s	22,9	59s	21,4	27s	19,9
Narrow	58s	22,5	53s	19,5	26s	17,6
Broad	64s	21,2	48s	18,3	29s	11,9
Very Broad	87s	25,5	55s	32,8	40s	26,7

A two-factorial ANOVA test for repeated measures revealed that the average times varied significantly between the *hierarchies* (F(3,39) = 3.6, p = .020). To find out where and how these times vary between hierarchies, the mean times of the hierarchies were compared pair-wise, Bonferroni-adjusted for multiple comparisons. With the small device, a significant difference was found between the very narrow and the very broad hierarchy (p = .002), which was also expected. Unfortunately, the observed variance within the hierarchies was so large compared to the relatively small differences between the hierarchies that no significant differences were found, except the one mentioned.

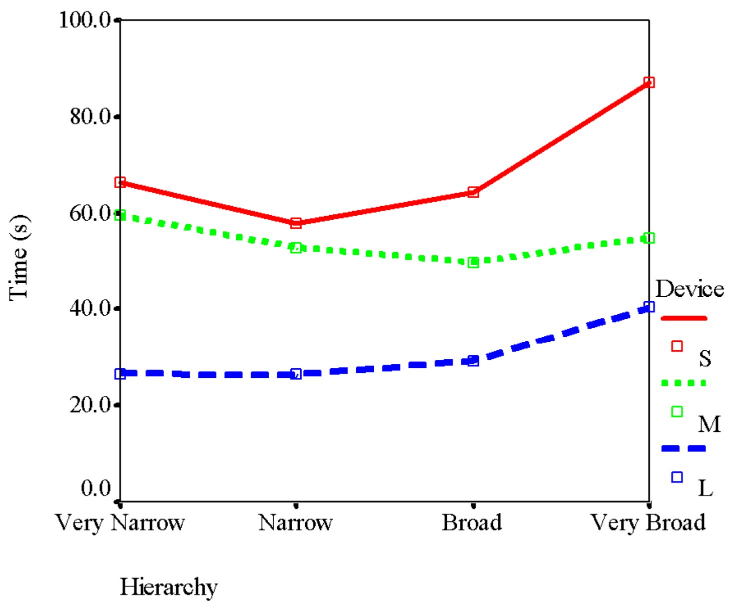


Figure 11: Time to find the target item, in seconds, per hierarchy, by device. The difference between the devices in terms of time is clear, but the expected difference between the four hierarchies is hardly visible.

After testing whether the times vary between the hierarchies, it was also investigated whether the times vary *between the devices*. Using the same two-factorial ANOVA test for repeated measures, it was found that the average times actually did vary significantly between the devices (F(2,26) = 49.3, p = .000). To find out if a trend could be made visible from device to device, the times for the three were compared pair-wise, Bonferroni-adjusted for multiple comparisons. The means for the three devices all differed significantly from each other. Participants performed the best with the large device, 30.6s on average, and worst with the small device, 68.9s on average. The performance with the medium device was in between that with the other two devices, with an average of 54.2s.

As reported above, the small device gave slower results for all hierarchies and the large device gave faster results for all hierarchies. This means that it is not possible to speak of a two-way (interaction) effect in this measurement. Although both hierarchy and device had a significant effect on task time, no significant interaction effect in time was found between the hierarchies and devices. (F(6,78) = 1.44, p = .210).

Based on these results it is possible to partly accept hypothesis H1a. It is shown that there is actually a significant effect in favor of the narrow hierarchy, albeit it only on the small device. However, hypothesis H2a cannot be accepted based on these results, since no significant interaction effect was found.

It is interesting to note in this respect that during the experiment it seemed that participants learned from their mistakes over time. It seemed as if they always needed more time the first time they faced a certain task, no matter which hierarchy they were using. Users always performed the same task with all hierarchies and because the hierarchies are somewhat similar to each other, it is logical that users can better predict the location of an item after they have its location once. Luckily, this is not very harmful for the theoretical setup of the experiment; after all, such effects of practicing were taken into account in the design of the experiment, which is the main reason why a balanced design was chosen. Each participant used the hierarchies and devices in a different order, so that learning effects are equally spread over all conditions. The main problem of this approach is that it also generates a lot of noise in the variance of the testing, especially if the learning effects are larger than expected. Therefore, the data was additionally analyzed to see whether such learning effect might be visible.

It turns out that such an effect also shows up in the data. No matter what hierarchy was used first, participants structurally needed more time with the first hierarchy they encountered on a device, with an average of 17 seconds more than with the other hierarchies, see Figure 12. A two-factor ANOVA test for repeated measures with the order as a covariate revealed that this finding was also significant (F(9,30) = 2.82, p = .016), but the type of device does not affect it.

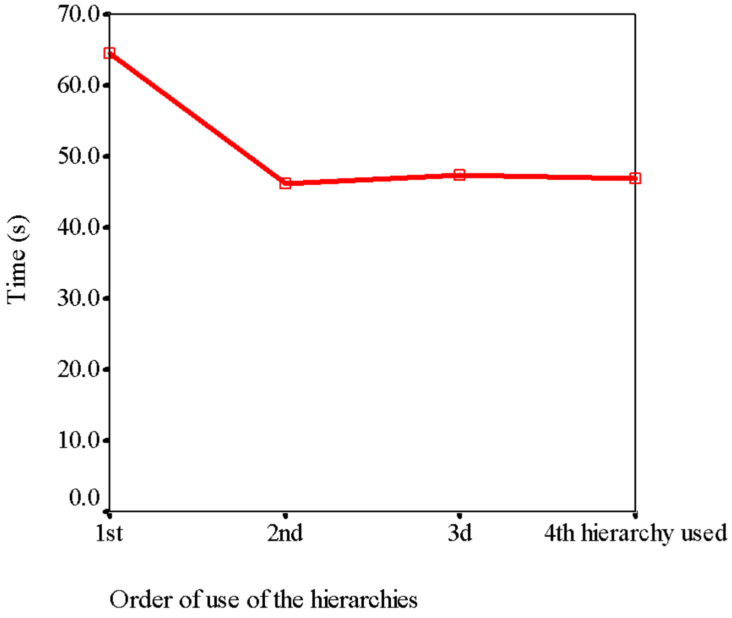


Figure 12: Effects of ordering on the total search time. Users structurally need more time with the first hierarchy than with the other three, independent of the structure of the hierarchy.

This does not have a significant effect on the rest of the experiment, but it is a good explanation of why there is so much variance across conditions. A participant who started with the Very Narrow hierarchy will have approximately 17 seconds added in this condition compared to the other three hierarchies, whereas another participant has this time added to another condition. In effect, all conditions thus have such extra time that was not accounted for by the used hierarchy and device, which made it difficult to find significant results.

4.1.2 Errors

Every time the user pressed the back-button to go up one level in the hierarchy, he or she either realized that he or she was in the right category and incorrectly decided to go back (Type I error), or was in the false category and made the right decision to go back (Type II error). Both types of errors are navigation errors. Therefore, the number of times users pressed the 'back' button was used as a measure for the errors made. Both types of errors were counted equally. Hypotheses H2a and H2b predicted that the number of errors would vary with the hierarchy used in general, and in the form of an interaction between the hierarchy and device used.

The counts of errors were analyzed with an ANOVA test for repeated measures. This test revealed that the mean values of these errors neither varied significantly for the four hierarchies (F(2.28,27.4) = 2.388, p = .085)[1] nor for the three devices (F(2,24) = .633, p = .54). The random variance for users overshadows variance analyses for both the four hierarchies and the three devices. Some users made many errors, other users almost none, which did not depend on hierarchy or device, but seemed to occur randomly.

A test for an interaction effect for device and hierarchy on the number of errors was also not found (F(4.43,53.2) = .322, p = .924)[1]. This means that the number of errors made during the experiment was not influenced by certain combinations of hierarchy and device, although such interaction expected.

This means both H1b and H2b cannot be accepted. The number of errors did not vary significantly between any of the conditions.

4.1.3 SATISFACTION

Satisfaction was tested by measuring personal preferences from the participants. This can give an answer to hypotheses H1c and H2c that deal with satisfaction.

The participants were asked which hierarchy they preferred after working with all four hierarchies on a device. This gave an interesting result: most of the users indicated that they preferred one of the hierarchies Very Narrow and Narrow. These results are depicted graphically in Figure 13. A significant difference was found between the preferences for the four hierarchies ($?^2 = 15.8$, p = .001), as seen in Figure 13.

What also can be seen in the figure is that the number of users who prefer the narrowest hierarchy is lower for the larger devices (which is an interaction effect between device and hierarchy). Seven participants preferred the narrowest hierarchy on the small device, but only four participants preferred it on the large device. Instead, they chose for the broader hierarchies. Unfortunately, a test for this trend proved that the trend is not statistically significant ($?^2 = 1.95$, p = .924).

[1] The assumption of sphericity was not met, therefore, Huynh-Feldt corrected degrees of freedom have been used in the analysis

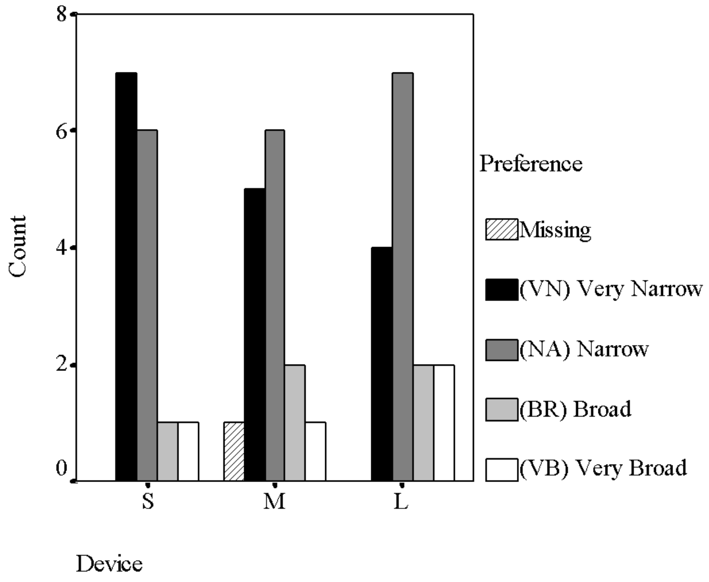


Figure 13: User preference for hierarchy, for each of the devices.

The above findings mean that we can accept hypothesis H1c, satisfaction for the narrower hierarchies is higher than for the broader hierarchies. However, we cannot accept hypothesis H2c that predicts an interaction effect between device and hierarchy, because no significant difference was found in preference between the three devices.

As a second way to measure satisfaction, users were asked in each condition how they would rate the website on a scale from 1 to 5, with 1 being very good and 5 being very bad. An ANOVA test controlled for repeated measures revealed that the mean values for these ratings do not vary significantly between the four hierarchies (F(3,39) = 2.424, p = .092). Satisfaction also did not vary significantly between the three devices (F(2,26) = 2.180, p = .133). An interaction effect also did not prove to be significant (F(6,78) = .920, p = .485).

4.2 NUMBER OF KEY PRESSES

Hypotheses H3 and H4 dealt with the number of key presses that is necessary with the different devices and hierarchies to reach the target item. This was measured by counting all key presses or clicks, including scrolling, selecting and back-navigation. On a larger screen,

A two-factorial ANOVA for repeated measures showed that the number of key presses varied significantly across the conditions with the four different hierarchies (F(1.87, 22.4) = 128.15, p = .000)[1]. The number of key presses is increases with every increase in number of items per page, with the least key presses in the Very Narrow-hierarchy, and the most in the Very Broad hierarchy. Additional analysis showed that the number also varied significantly across conditions with the three devices (F(2, 24) = 68.41, p = .000). The number of key presses is higher with the small device than with the medium device, and the number of key presses with the medium device is higher than with the large device.

Although not predicted, an interaction effect was also found. This interaction effect shows that the number of key presses for a certain hierarchy does not vary equally for all devices. For example, the effect of the Very Broad hierarchy is much stronger for the Small device than for the Medium device. The interaction effect also proved to be significant (F(1.87, 22.5) = 8.374, p = .000) [1].

This means both hypotheses H3 and H4 can be accepted and additionally an interaction effect can be mentioned.

4.3 CERTAINTY OF CHOICE AND PERCEIVED COMPLEXITY

Hypothesis H5 dealt with the certainty of choice for the different hierarchies and Hypothesis H6 dealt with the perceived complexity. The results in respect to these hypotheses are presented below.

4.3.1 CERTAINTY OF CHOICE

In an additional task to measure certainty of choice, participants were asked after each selection how sure they were that the target item was located in the category that they just selected (1 = very certain, 5 is very uncertain). These ratings were averaged for each hierarchy and device and analyzed with a two-factorial ANOVA test for repeated measures. This revealed that no significant difference existed between the three devices (F(2,14) = .542, p = .594). There also did not appear to be a significant difference between the four hierarchies (F(3,21) = 1.457, p = .255). Finally, no interaction effect was found either (F(6,42) = .112, p = .539).

It was predicted that a broader hierarchy leads to more certainty in choice than a narrower hierarchy. Since the certainty of choice did not vary significantly between the four hierarchies, hypothesis H5 cannot be accepted.

4.3.2 PERCEIVED COMPLEXITY

To measure perceived complexity, users were asked to estimate the number of levels of the hierarchy. In H6, it was predicted that the perceived complexity of a hierarchy increases with the increase of the number of levels. This does not vary significantly between the devices (F(2,24) = 1.578, p = .225), but the findings for the different hierarchies are interesting. It is not expected that the perceived number of levels is the same for all hierarchies, but the perceived values should approximate the actual values.

Interestingly, all perceived values, except for the Narrow hierarchy, are significantly different from their actual values: the very narrow hierarchy was perceived to have 5.3 levels which is less than the 6 levels it actually has. Users systematically underestimated the number of levels of this hierarchy. This means that the perceived complexity of this hierarchy is actually lower than its actual complexity, as displayed in

. In contrast, the Broad and the Very Broad hierarchy were both estimated to have more levels than they actually had. This overestimation is even more interesting. It is known that users underestimate large numbers, but that they also overestimate small numbers is strange, since people are quite able to count until two or three.

Table 4: Perceived number of levels per hierarchy (all differ significantly from their actual values, except the 'Narrow'-Hierarchy)

Hierarchy	Number of Levels	Perceived levels	Std. Error
Very Narrow	6	5.310	.139
Narrow	4	4.083	.143
Broad	3	3.583	.186
Very Broad	2	2.381	.149

The participants were also asked whether they remembered which links they followed to reach the target item. Participants could often not remember the exact titles of the links they followed (which could be expected). In addition to that, in the narrower hierarchies, users often also forgot to mention some levels. On the other hand, in broader hierarchies, some participants 'invented' more links than they had actually seen. This matches well with the result above that participants overestimate the number of levels in the broad hierarchies and underestimate the number of levels in the narrowest hierarchy.

Based on these data, hypothesis H6, that claimed that the perceived complexity will increase with an increase of the number of levels, has to be rejected. Instead, it turned out that the results were actually opposite to what was expected.

4.4 Expertise Effects

Hypothesis H7 predicted a change in user preference for a certain hierarchy with the increase of expertise. The effect of expertise in terms of satisfaction and preference are presented below.

4.4.1 Expertise and User Satisfaction Ratings

Expertise was taken into account in the experiment by asking participants to return a second and third time to perform similar tasks as they performed the first time. Since there were no interaction effects found between device and hierarchy, it was chosen not to test expertise effects for all devices. Instead, the experiment was only repeated with the medium and large device.

For satisfaction (ratings), the ratings that each participant gave to the different hierarchies were analyzed using a three-factor ANOVA-test for device, hierarchy and expertise level. This analysis reveals that there were no changes in rating for certain hierarchies found for different expertise levels.

The test did reveal a main effect for hierarchy (F(3,36) = 6.13, p = .002), which means that over the three sessions, the satisfaction rating per hierarchy have become more exact; was the variance within each hierarchy too large before to give significant results, now the differences between the four hierarchies have become a bit more clear. To find out more precisely where and how the ratings differed between the hierarchies, a pair-wise comparison was performed (Bonferroni-adjusted). This indicated that the significant effect was mainly due to a significantly better rating for the very narrow than for the very broad hierarchy (p = .004). The satisfaction ratings for each hierarchy are depicted for the three expertise levels in Figure 14.

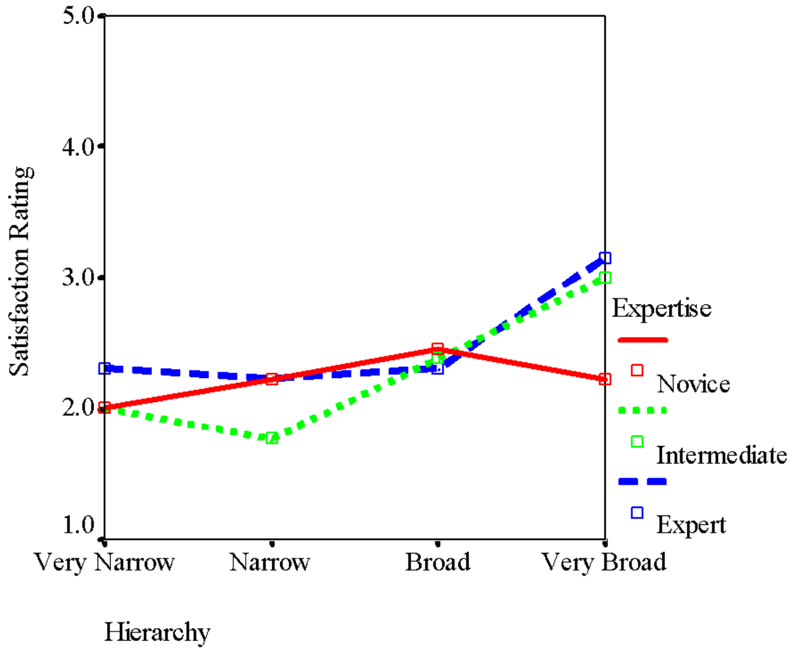


Figure 14: Satisfaction ratings (the lower the better) by hierarchy, for each expertise level. The trends that are visible between the expertise groups are not significant. Nonetheless it seems that although novices rated the very broad hierarchy relatively good, this rating becomes worse at the second and third occasion they use the hierarchy.

4.4.2 Expertise and user Preferences

It was expected that experts change their search strategy and criteria in such a way that they do not need to review as many items in a list as novice users. Therefore, they were expected to prefer broader hierarchies. They would be able to appreciate this broad structure because they are able to deal with the large lists involved, and have the subsequent advantages of the fewer levels linked to the broader hierarchies. It turned out, however, that the expertise level did not have a significant influence on the preference for one hierarchy over an other ($? ^2$ = 5.42, df = 6, p = .490). Not only novices, but also intermediates and experts preferred the two narrow hierarchies to the two broader hierarchies (see Figure 15).

What also can be seen in Figure 15 is a trend that the preference for the Very Narrow hierarchy seems to increase and the preference for the very broad hierarchy is completely reduced to zero. However, this result cannot be statistically verified.

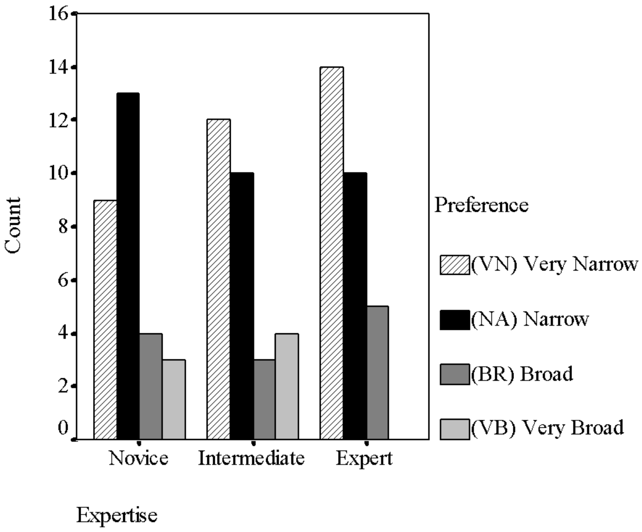


Figure 15: User preferences for hierarchy, per expertise level.

Although it was expected that users with more expertise would change their preference in the direction of a broader device, this seemed not to be the case. Therefore, we have to reject hypothesis H7.

5 Discussion of the Experiment Results

5.1 TIME AND ERRORS

Comparing the predicted effects of hierarchy and device on time (Figure 7) and the actual findings during the experiment (Figure 11), it seems that the hypotheses concerning this variable were too optimistic. It was expected that a clear trend would be visible, but it seems this trend is not as large as expected. On top of that, the effect of which hierarchy was used first is relatively large in terms of time (17 seconds) and increases the amount of variance that cannot be explained by the type of hierarchy. It is not possible now to say that one hierarchy is better in terms of time or errors than another. This can mean either that there is no such effect or that the effect is relatively small and therefore not visible in this experiment. Before being able to give an answer on the "optimal hierarchy", we will have to take a look at the other factors that were investigated. Search times and errors do not give a conclusive answer.

5.2 USER PREFERENCES

The experiment described above provides a number of more promising results concerning user preferences. It turns out that narrower hierarchies are better suited for the mobile user than broader hierarchies. Although this does not show in the experiment in terms of task time and number of errors, users very clearly prefer the narrower hierarchies on all three devices. This contradicts many previous findings, but is more in line with the expectations based on specific mobile usability issues.

As predicted in hypothesis H3, the number of key presses is lower when the number of items on one level is lower too (independent of the exact search strategy). Of course, this reduced amount of key presses can also play a significant role in determining a user's preference. The number of key presses seems to be in line with the user preferences, as the narrowest hierarchy was preferred and needed the fewest key presses. Indeed, as a participant remarked "I prefer to be lazy and go through a few more levels, rather than reading all the items in the hierarchies with more items per level".

5.3 CERTAINTY OF CHOICE AND PERCEIVED COMPLEXITY

Certainty of choice did not seem to be an issue in this experiment. Although a few participants got entirely lost once or twice, all participants generally were very sure of their choices. Actually, the average value of the reported certainty for the participants was 1.39, which is very close to the maximum possible value of 1. Even if in reality there is a difference in certainty of choice, it would be difficult to measure, because the values of the self-reported measure are so close to each other.

Another interesting result was found in the perceived complexity of the four hierarchies. It turned out that the number of levels of the very narrow hierarchy was consistently underestimated whereas the number of levels of the broader hierarchies was structurally overestimated. This is contrary to results from Jacko and Salvendy (1996), who found that narrower hierarchies result in higher perceived complexity.

5.4 EXPERTISE EFFECTS

Although other researchers found distinct expertise effects in their experiments (Jacko, Salvendy & Koubek, 1995; Paap & Cooke, 1997), no such effect was found in this experiment. This could be due to the fact that users were still not really experts after the three hours of testing. Perhaps the system was too complex to become an expert in such a relatively short time. In the experiment it was visible that even the 'expert users' did not know the hierarchies fully and completely and did not perform significantly faster than novice users. Most previous researchers used a smaller hierarchy (fewer items in total) to test with, in which case learning effects can be noticed faster. Some even allowed their test participants to memorize the whole hierarchy beforehand, which makes them system experts by default. In this experiment, a hierarchy was used with as many as 4096 items on the deepest level, which was only done before by Landauer and Nachbar (1985), who did not focus on expertise effects.

5.5 CATEGORY LABELS

The selections participants had to make in the experiment by Landauer and Nachbar (1985) were entirely based on interval decisions (like "in what interval does the number 2391 fall?"), which make target selection completely unambiguous. In real-life situations labels are generally not so clear; users will often make errors and have to backtrack. In terms of Pirolli and Card (1995), the scent of real-life labels can be misleading. The quality of category labels has a significant effect on the balance between depth and breadth, as demonstrated by Miller and Remington (2004). In the present experiment, category labels were not unambiguous, as can be seen in the fact that users did not always follow the optimal path (instead, they generally made some errors and had to go back and correct their error). This more closely resembles a real-life situation where category labels are not perfect and choices are ambiguous. The current experiment shows that participants preferred narrower hierarchies to broader ones in such an ambiguous situation.

5.6 CONCLUSIONS OF THE EXPERIMENT

All in all, the experiment shows that the most preferred hierarchy for use with mobile devices is one with only few items on each level. If many items have to be ordered, it is better to order them in a hierarchy with more levels (a deeper structure) than in a hierarchy with more items per level. These findings do not differ significantly for mobile devices with larger screen sizes. Users with larger devices will generally find items faster than users with smaller devices, because selection and navigation is easier on such larger devices, but this is independent of the hierarchy that is chosen.

This is good news for developers of future mobile applications and websites: Mobile users have clear wishes. They want structured information that is organized in narrow hierarchies, preferably with no more than 4 to 8 items per level. Contrary to desktop applications where many options are usually presented at once, it is better to use a layered design for mobile use. Options should be grouped into small categories, which in turn can be grouped into higher-order categories. Each level should not contain more than 8 items for optimal performance.

Users clearly prefer a compact, layered navigation structure to a broader structure where more options are visible at once. This applies to mobile devices in general (both to mobile phones and to PDAs); it is not necessary to differentiate between devices with varying screen sizes. Additionally, expertise does not seem to have a substantial effect on user preferences in this case.

6 Conclusions

This report started with one main question and four sub questions as the focus of the report. One literature review and one experiment later, it is time for checks and balances. The main question was formulated as follows:

"How can the usability of an information structure of an application or website be optimized for the special situation on mobile devices?"

A number of factors have been taken into account to address this problem. The literature review started with ways to structure information in information hierarchies, with a discussion on the merits of broader and narrower hierarchies (trees) with more items per level and fewer levels in broader hierarchies compared to narrower hierarchies. From this discussion on information hierarchies, the discussion moved to user search strategies and navigation models, which are important aspects of how the user navigates through an application or website. Taking the user's context into account where literally a whole world can influence the interaction between the user and the information system expanded this user discussion. Especially the technological means at the user's disposal at a certain time, the location of the user at that time and the user's identity and personal characteristics in general create the specific properties of an interaction. Most of these points have been discussed in the literature review, and the topic of technological means available, especially mobile devices, has additionally been lifted out as the subject of an experiment as described in the chapters 3, 4 and 5.

The main question of the study on usability of information structures was split into four sub questions. These questions were the foundation of this report, and an answer can be formulated for each of them here.

6.1 HOW CAN AN INFORMATION STRUCTURE HELP MOBILE USERS WITH INFORMATION NAVIGATION?

Any information structure can help users to find what they are looking for. But first, it is interesting to know how and why a user searches information. This search for information depends a lot on the task of the user. Is the user is looking for a specific item on a specific topic, then he or she will have the motivation to overcome high mental load, follow long paths and backtrack if necessary to eventually find the target. On the other hand there is the general-purpose browser who is interested in 'anything sounds interesting', who will drift away from complicated navigational structures and look for information that is easier accessible. These kinds of users (and the users in between these extremes) vary in their goals and their motivation levels, but are essentially both aided by a clear navigation structure.

The structure provides the user with a way to access the information. Such a structure can also have a negative side, which is mainly found in the way the structure is built up. A structure can cause navigation problems when the information is structured illogically or when the navigational labels are incorrect, incomplete or inappropriate. Then the question becomes how information *should* be structured in a logical, correct, complete and appropriate way. One way to limit the scope of search for a user is to provide a hierarchical organization of items. In this study it is assumed that target information is located in the end nodes of such a hierarchical structuring, as is often the case in user guides, product search and product information, applications with many functions, or websites with more than one navigational level. Structuring the information in such a hierarchy can be done by considering two important components: the exact configuration of the hierarchy of the information on the one hand and the naming of the labels of categories, sub categories and items on the other hand. The hierarchical structuring is divided into two interrelated factors, the depth and the breadth of the hierarchy. The breadth referred to the amount of items that are listed on a single level, the depth refers to the total amount. Literature based on both experiments and theoretical models for desktop use shows that with a broad hierarchy with moderately many items per level (with just a few levels) desktop users perform better than with a narrower hierarchy, as presented in chapter 2.2.1. So, an information structure that is organized in a broad way helps desktop users with their information navigation. Mobile users are a specific case, which was the subject of the experiment of chapter 3, 4 and 5. Mobile users are not aided by a broad hierarchy, but instead prefer a narrower hierarchy that allows them to see all available options on the screen at once, but that is more the subject of the third sub question based on the main question, below.

With the configuration of the hierarchical structuring of a website or application being the first component of information structure, the second component is the labeling of the categories. As described by Pirolli and Card (1995), the quality of such labels determines whether a user will find an item fast, slowly, or not at all. Navigation is an expected-cost (an additional click and loading time) versus expected-rewards (finding the target item) tradeoff, where the expectedness of the reward depends on the quality of the labels leading to the target item. The quality of labels is determined by many factors, where consistency is the most important aspect. Good and consistent labeling is valuable for both desktop and mobile users.

This combination of a well-chosen hierarchical structure with a well-chosen labeling system helps users with their information navigation, for use on both desktop computers and mobile devices.

6.2 HOW DO THE SPECIFIC PROPERTIES OF MOBILE DEVICES (SOCIAL AND PHYSICAL CONTEXT, TRANSFER SPEEDS) INFLUENCE THE USABILITY OF AN INFORMATION ARCHITECTURE?

Mobile devices are limited in their capabilities compared to desktop computers. The mobility of such devices means that the context in which the device is used can change, see also chapter 2.4, about the user’s context. The changing locations (physical context) of the user context means that the user is able to use the device under different conditions where issues like lighting and temperature play a role. The user’s social context can also change rapidly when using a mobile device: it can be used alone, but also amidst friends, or strangers. Both kinds of conditions place burdens on the external design of the device. When lighting is not available, the device display should light up, when the temperature is low and the user is wearing gloves, the device should still be manageable under these tougher conditions. When the user is among strangers, the viewing angle could be small to increase privacy; when among friends, the viewing angle could be larger to share the information on the screen with them. The influence of these aspects on the user’s task performance is large, but I expect they do not severely influence the usability of an information architecture, which however has not been the topic of any past or present literature.

Limited battery life and limited storage possibilities are also aspects at which mobile devices differ from their desktop counterparts, but these are unrelated to task performance in information structures. The factors screen size and input device do have an effect; these two factors are the topic of the question discussed below.

Finally, low transfer rates that accompany mobility do have a significant effect, which is directly related to the models of search as discussed in chapter 2.4.1. The “computer response time” is one of the factors that influence the total search time, which is less influential as the depth decreases. In other words, if the transfer times are long a deep structure leads to longer search times than a broad structure. The effects of screen size seem to reverse this effect, though.

6.3 HOW DO SCREEN SIZE AND INPUT METHOD INFLUENCE USERS’ PREFERENCE FOR A CERTAIN INFORMATION ARCHITECTURE?

The effects of screen size on the general usability of different information hierarchies have been investigated in the experiment that is described in the previous chapters 3, 4 and 5. Users had to use mobile devices with limited screen sizes to find target items on the lowest levels of these hierarchies. Time, errors and satisfaction were measured, as well as the number of key presses, the relative certainty of choice and the perceived complexity of the hierarchy.

The experiment shows that most users of the mobile devices preferred a hierarchy with only few items on each level: the ‘narrowest’ hierarchy in the experiment with 4 items per level and 6 levels. Unfortunately, the measurements in time, errors and certainty of choice varied too much between the subjects and conditions to obtain reliable results. The effect in users’ preferences however is quite strong, and independent of exactly which of three devices is used. Consequently, if many items have to be ordered, it is better to order them in a hierarchy with more levels (a deeper structure) than in a hierarchy with more items per level. Measurements of perceived complexity and number of key presses using the four different hierarchies point in the same direction; the narrowest hierarchy is seen as less complex than it actually is and requires the fewest key presses to reach the target item. The broadest hierarchy is seen as more complex than it actually is and needed the highest number of key presses to reach the target item. This is a very interesting outcome, because literature on desktop computer use shows that hierarchies with many items per level are generally preferred. It is in line with the expectations of the experiment though, which were based on the (literature on) limitations that small screens have.

The positive or negative effects of certain input methods (for navigation purposes, no text input was necessary throughout the experiment) were not directly investigated in the experiment. However, the devices that were used in the experiment did not have exactly the same input method. The largest of the three, the PocketPC, was equipped with a stylus, whereas the other two were equipped with a joystick for navigation. A difference in preferences between the devices was not statistically significant, but a small trend is noticeable where preferences shift somewhat towards broader hierarchies for the large device. More various possible input methods would have complicated the experiment design in an intolerable way, and were therefore not directly taken into account. Literature does show that pen navigation is faster than joystick navigation when there are many items on the screen (Mizobuchi, Mori, Ren & Michiaki, 2002). As long as the targets that have to be selected are not too small, users of a device that is equipped with a stylus should be able to work with a broader hierarchy without negative consequences on the usability.

Based on the performed experiment it is safe to say that screen size has an influence on users' preferences for a certain hierarchy, but only if the difference in screen size is large enough, like the difference between a mobile device and a desktop computer. The variation in screen size among mobile devices is relatively small; therefore no significant preference-effects between these devices showed up conclusively in the experiment. Input methods for navigation are claimed to have an effect as well, in that a pen or stylus allows for more items on the screen without hindering the user in his navigation.

6.4 WHAT IS THE INFLUENCE OF INCREASING EXPERTISE ON THIS PREFERENCE?

The influence of increasing expertise was the second topic of the experiment that was described in the previous chapters 3, 4 and 5. The available literature was not conclusive on this topic. Some authors argued that novices would perform better with broad hierarchies because they allow them to see many items simultaneously, avoiding navigation mistakes and lostness (Jacko, Salvendy and Koubek, 1995). Other authors claimed that novice users would perform better when using a narrow hierarchy, because such hierarchies would 'funnel' them in the right direction, thereby avoiding difficult decisions that would guide them in the wrong direction (Paap and Cooke, 1997).

Users were asked after the experiment to return two more times to do the experiment again. This allowed for a comparison between novice users (the first time), intermediate users (the second time) and expert users (the third time). Surprisingly, no significant differences were found between the preferences of the three kinds of users. A reason for this might be the relatively difficult hierarchy that was used, where even three sessions of an hour turned out to be not enough to qualify as an expert and to know the system thoroughly. Another reason could be that there *is* no difference between novice and expert users, but that all users prefer the narrower hierarchy on a mobile device.

All in all, it is still hard to say what the exact influence of increasing expertise is on the preference. Most users in the experiment preferred the narrowest hierarchy available, and a shift to a broader hierarchy was expected, but did not occur. Users still prefer the narrowest hierarchy after three hours of use, which speaks in favor of this hierarchy in any case.

Everything taken together, the main conclusion of this study is that narrower hierarchies perform better on mobile devices than larger hierarchies. Other factors have and might have influenced this performance (user factors like expertise, context factors like the laboratory setting in which the experiment was performed, or technological factors like the specific properties of the device used); in the end, the laboratory study shows that users prefer the smallest hierarchy – four items per level, six levels deep – over the other hierarchies that were tested.

7 Suggestions for Future Research

When doing research in a laboratory setting, one does not have the possibility to vary many variables, as the complexity of the experiment design would rapidly increase beyond control. A lab experiment is therefore always limited in its scope. This experiment is limited in the sense that “only” four hierarchies were tested, on “only” three devices, with “only” fiftheen participants, on “only” three occasions, on “only” one location. Nonetheless, the variance between conditions and between subjects was already so large that it was not always possible to make distinctions between each and every condition. Testing more variables would further increase this variance, making it even more difficult to find reliable results. Noting that the scope of the experiment is limited, there are endless possibilities of extending this research with further tests and experiments that would give more insight in the use of information structures on mobile devices.

These possibilities are indirectly already mentioned in the text above, where the word ‘only’ indicates a possibility for extending the experiment. Most interestingly would be to experiment with exactly the same hierarchies on a wider range of devices or at least a wider range of screen sizes. The screen sizes in the current experiment ranged from a display capable of showing 4 lines of text to a display capable of showing 16 lines of text. It would be very interesting to see how the same hierarchies would perform on a even larger display, for example on a regular desktop computer which is typically able to show 40 to 50 lines of text.

A second extension to the experiment could be to test with different hierarchical structures. Right now, four hierarchies were tested, but this is still only a fraction of all the possible hierarchies that exist. The most obvious additional hierarchy that could be tested would be a binary hierarchy, where every level exists of only two items. Since the participants in the present experiment indicated that they prefer a narrow hierarchy to a broader one, it seems logical to test whether this continues into the extreme case of a binary hierarchy, 2^{12}. Another possibility would be to test hierarchies that lay in between the now-preferred hierarchies. Users currently prefer narrow hierarchies, with 4 to 8 options per level. An experiment with more hierarchies, each with 8 or less items per level, would further specify the exact place of the optimum. Such an experiment design would be difficult to test though: the current experiment already shows a lot variance that is independent of the conditions, making it hard to find significant results. When the hierarchies are even closer to another than they were in the current experiment, the results will also lay closer to another and a difference would be even harder to detect. Hierarchies with between 4 and 8 options would also introduce a (minor) technical problem in that a symmetrical distribution of the items over all levels of the hierarchy would not yield an exact number of 4096 items in total. 5 Items per level for example would only be possible if it is sometimes mixed with 4 or 6 items per level, which biases the outcomes of such an experiment. So far, all hierarchies that were described were completely symmetrical and the number of items was equal on each level of the hierarchy. What would also be interesting is to test non-equal distributions in such hierarchies as well. Norman and Chin (1988) performed such an experiment on the desktop, in which they used a convex structure (more items in the middle, fewer items at the beginning and at the end) and a concave structure (fewer items in the middle, more at the beginning and at the end). This led to positive results in their experiment, and it might be interesting to see how such structures would perform on mobile devices.

Another very interesting extension of the current experiment would go in the direction of field studies. When an experiment is performed in a lab setting, users are typically biased by the strange surroundings that they are in. The sheer amount of cameras, microphones, computers, et cetera can be quite impressing and influence the participants so that they do not behave the same way as they would normally. Of course, the presence of a test leader and the fact that they are asked to perform very specific tasks (in exchange for money) also change their motivation and expectations towards the tasks and the goals. In a follow-up study, testing out the devices "in the real world" could reduce this laboratory effect, where the surroundings of the user are more natural and his or her activity is less obviously monitored. This would improve the validity and increase the generalizability of the results of the experiment. Additionally, such answers could be answered as "does reduced lighting, or temperature, or loud noise, changing social contexts, or any other context factor, influence the user's preference for a certain hierarchy?" Context effects are generally difficult to study in the lab, especially when the focus is on social context. Although it is predicted that these factors will not have an effect on the preferred information structure, this could not be based on existing literature or experimental results, which means such an experiment would certainly be interesting, even if the results are negative.

Finally, item labeling might be an interesting topic for follow-up studies. The importance of the quality of item labeling should not be underestimated. In the present experiment, the quality of the labeling was kept as equal as possible between the four hierarchies. This was done after restructuring the hierarchies, where items lost some form of context because they lost their headings. The items were then renamed so that each label was self-explanatory even without the context of the heading that used to accompany it. This way, although the quality might still not be one hundred percent equal for all items, the quality is approximately the same throughout the four hierarchies. An interesting follow-up experiment could test an information hierarchy under the condition that the quality is *not* the same for all items. This makes it more difficult for users to find the target item and, according to Miller and Remington (2004), also has a varying influence on search times depending on the structure that is used. Such a test of different label qualities might give interesting results, especially in the case of mobile devices where users' workload already increases due to other limiting factors.

All in all, there are many ways in which the experiment described in this study can be extended. This experiment gives a good first impression of how mobile users navigate through information structures, but the literature on this topic is still far from complete and there is place enough to increase the knowledge about mobile users and their specific needs.

8 References

Albers, M.J. & Kim, L. (2000). User Web Browsing Characteristics Using Palm Handhelds for Information Retrieval. *Proceedings of IEEE professional communication society international professional communication conference and Proceedings of the 18th annual ACM international conference on Computer documentation: technology & teamwork*, 125-135.

Bederson, B.B., Clamage, A., Czerwinski, M.P. & Robertson, G.G. (2004). DateLens: A Fisheye Calendar Interface for PDAs. *ACM Transactions on Computer-Human Interaction*, 11(1), 90-119.

Björk, S. & Redström, J. (1999). An Alternative to Scrollbars on Small Screens. *CHI '99 extended abstracts on Human factors in computing systems,* 316-317.

Brewster, S. (2002). Overcoming the Lack of Screen Space on Mobile Computers. *Personal and Ubiquitous Computing*, 6, 188-205.

Buchanan, G., Farrant, S., Jones, M. & Thimbleby, H., Marsden, G., Pazzani, M. (2001). Improving Mobile Internet Usability. *Proceedings of the Tenth International World Wide Web Confere*nce, 673-680.

Buyukkokten, O., Garcia-Molina, H., Paepcke, A. & Winograd, T. (2000). Power Browser: Efficient Web Browsing for PDAs. *Proceedings of the SIGCHI conference on Human factors in computing systems 2000, 430-437.*

Buyukkokten, O., Garcia-Molina, H. & Paepcke, A. (2001a). Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones. *Proceedings of the SIGCHI conference on Human factors in computing systems 2001*, 213-220.

Buyukkokten, O., Garcia-Molina, H. & Paepcke, A. (2001b). Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. *Proceedings of the 10th international conference on World Wide Web.*

Card, S.K. (1982). User perceptual mechanisms in the search of computer command menus. *Proceedings of the 1982 conference on Human Factors in computing systems*, 190-196.

Catledge, L. D. & Pitkow, J. E. (1995). Characterizing Browsing Strategies in the World Wide Web. *Proceedings of the Third International World-Wide Web Conference, Computer Networks and ISDN Systems*, 27(6), 1065-1073.x

Choo, C.W., Detlor, B. & Turnbull, D. (2000). Information Seeking on the Web: An Integrated Model of Browsing and Searching. *First Monday*, 5(2), www.firstmonday.org

Clarke, J.A. (1993). Cognitive style and computer assisted learning: Problems and a possible solution. *Association for Learning Technology Journal*, 1, 47–59.

Consensus Project (2003). Application programming guidelines for the complexity of applications on device class level, 3G Mobile Context Sensitive Adaptability – User Friendly Mobile Work Place for Seamless Enterprise Applications (CONSENSUS). *IST-2001-32407*

Cox, T. & Griffith, A. (1995). The Nature and Measurement of Work Stress: Theory and Practice. in *Evaluation of Human Work: A Practical Ergonomics Methodology (Second Edition)*, Wilson, J.R. & Corlett, E.N. (eds.), Philadelphia: Taylor & Francis Inc., 783-803.

Dey, A. K. and Abowd, G. D. (2000). Towards a better understanding of context and context-awareness. *Computer Human Interaction 2000 Workshop on the What, Who, Where, When, Why and How of Context-Awareness*.

Dray, S. M., Ogden, W. G., & Vestewig, R. E. (1981). Measuring Performance with a Menu-Selection Human-Computer Interface. *Proceedings of the Human Factors Society, 25th Annual Meeting*, 746-748.

Duchnicky, R.L. & Kolers, P.A. (1983). Readability of text scrolled on visual display terminals as a function of window size. *Human Factors*, 25, 683-692.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*, 381-391.

Ford, N. (1985). Learning styles and strategies of postgraduate students. *British Journal of Educational Technology*, 16, 65– 77.

Ford, N., Miller, D., & Moss, N. (2005). Web search strategies and human individual differences: Cognitive and demographic factors, Internet attitudes, and approaches. *Journal of the American Society for Information Science and Technology*, 56(7), 741-756

Ford, N., Wilson, T.D., Foster, A., Ellis, D. & Spink, A. (2002). Information seeking and mediated searching. Part 4: cognitive styles in information seeking. *Journal of the American Society for Information Science and Technology, 53(9), 728-735.*

Giller, V., Melcher, R., Schrammel, J., Sefelin, R & Tscheligi, M., (2003). Usability Evaluations for Multi-device Application Development Three Example Studies. *Human-Computer Interaction with Mobile Devices and Services, Proceedings of the 5th International Symposium, Mobile HCI 2003*, 302-316.

HEROLD Yellow Pages Austria, http://www.herold.at.

Hollands J.G. & Merikle, P.M. (1987). Menu organization and user expertise in information search tasks. *Human Factors*, 29(5), 577-586.

Hölscher, C. & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks, 33*, 337–346.

Hyman, R. (1953). Stimulus Information as a Determinant of Reaction Time. *Journal of Experimental Psychology*, 45, 188-196.

Jacko, J.A. & Salvendy, G. (1996). Hierarchical menu design: Breadth, depth, and task complexity. *Perceptual and Motor Skills*, 82, 1187-1201.

Jacko, J.A., Salvendy, G. & Koubek, R.J (1996). Modeling of menu design in computerized work. *Interacting With Computers*, 7(3), 304-330.

Järvelin, K. & Ingwersen, P. (2004). Information seeking research needs extension towards tasks and technology. *Information Research*, 10(1), http://informationr.net/ir/10-1/paper212.html.

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K. & Buchanan, G. (1999). Improving Web interaction on small displays. *Proceeding of the eighth international conference on World Wide Web*, 1129-1137.

Jones, M., Buchanan, G. & Thimbleby, H.W. (2002). Sorting Out Searching on Small Screen Devices. *Proceedings of the 4th International Symposium on Mobile Human-Computer Interaction*, 81-94.

Kamba, T., Elson, S.A., Harpoid, T., Stamper, T. & Sukaviriya, P. (1996). Using small screen space more efficiently. *Proceedings of the SIGCHI conference on Human Factors in computing systems: common ground*, 383-390.

Khan, K., & Locatis, C. (1998). Searching through cyberspace: The effects of link display and link density on information retrieval from hypertext on the World Wide Web. *Journal of the American Society for Information Science*, 49(2), 176–182.

Kiger, J.I. (1984). The depth/breadth trade-off in the design of menu-driven user interfaces. *International Journal of Man-Machine Studies,* 20, 201-213.

Krug, S. (2000). *Don't make me think!: a common sense approach to Web usability*. Boston: Pearson Custom Publishing.

Landauer, T.K. & Nachbar, D.W. (1985). Selection from alphabetic and numeric menu trees using a touch screen: breadth, depth, and width. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 73-78.

Lazander, A.W., Biemans, H.J.A., & Wopereis, I.G.J.H. (2000). Differences between novice and experienced users in searching information on the World Wide Web, *Journal of the American Society for Information Science*, 51(6), 576-581.

Lee, E. & MacGregor, J. (1985). Minimizing User Search Time in Menu Retrieval Systems. *Human Factors*, 27 (2), 157-162

Lee, E., MacGregor, J., Lam, N. & Chao, F. (1986). Keyword-menu retrieval: an effective alternative to menu indexes. *Ergonomics*, 29(1), 115-130.

MacGregor, J., Lee, E. & Lam, N. (1986). Optimizing the structure of database menu indexes: A decision model of menu search. *Human Factors*, 28(4), 387-399.

MacKay, B. & Watters, C. (2003). The Impact of Migration of Data to Small Screens on Navigation. *IT & Society, 1(3),* 90-101.

MacKenzie, I.S., Sellen, A. & Buxton, W. (1991). A Comparison of input devices in elemental pointing and dragging tasks. *Proceedings of the SIGCHI conference on Human Factors in computing systems: Reaching through technology*, 161-166.

Marsden, G., Cherry, R. & Haefele, A. (2002). Small Screen Access to Digital Libraries. *CHI '02 extended abstracts on Human factors in computing systems,* 786-787.

Miller, C.S. & Remington, R.W. (2004). Modeling Information Navigation: Implications for Information Architecture. *Human-Computer Interaction, 19*, 225-271.

Miller, D.P. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Proceedings of the Human Factors Society 25th Annual Meeting*, 296-300.

Mizobuchi, S., Mori, K., Ren, X. & Michiaki, Y. (2002). An Empirical Study of the Minimum Required Size and the Minimum Number of Targets for Pen Input on the Small Display. *Proceedings of the 4th International Symposium on Mobile Human-Computer Interaction,* 184-194.

Nielsen (2000). WAP Backlash. Online at http://www.useit.com/alertbox/20000709.html

Norman, K.L. (1991). *The Psychology of Menu Selection: Designing Cognitive Control of the Human/Computer Interface*. Norwood: Ablex Publishing Corporation.

Norman, K. L. & Chin, J. P. (1988). The Effect of Tree Structure on Search in a Hierarchical Menu Selection System. *Behaviour and Information Technology*, 7, 51-65.

Norman, D. (1988). *The Psychology of Everyday Things*. New York: Basic Books.

Osborne Rao, D. (2000) A study of input devices on personal digital assistants (PDAs), Serco Usability Laboratory, available at http://www.usability.serco.com/research/research.htm

Paap, K.R. & Cooke, N.J. (1997). Design of menus. In Helander M., Landauer, T. K. & Prabhu, P. (Eds.). *Handbook of Human Computer Interaction (second edition)*. Amsterdam: North-Holland, 533- 572.

Paap, K.R. & Roske-Hofstrand, R.J. (1986). The optimal number of menu options per panel. *Human Factors*, 28(4), 377-385.

Palmquist, R.A. & Kim, K.S. (2000). Cognitive Style and On-Line Database Search Experience as Predictors of Web Search Performance. *Journal of the American Society for Information Science*, 51(6), 558 –566

Pask, G. (1976). Styles and strategies of learning. *British Journal of Educational Psychology*, 46, 128-148.

Pierce, B.J. Parkinson, S.R. & Sisson, N. (1992). Effects of semantic similarity, omission probability and number of alternatives in computer menu search. *International Journal of Man-Machine Studies*, 37, 653-677.

Pierce, B.J., Sisson, N. & Parkinson, S.R. (1992). Menu Search and Selection Processes: a Quantitative Performance Model. *International Journal of Man-Machine Studies*, 37, 679-702.

Pirolli, P. & Card, S.K. (1995). Information foraging in information access environments. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 51-58.

Resiel, J.F. & Shneiderman, B. (1987). Is bigger better? The effects of display size on program reading. In: Salvendy, G. (Ed.), *Social, Ergonomic and Stress Aspects of Work with Computers*. Amsterdam: Elsevier,113-122.

Riding, R.J., & Cheema, I. (1991). Cognitive styles – An overview and integration. *Educational Psychology*, 11, 193-215.

Rosenfeld, L. & Morville, P. (2002). *Information architecture for the World Wide Web*. Sebastopol, O'Reilly & Associates, Second edition.

Saracevic, T. (1996). Modeling interaction in information retrieval (IR): A review and proposal. *Proceedings of the American Society for Information Science*, 33,3-9.

Searchenginewatch.com (2004). "Search Engine Size Wars V Erupts". Online at http://blog.searchenginewatch.com/blog/041111-084221.

Snowberry, K., Parkinson, S.R. & Sisson, N. (1983). Computer display menus. *Ergonomics*, 26(7), 699-712.

Spielberger, C. D. (1983). *State-Trait Anxiety Inventory (Form Y)*. Palo Alto: Mind Garden.

Swierenga, S.J., (1990). Menuing and scrolling as alternative information access techniques for computer systems: interfacing with the user. *Proceedings of the Human Factors Society 34th Annual Meeting*, 356-359.

Sternberg, S. (1969). High-speed scanning in human memory. *Science*, 153, 652-654.

Tullis, T.S. (1985). Designing a menu-based interface to an operating system. *Proceedings of the SIGCHI conference on Human factors in computing systems,* p. 73-78.

Vakkari, P. (1999). Task complexity, problem structure and information actions, Integrating studies on information seeking and retrieval. *Information processing and Management,* 35, 819-837.

Wallace, D.F., Anderson, N.S., Shneiderman, B. (1987). Time stress effects on two menu selection systems. *Proceedings of Human Factors Society, 31st Annual Meeting*, 727-731

Wang, P., Hawk, W.B. & Tenopir, C. (2000). Users' interaction with World Wide Web resources: an exploratory study using a holistic approach. *Information Processing and Management, 36*, 229-251.

Wickens, C.D. (1999). Automation in air traffic control: The human performance issues, in M.W. Scerbo and M. Mouloua (Eds.) *Automation Technology and Human Performance: Current Research and Trends*, Hillsdale: Lawrence Erlbaum, 2-10.

Wilson, T.D. (1997). Information behaviour: an interdisciplinary perspective. *Information Processing and Management*, 33(4), 551-572.

Wilson, T.D. (1999). Exploring models of information behaviour: the ‘uncertainty’ project. *Information Processing & Management*, 35(6), 839-849.

Wobbrock, J.O., Forlizzi, J., Hudson, S.E. and Myers, B.A. (2002). WebThumb: Interaction Techniques for Small-Screen Browsers. *Proceedings of the 15th annual ACM symposium on User interface software and technology*, 205-208.

Zaphiris, P.G. (2000). Depth vs. Breadth in the Arrangement of Web Links. *Proceedings of the Human Factors Society 44th Annual Meeting,* (1) 453-456.

9 Appendix

The tasks that were used in the first week:

1. Ich brauche noch Betten für die neue Wohnung.
2. Ein Wäschetrockner würde mir viel Zeit und Platz im Haushalt sparen.
3. Ich brauche für den Winter noch neue Eislaufschuhe.
4. Wenn man viel Post verschicken muß, sind Frankiermaschinen die Lösung.
5. Ich suche einen neuen Rechner und zwar ein Komplettsystem.
6. Ich möchte mir ein Blutdruckmessgerät kaufen.
7. Beim Bergsteigen sind Seile unbedingt notwendig.
8. Ich suche neue Wasserpflanzen für meinen Teich.
9. Ich suche ein Facharzt in Richtung Angiologie.

The tasks that were used in the second week:

1. Auf meiner Party wird ein Tonanlage gebraucht.
2. Ich möchte gerne mehr wissen über Gebühren & Entgelte von Abfallsammlung.
3. Ich brauche Düngemittel für meinen Garten.
4. In meiner Wohnung möchte ich gerne die Tische decken mit Echtsilberbesteck.
5. Für meinen Urlaub suche ich Campingplätze wo ich ein Paar Nächte wohnen kann.
6. Wenn ich Tauchen gehe, brauche ich einen neuen Tauchanzug.

And the tasks that were used in the third week:

1. Falträder sind viel mobiler und in der Stadt auch noch viel praktischer als Autos.
2. Ein wichtiges Teil des Reiten ist die Pferdepflege. Dafür braucht man sogenannte Pferdeprodukten.
3. Ich möchte gern mal in einen Kaffeerestaurant gehen.
4. Ich möchte gerne mehr wissen über die Pappe die Buchbinder gebrauchen, Buchbinderpappe.
5. Ich suche ein neues Dampfbügeleisen.
6. Ich möchte gerne ein Kamin einbauen. Daher möchte ich zuerst mal mehr wissen über dem Kaminbau.