Usability Evaluations for Multi-Device Application Development
Three Example Studies

Verena Giller, Rudolf Melcher, Johann Schrammel,
Reinhard Sefelin, Manfred Tscheligi

CURE Center for Usability Research & Engineering,
Hauffgasse 3-5,
1110 Vienna, Austria
cure@cure.at

Abstract. This paper discusses three example studies, that informed user interface guidelines, developed for a set of different classes of mobile devices. The results of these studies show answers to typical design problems arising during the development of mobile applications. Furthermore the studies are meant to be examples showing which kind of studies are required in order to develop a sufficient pool of user interface guidelines covering almost all sorts of mobile devices.

1 Introduction

This paper presents the results of three example studies, which compare different ways of navigation, selection and interaction implemented on five classes of mobile devices. These studies were part of an EU funded project (CONSENSUS [3]). CONSENSUS develops a mark-up language supporting the automatic adaptation of user interfaces for mobile devices. The mark-up language development process, however, will not be discussed in this paper. The studies discussed in our paper are three examples of a large number of empirical studies, which informed the development of user interface guidelines on which the adaptation process is based.

As one important prerequisite of these studies a device classification was developed, which enabled us to draw generic conclusions from our studies. The classification of the devices was based on an analysis focusing on those device characteristics that have an influence on users' behaviour and on their perception of the user interface. This analysis led to three main dimensions defining the eight device classes, which we finally developed. These dimensions are: presentation structure, supported input modality, mark-up language. Van Welie and de Groot [11], who also developed a device classification, used a very similar approach, which was also based on these three dimensions.

The classification on which we based our empirical studies included eight classes. The classes ranged from class 0 (speech input and output only) to class 7 (laptop-PC).

Speech interfaces were included because in a lot of mobile contexts (take for example car driving) hands-free interaction is required. In this paper only studies with devices of the classes 0 to 4 will be discussed.

The first study, which we discuss in this paper compares different ways of text presentation. The second study investigated the optimal depth and breadth of trees enabling users to navigate through content structures. Finally the third study discusses how we defined optimal and maximal numbers of list entries for two different kinds of speech-lists. The studies were conducted with one device of class 1 (typical mobile WAP-phones; representative: SIEMENS, ME45), one of class 2 (mobile phones with large colour displays; representative: NOKIA, 7650) two different devices of device class 3 (a COMPAQ: iPAQ Pocket PC (browser type: Konqueror) and a handspring treo-Communicator (browser type: Blazer) and one of class 4 (clamshell devices; representative: NOKIA, Communicator 9210i). Our third study dealt with speech interaction via fixed line telephones or mobile phones (class 0).

2 Optimising text presentations for reading tasks

2.1 Motivation

Today’s web guidelines, which are dealing with text reading come to the conclusion that texts should be as short as possible but that users will scroll content pages if they expect further information which is relevant to their tasks (see e.g. [6], page 115 and [9], page 77). The splitting of texts into chunks of two or more pages should be avoided.

Screens of mobile devices, however, are much smaller than the screens for which “ordinary” web pages are designed. These differences may lead to different rules regarding the optimal lengths of pages for these devices. Since the handling of screen elements like scroll bars displayed on mobile devices cause more effort compared to the handling of such elements on a computer screen, users might tend to prefer a pagination mechanism. Therefore our hypotheses was that mobile devices are closer to the book than to the desktop computer and need therefore a metaphor of turning pages rather than the one of a paper roll.

These and other considerations led us to the set up of this study. We wanted to answer the question, whether users prefer a pagination mechanism or a scrolling mechanism to read longer texts. Furthermore we wanted to investigate whether these preferences are device class dependent.

The study was necessary because at the moment there is no empirical data available, dealing with these questions. Although Buchanan et al. [1] compare different possibilities to scroll though lists of headlines, these tests are of small relevance for our questions. Firstly this study is not dealing with longer texts and secondly it is using interaction techniques, which are not supported by today’s state of the art wap-browsers.

Another goal of this study was to see whether line breaks inside words displayed on devices of device class 1 and 2 can increase users' reading speed and whether they are reducing their subjective reading satisfaction.

2.2 Methodology

Devices of device class 1 to 4 were used for this study. The representative of class 3 was a "COMPAQ: iPAQ Pocket PC". The test sessions started with a briefing phase where demographic variables as well as variables concerning users' experiences with the four device classes were gathered. Subjects who were not familiar with device class 2 and 4 got a special introduction of the handling of these two device classes. All subjects were familiar with device classes 1 and 3.

The test of each device class started with the reading of a sample text. This sample text gave users the possibility to get used to the devices and to the task of reading a text with it. After that we started the test phase during which users' reading speed, their comprehension of the text and their reading satisfaction were measured. With the devices of device class 1 and 2 we compared pagination with scrolling and we compared also texts, which included line breaks inside words, with texts, which did not include such line breaks.

After subjects had completed a reading task (the reading time was measured in seconds), we asked them three questions concerning the content of the text. This enabled us to measure whether they had understood the text and, more important, we forced our subjects to read carefully and to make sure that they get the text's main messages.

The alternative display method was presented to the subjects directly afterwards. That means that a subject who was first confronted with a text, which he/she had to scroll, was then confronted with another text, which had to be paged through. A special index (see [4]) made sure that the complexity and readability as well as the number of words of both texts was almost equal. Texts used for tests with mobile phones included less words and characters than those used for tests with PDAs (mobile phones: approximately 400 words, PDAs: approximately 550 words). The texts were in German. We used modified articles of an Austrian newspaper about economical topics. The texts were slightly changed in order to make sure that they are fulfilling our requirements.

After the user had finished this alternative text (again the reading time was measured) the same procedure as described above was repeated. Finally we asked users to draw a line for each of the two possibilities. The length of the line should express their reading comfort. Users who were satisfied with the pagination method but not with the scrolling method drew a longer pagination-line compared to the scrolling-line and vice versa. The sheet on which users should draw the line included a clear marker for the starting point of the line that users should draw. These lines enabled us to measure the exact proportion of users' assessments of both methods.

This procedure was repeated for all the six alternative text displays on our four devices. The test design made sure that order and learning effects could not occur. Therefore the orders of devices and of text presentation styles were randomised. A short qualitative interview was conducted at the end of each session.

The pagination was implemented in the following way: At the end of the last line of the page/card three dots were displayed. Users then had to use a next button to go to the next page/card. Breaks inside of sentences only occurred once per text when the texts were displayed on devices of classes 3 and 4. The chunks of text displayed on devices of class 1 and 2 were implemented as single cards rather than as a whole deck. This decision was based on the assumption that users prefer to wait relatively short periods of time for each chunk and that loading the whole deck at once would lead to a too long waiting period.

The line breaks inside the words did not include hyphenation because this feature currently is not supported by state of the art WAP-browsers.

2.3 Results

10 subjects participated in our test sessions. 4 of them were female and 6 male. Their average age was 27.5 years (Std. Dev.: 7 years). All subjects were experienced users of device class 1 and 3.

Figure 1 shows a comparison of users' reading speed with the four device classes. On the left diagram scrolling is compared with pagination. The figure shows that big differences occur only when devices of device class 1 were used.

For the statistical analysis we used a two-way ANOVA with device class and presentation style (scrolling vs. paging) as within-subjects factors. Both main effects as well as their interaction were significant (Device classes: F(3.27)=7.03, p=0.001; Presentation styles: F(1.9)=21.78, p=0.001; Interaction: F(3.27)=30.77, p<0.0005). Due to the different text lengths the main effect of the device classes cannot be interpreted unambiguously. Post hoc analysis of the simple main effects showed that the time-difference between scrolling and paginating are only significant for device class 1 (F(1.9)=108,4, p<0.0005. The reason for this might be that users who were confronted with devices of device class 1 and who had to page through the texts had to deal with relatively high total loading times. Since one card only displayed a rather small number of characters (approximately 45) a relatively high number of cards had to be loaded in order to read the whole text.

Figure 2 shows the subjects' reading comfort expressed in line lengths.

Again a two-way ANOVA was used to analyse the data. The main effect of the presentation styles (F(3.27)=15.12, p=0.004) and the interaction effect (F(1.9)=7.42, p=0.001) were significant, whereas the main effect of the device classes was not (F(3,27)=2.64, p=0.07). A post hoc analysis of the simple main effects showed, that similar to the time results the difference between scrolling and pagination is only significant for device class 1 (F(1.9)=30.69, p<0.0005).

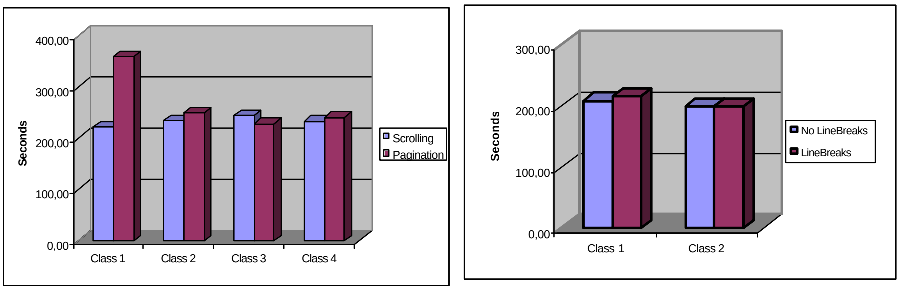

Fig. 1. Mean reading speed in seconds: left: Comparison of pagination and scrolling for device classes 1-4; right: Comparison of line breaks inside words and no such line breaks for device classes 1 and 2

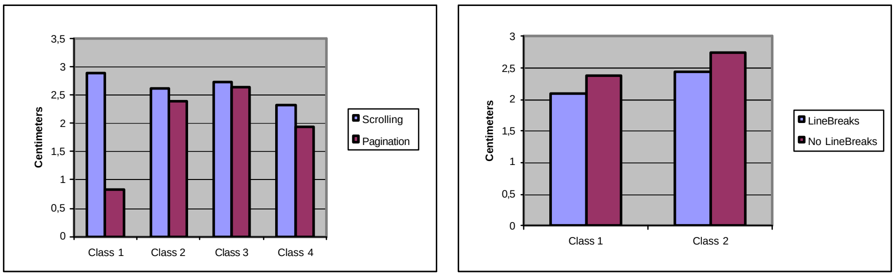

Fig. 2 Mean reading comfort in centimeters: left: Comparison of pagination and scrolling for device classes 1-4; right: Comparison of line breaks inside words and no such line breaks for device classes 1 and 2 (the longer the line the higher the expressed satisfaction)

The results of this study lead to the conclusion that in general a designer should have good reasons to use a pagination mechanism instead of a scrolling one. This is especially true when he/she is designing for devices of class 1.

The tests were conducted with typical content pages, which did not contain interaction elements. Our hypothesis was that possibly users prefer to page through the text displayed on mobile devices because the handling of elements that are supporting scrolling may be more difficult than the handling of these elements with a typical desktop PC. Another advantage of pagination might have been that scrolling easily can lead to a loss of orientation between the lines and that, on the other hand, the pagination mechanism provides users with pages whose content does not move.

However, the empirical data gathered could not prove that these two possible advantages outperform the disadvantages of long loading times and of an interaction mechanism, which is not consistent with today's web sites. That means that as long as loading times are at today's level there seem to be no clear reasons for pagination. Some user statements collected during our tests also support this. So a lot of users complained that the loading times are too long and that they are often loosing the plot of the texts. Another advantage of the scrolling concept, which was often mentioned by our subjects, was that scrolling gives them more control over the text. So one user said, "When

I am scrolling I can control the speed. On the other hand the pagination thing gives me the feeling that the computer controls how fast I am allowed to read."

Future studies will have to prove whether the results of device class 1 and 2 change if the chunks of text are implemented as decks rather than as single cards.

Figure 1 and Figure 2 (right pictures) show that the differences, which are due to line breaks inside words, are relatively small. Moreover the reading speed differences vary between the two device classes. On the other hand we see that these differences are consistent over both measurements of the users' reading satisfaction. A two-way repeated measures ANOVA results in a significant main effect for the factor line-break (F(1.9)=5.98, p=0.037). That means that line breaks inside of words lead to a lower reading satisfaction.

3 Content structures: Depth versus Breadth

3.1 Motivation

Navigation is one of the most critical factors of user interface design. A very important aspect of navigation is its structure, which is determined by the number of levels (depth) and by the number of items per level (breadth). In this context the question arises, whether it is better to offer a deep structure with few items per level or a broad one with many items per level. In the literature there are several recommendations available concerning this issue, but they mainly refer to desktop systems (see e.g. [6]).

This study aimed at estimating the influence of navigation structures on the searching performance and on the subjective satisfaction of users.

Our hypothesis was that all items on the same level should be perceptible at a glance, without forcing users to scroll. Therefore the optimal breadth would be determined by the screen size. We estimated the optimal depth of the navigation structure of device class 1 and 2 on the basis of available WAP-guidelines (see e.g. [10]). Regarding device class 3 and 4 we assumed that the structure should not be more than four levels deep. Starting from the premises mentioned above we defined an assumed optimal structure for each device class (see the grey coloured fields in Table 1).

	Depth	Breadth		Depth	Breadth
Class 1 ST1	4	3	**Class 3 ST1**	4	12
Class 1 ST2	2	6	**Class 3 ST2**	2	24
Class 1 ST3	3	4	**Class 3 ST3**	8	6
Class 2 ST1	3	8	**Class 4 ST1**	4	12
Class 2 ST2	2	12	**Class 4 ST2**	2	24
Class 2 ST3	6	4	**Class 4 ST3**	8	6

Tab. 1. Tested structures per device class. Variations in breadth and depth (ST1, ST2, ST3). The grey fields indicate the assumed optimal structures

3.2 Methodology

The goal of this study was to compare the assumed optimal structure with two alternative structures differing in depth and breadth. We measured users' searching performance and their subjective satisfaction.

Most of the items used for the different structures were terms of yahoo's content classification. To get reliable data, users had to perform three different search tasks per structure. Subjects were asked to find different items at the deepest levels of the structure. Two target items where located in the same main path and one in a second main path. That means that to reach the second target item users had only to go back to the middle of the first path, before they could enter the correct sub-path leading to this second item.

After each task users had been asked to estimate the complexity of the navigation on a 5 point rating scale. After each device users had to compare the three different structures in terms of their navigation- and selection-comfort. Again we asked the subjects to draw a line for each structure. In this case the sheet given to the subjects included a clearly defined starting point and a line representing users' subjective threshold of pain. Users should express their comfort relative to this threshold. (This threshold-line is also represented in the right hand picture of Figure 3.) Therefore, in contrast to the first study, in this case, a shorter line meant higher comfort and vice versa. Additionally we conducted a short qualitative interview where we asked users to explain their preferences. Again, the tests started with a briefing session.

Devices of device class 1 to 4 were used for our tests. The representative of class 3 was a "handspring treo-Communicator". Devices and structures had been randomised between subjects to avoid order and learning effects. Note that the items of the structure "ST1" on class 4 were displayed side by side in order to use the whole screen real estate.

3.3 Results

10 subjects participated in these sessions. 5 of them were female and 5 male. Their average age was 26 years (Std. Dev.: 5 years). Figure 3 (left picture) shows the average satisfaction ratings of the different types of structures per device class. The higher the

rating the higher was the user satisfaction. As expected, with the exception of device class 3, there is a small trend towards a preference of the structure “ST1”.

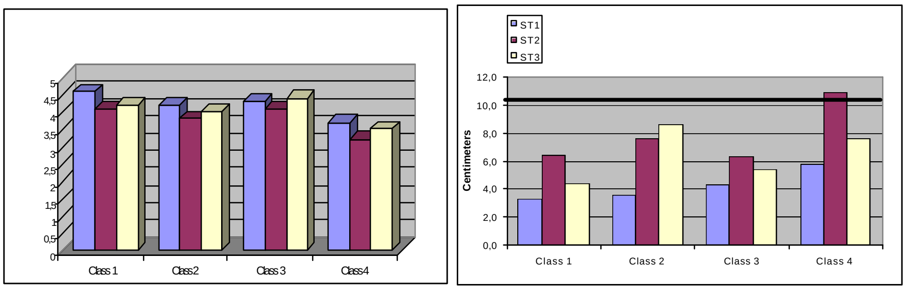

Fig. 3. Left: Average user satisfaction per device. (1=uncomfortable; 5=comfortable); right: User satisfaction expressed in line lengths (the shorter the line the higher the user satisfaction

The right picture of Figure 3 shows users’ relative ratings expressed in line lengths. The shorter the line, the higher was the users´ satisfaction. Here the differences of users’ preferences are stronger. Statistical analysis showed significant main effects for both, device classes (F(3.27)=3.16, p=0.041) and used structure (F(2.18)=4.52, p=0.026). The interaction effect did not show a significant result (F(6.54)=0.069 p=0.662).

To explore these results in detail we performed post hoc comparisons of the main effects of the three structures. We observed a significant difference between the structures “ST1” and “ST2” (F(1.9)=21.38, p=0.001). No significant difference could be observed between “ST2” and “ST3” (F(1.9)=0.90, p=0.367) and between “ST1” and “ST3” (F(1.9)=2.80, p= 0.129).

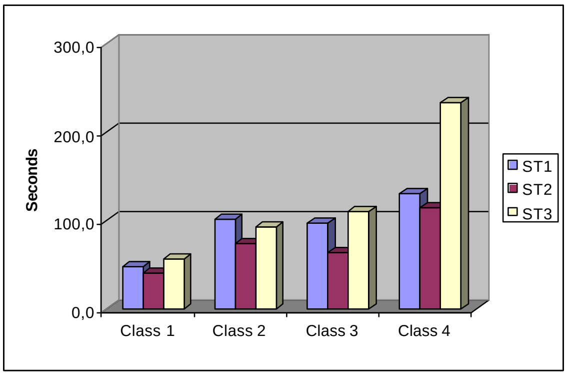

Fig. 4. Averaged search performance per device and structure (in seconds)

Finally we calculated the mean search performance per device and structure (see Figure 4). In all the cases with the exception of device class 2 the most time consuming structure was “ST3”. The outlier (device class 2) was due to a semantic problem (most of the subjects initially searched in the wrong category). The big differences of the search performances of device class 4 can be attributed to the deep structure and also to some semantic problems. Note that the total number of items differed between the structures presented to our users. Therefore, the search performances can only be interpreted as possible explanations of users' preferences but not as a source of recommendations of the optimal structure.

Quantitative as well as qualitative data confirm our assumption that all items on the same level should be perceptible at a glance. Although the structure "ST2" was faster in terms of task performance (see Figure 4), subjects preferred the structure "ST1" (see Figure 3).

Averaged user's preferences of structures "ST2" (broad) and "ST3 (deep) are not that clear and consistent. Figure 3 shows, that over all device classes, with the exception of class 2, users preferred "ST3". The deviation of device class 2 can be attributed to the specific navigation functionality of the used representative of device class 2. In comparison to the other devices it was hard for the users to navigate back to the superior levels. The subjects in particular experienced this problem when they had to navigate through the structure "ST3". In this structure (6 levels x 4 items) they had to navigate six levels deep. For this reason and because of the rather small sample the post hoc analysis of the line lengths delivered only significant differences between "ST1" and "ST2".

In sum the data show at least the tendency that users prefer deep structures to broad ones although broad ones lead to faster search performances. The most striking reason is the more concise arrangement of items. This tendency is also reflected in the user statements gathered during the qualitative interviews.

4 Maximal and optimal lengths of speech lists

4.1 Motivation

Speech applications are like conversations between the user and the computer. Conversations are characterized by turn-taking, shifts in initiative, and verbal and non-verbal feedback to indicate understanding.

There are only a few elements, which a designer of voice applications can use to enable the user to interact with a system. These elements are (1) direct speech input and (2) the selection from lists of n items. Often these two possibilities have to be combined. So, for example, a system may first ask the user to utter a certain item and presents then a list with those items, which match best with the user's speech input.

We distinguish between two kinds of list selection: (1) selections, where the user knows which item he/she wants to select (known target item) and (2) selections, where the system presents a list of available items from which the user has to choose (unknown target item).

Although the adaptability of Miller's [5] well known magic number 7±2 for the design of visual displays certainly is debatable, it is still well known and accepted when it comes to the design of telephone systems. Nevertheless, systems which allow both speech input and output require less memory load than systems which are operated with the telephone keypad because the user does not have to remember a number associated with the item she/he wants to choose.

The goal of this study was to investigate how many items can be presented to users without annoying and overloading them. Both kinds of the lists, which we discussed above, were tested.

4.2 Methodology

The test was divided into two parts: The first part defined the maximal number of items, which can be presented in a list when the target item is unknown. The second test also defined a maximum number of listed items, but in this case the user already knew which item he/she wanted to select.

The tests were realised with a wizard of oz prototype (see e.g. [8], page 541). During the tests one person was sitting in another room and was simulating the system. She did that by operating a computer on which all the system's commands were saved as wav-files. The wizard, used a special software to start the wav-files, which the system presented according to the user's speech inputs. This wav-file then was transmitted to the user via a telephone line. This approach enabled us to avoid biases due to voice recognition problems. Note that the only interaction device of our subjects was the telephone receiver. There was no additional display and subjects could not use the keypad of the telephone to make their selections.

After a briefing session and before the actual test was started users had to go through a sample test, where they had first to select their favourite season from a list. Then they should name one number out of ten. After that we started with the first part of the test.

First test (unknown target item)

Subjects were confronted with four lists. These lists contained a selection of 4, 8, 15, 20 convenience foods. (The items included three to nine syllables.) The facilitator explained the background of this task to the subjects. They should imagine that they are performing a part of a larger product-ordering task.

First users were confronted with the following text, which the system spoke in German:

Step1: "*Please select one product from the following list. The list contains 4 [8, 15, 20] products. After you have heard all list entries, please repeat the product which you want to order.*"

The subjects were instructed that they should really choose the product, which they would like to have for dinner or for lunch. The four list lengths were presented in different orders to the subjects to avoid order effects.

If the user then repeated one of the products which was part of the list that was presented to him/her the system answered:

Step2**:** "*You have selected the product XY? Is that correct?*"

If the user then answered yes, the task was finished. If the user gave an answer that was not part of the list, which was presented to him/her, or if the user asked the system to repeat or if the user did not say anything, the system answered:

Step3**: "***I am sorry, I could not [understand, hear] you. I will repeat the list. When I mention the product that you want to order, please repeat the product name immediately.*"

After the user had uttered the correct product name the system went to "step 2" and finished this task.

After each list we asked users to rate the subjective complexity of the task on a five-point scale. After the subject went through the four lists we encouraged him/her to express how close the selection process was to the subject's subjective threshold of pain. Again we made subjects draw lines to express this closeness. On the sheet, which we gave to our users, the line, marking users' subjective threshold, was 9,5cm from the starting point of the lines drawn by our subjects. The starting point of the lines was marked on the sheets that we gave to our subjects. Therefore, the line-lengths indicated not only the selection comfort but also whether the particular list length exceeded the user's subjective threshold. Finally we conducted a short qualitative interview and asked users to explain their ratings. Furthermore we asked them to define the maximum number of list items, which is acceptable for this kind of list selection.

Second test (known target item)

In this case we asked users to imagine that they want to edit customer details from a database. In order to do this they had to select a customer. After this selection we abandoned the task.

Users got the following instruction from the system in German:

Step 1: "*Please say the name of the customer that you want to edit*."

In the second step the system presented lists of different lengths to the users. These lists contained the items, which the system associated with the user's input. Four lists were presented to the users. The lists contained 4, 8, 15 and 20 items. (The items included one to three syllables.) Again the lists were presented in different orders to our subjects. The correct name was always two entries before the end of the list. The system's instruction were as follows:

Step 2: "*I could not understand you. I will now read a list with the names, which you possibly uttered. When I mention the name that you have said, please repeat the name immediately*."

Then the system presented the lists to the user. After the user has uttered the correct name the system replied as follows:

Step 3: "*You have selected the name XY? Is that correct?*"

The user then answered yes, which was the end of the task.

After each task and after the four lists again the same procedure as described above was repeated.

4.3 Results

The tests were conducted with ten subjects (7 males and 3 females). Their average age was 28 years (Std. Dev.: 13 years). Figure 5 shows users' ratings of the complexity of the list selection tasks. On the left hand side the figure shows the results of the selection with unknown target items (convenience foods), on the right hand side it shows the results of the selection with known target items (names). The next figure (Figure 6) shows the users' comfort of the selection expressed in line lengths. A length of more than 9.5cm means that at this point users would prefer another kind of selection (for example a list, which is divided into sub-lists). Both figures show that when users know the target item 15 items still seems to be OK, whereas on the other hand this number is too high for a selection of an unknown target item.

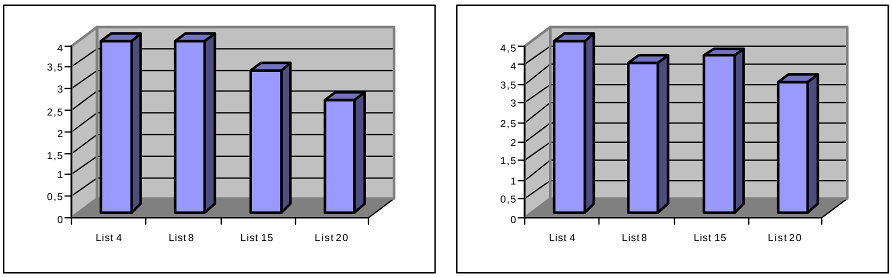

Fig. 5. Mean complexity and comfort of selection expressed in ratings from 1 to 5 (1: uncomfortable, 5 comfortable) left: Selection of unknown items, right: Selection of known items

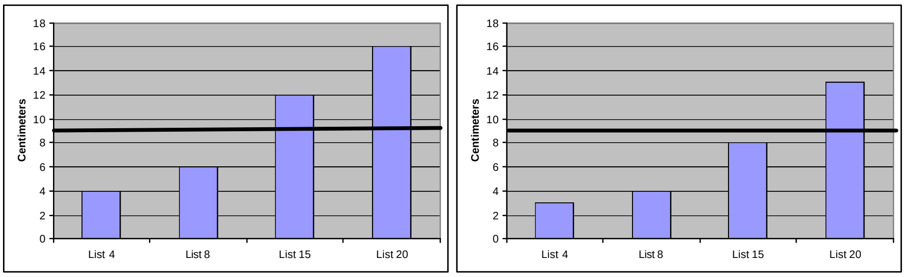

Fig. 6. Mean complexity and comfort of selection expressed in line lengths (centimetres) left: Selection of unknown items, right: Selection of known items (The shorter the line the higher the user satisfaction

Figure 5 and 6 show that if the target item is unknown, there is a clear border, which lies between list lengths of 8 and of 15 items. For lists where the user knows the item he/she wants to select, there does not exist such a clear borderline. This can also be seen when we have a look at the "thresholds of pain", which were defined by our subjects during the qualitative interviews and at their confidence intervals. The mean values of these thresholds can be seen in Figure 7. The 95% confidence interval for a selection with an unknown target item lies between 8.4 and 13.2 items, whereas the 95% confidence interval for a selection with a known target item lies between 8.2 and 18 items.

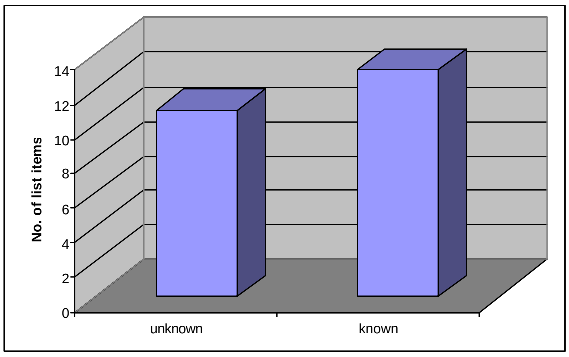

Fig. 7. Mean "thresholds of pain" as defined by our subjects

These results show that the optimal number of list entries of lists with unknown target items should lie around 8. Lists longer than eight are possible but should not exceed approximately 13 items. This absolute maximum is derived from the confidence interval of users' thresholds of pain and from the fact that there is a clear borderline between lists of 8 and lists of 15 items in terms of selection comfort. Two users said also that after the 15th item they forgot to listen because then the task was too demanding for them.

The results of the second type of list were slightly different. Again, 8 items seem to be the optimum. So we can say that this number of list entries is optimal independent form the type of list, which is presented to the user. However, in this case the maximum number of possible entries seems to lie higher than in the case discussed above. On the one hand the confidence interval of users' thresholds is much broader and on the other hand there are two equal steps (see Figure 6) from lists of eight to lists of 15 and from lists of 15 to lists of 20. Therefore the maximum number of possible list entries should lie around 18.

Note that all the items of our lists included different numbers of syllables and words. Future studies will have to investigate the influence of these numbers on users' subjective satisfaction. An influence seems to be possible since Baddely et al. [2] could show that the number of syllables influences the number of words that can be stored in humans' working memory (word length effect).

5 Conclusions

In the last chapters only three of our studies were reported. We discussed a study on text reading, where we saw that for content pages/cards scrolling seems to be more appropriate than pagination. In the second study we proved the hypothesis that in a tree structure all items on the same level should be perceptible at a glance. Finally we defined optimal and maximal numbers of list entries for two different kinds of speech-lists.

Our overall approach was to develop a mosaic of empirical tests, which are well fitting together. In synergy with the already available empirical and qualitative data they led to a picture of do´s and don't´s included in our user interface guidelines.

Examples of other research questions, which we answered by empirical studies and whose answers were fed into our guidelines are listed below:

- Task efficiency of direct text input tasks
- Thresholds of pain of unsorted WAP- and html-lists
- Thresholds of pain of sorted WAP- and html-lists
- Navigation through forms (scrolling vs. pagination)
- Comparison between navigation by search entries and by tree-navigation
- Comparison of speech feedback mechanisms of number input
- Definition of mean viewing distances per device class

The three examples presented in this paper show how empirical studies informing the development of user interface guidelines can be conducted and to which kind of results they lead. They also show that studies like these can only answer very detailed questions and that there is an almost infinite number of further open questions. You cannot assume to cover a broader range of questions with one study because studies like these include a lot of devices and require carefully chosen set ups.

We want also to emphasise that because of differences inside the device classes the results have to be analysed very carefully. Often these differences are not obvious before the actual study has been conducted. Therefore it is important that appropriate conclusions are drawn from these data.

An example of such a problem was given in chapter 3 (second example study). There we saw that some users had difficulties to navigate backwards with one of the representatives of our device classes and some users experienced also other hardware related difficulties. That means that users' performance with two devices of one and the same device class may differ because of hardware- or browser-specific differences, which are not part of the device class specification.

Future research should focus on the definition of classes, which are minimising these problems. As long as this concerns the browser capabilities the number of differences, which have to be considered, may be manageable. However, when it comes to hardware differences the number of differences is almost infinite and steadily growing.

So the challenge of the future development of device classes will be to include such differences and to define them and their effects on the efficiency and user satisfaction of each task. Furthermore new classes will have to be developed because of new products and types of interface designs that are entering the market. Therefore the three dimensions, which currently are used for the definition of device classes, will have to be evaluated on a current basis to ensure that they fulfil the requirements of a device classification including the latest developments.

6 Acknowledgements

The empirical studies and the guideline development were conducted as a part of the CONSENSUS project, which is funded by the European Commission. We would like to thank all our partners for their inputs and for their comments.

References

1. Buchanan, G., Farrant, S., Jones, M., Thimbleby, H., Marsden, G., Pazzani, M. (2001). Improving mobile internet usability. WWW10, May 1-5, 2001, Hong Kong, China
2. Baddely, A.D., Thompson, N., Buchanan, M. (1975). Word length and the structure of short term memory. Journal of Verbal Learning and Verbal Behavior, 14, 575-589
3. CONSENSUS: IST PROGRAMM/KA4/AL:IST-2001-4.3.2/CONSENSUS/CN:IST-2001-32407 3G Mobile Context Sensitive Adaptability User Friendly Mobile Work Place for Seamless Enterprise Applications. (www.consensus-online.org)
4. Dickes, P. and Steiwer, L. (1977). Ausarbeitung von Lesbarkeitsformeln für die deutsche Sprache. In Zeitschrift für Entwicklungspsychologie and Pädagogische Psychologie. Band IX, 1/1977, 20-28 (in German)
5. Miller, G.A.. (1956). The magic number seven plus or minus two: Some limits of our capacity for information processing. Psychological Review, 63(2), 81 - 87
6. Nielsen, J. (2000). Designing web usability. New Riders Publishing, Indianapolis, Indiana USA
7. Paap, K.R. and Cooke, N.J. (1997). Design of Menus. In Helander, M.G, Landauer, T.K., Prabhu, P.V. Handbook of Human-Computer Interaction (second edition)
8. Preece, J., Roger, Y., Sharp, H., Benyon, D., Holland, S., Carey, T. (1994). Human-Computer Interaction. Addison-Wesley
9. Spool; J.M., Scanlon, T., Schroeder, W., Snyder, C., DeAngelo, T. (1999). Web site usability: A designer's guide. Morgan Kaufmann Publishers, Inc.
10. Telenor Mobile Communications (2000). User Interface design guidelines for WAP applications. Version 1.4
11. Van Welie, M. and de Groot, B. (2002). Consistent multi-device design using device categories. In Proceedings of the 4th International Symposium, Mobile HCI 2002, Pisa, Italy