

User Evaluation of a Word Prediction System

Malin Wester
malwes@stp.ling.uu.se

Master's thesis
Uppsala University
Department of Linguistics
Language Engineering Program

30th September 2003

Supervisor of Master's Thesis:
Eva Ejerhed,
Department of Linguistics, Uppsala University

Abstract

Word predictors play an important role in facilitating the text generation process for persons with physical and linguistic disabilities.

The FASTY project aims at developing a word prediction system to assist motor, speech, learning and language impaired persons to produce texts faster, with less physical and cognitive load and with better spelling and syntax.

This thesis describes a user evaluation of the Swedish version of the FASTY word predictor. The purpose of the evaluation was to test the functionality and usability of the first prototype of FASTY. Not only the prediction was evaluated, but also the interface of the system into which it is integrated. The system was tested with seven Swedish test users, of which some are linguistically disabled and some physically disabled. The evaluation is based on a questionnaire and 141 log files generated by the test users who used FASTY to enter text.

The questionnaire showed that the users were generally confident that the system is useful. They thought that the system was easy to understand and that the adjustment possibilities were well structured. Bugs of the tested prototype did, however, restrict the usability of the system and it did therefore not seem to be well developed.

The log files were analysed using the standard keystroke saving rate (KSR) measurement. The result was alarmingly poor. The study did however show that different settings, such as sorting of the prediction list and number of predictions did have essential impact on the KSR.

Acknowledgements

The work described in this thesis was carried out at the Department of Linguistics at Uppsala University as a part of the FASTY project (Faster Typing for Disabled Persons, IST 2000-25420) in the framework of Information Society Technologies (IST).

First of all I would like to thank my supervisor Eva Ejerhed for well-motivated suggestions and very good support throughout this work and Mats Dahllöf at the Department of Linguistics at Uppsala University for useful suggestions regarding the text. I would also like to thank Anna Sångvall Hein for letting me take part in this project.

Further, I would like to thank all the participants in the FASTY project, in particular Mikael Wiberg, for providing useful tools and for his invaluable support. A special thank to Mikael Wiberg, Ebba Gustavii and Eva Petterson with whom I shared working room, for coping with my many questions and for their insightful comments.

I would also like to thank Lena Santesson-Hultén at Hjälpmedelscentralen and Helene Lidström at Folke Bernadotte-Hemmet for helping me to get in contact with potential test users.

Finally, I wish to thank all the users that have taken part in this evaluation for their patience due to all the technical troubles during the installation phase and for their valuable opinions regarding the system.

Contents

1	Introduction	1
1.1	Purpose	1
1.2	Outline of the thesis	1
2	Assistive technology and word prediction	2
2.1	Assistive technology	2
2.2	Augmentative and Alternative communication	2
2.3	Word prediction	3
2.3.1	Who benefits from a word predictor?	6
3	FASTY	7
3.1	The FASTY project	7
3.1.1	Main parts of FASTY	8
3.2	FASTY Word Predictor	10
3.2.1	Prediction technique	10
3.2.2	Data	10
3.2.3	The dictionary	11
3.2.4	Compound prediction	11
3.2.5	Grammar checking	12
3.2.6	Settings	12
4	Usability and user evaluation	14
4.1	Usability	14
4.2	User evaluation	15
4.3	Evaluation methods	16
5	User evaluation of FASTY	18
5.1	Test procedure	18
5.2	Test users	18
5.2.1	Ethical aspects	19
5.3	Qualitative study	20
5.3.1	Usability measures	20
5.3.2	Questionnaire	21
5.4	Quantitative study	22
5.4.1	Log files	22
5.4.2	Keystroke savings rate	25
5.4.3	Collection and calculation	27
6	Test result	28
6.1	Questionnaire analysis	28
6.1.1	Use of prediction and other functions during writing	28
6.1.2	Adaption of the system	35
6.1.3	Documentation, online-help, and tutorial	47

6.1.4	General impression	51
6.1.5	Open questions	69
6.1.6	Summary of the qualitative study	70
6.2	Log file analysis	71
6.2.1	Change over time	72
6.2.2	Number of predictions	73
6.2.3	Sorting	73
6.2.4	Placement and orientation	74
6.2.5	Grammar checking	74
7	Conclusion	75
	References	76
	Appendix	79
A	User interface images	79
B	Questionnaire glossary	89
C	Time table	90
D	Test report template	91
E	Keystroke saving rate and typing errors	92

1 Introduction

Verbal communication is a vital need for humanity. In the society of today, computers play an increasing role as a communication tool. Generation of text may, however, be very cumbersome for persons with physical or linguistic disabilities and the process of entering text may thus be slow. In these situations word prediction may be of great help to speed up the text generation rate.

There exist useful measurements to estimate the physical effort saved using word prediction to generate text. However, these measurements do not say anything about user satisfaction.

This thesis describes a user evaluation of FASTY, a system for increasing the text generation speed of disabled persons by predictive typing. The introduction explains the purpose of the thesis and describes the outline of it.

1.1 Purpose

The aim of the evaluation described in this thesis, was to test the Swedish version of the first prototype of FASTY word predictor with real users, and gather data about its usability and functionality. The main question to be answered is if FASTY meets the expectations, and is able to live up to the users' demands. It may also be interesting to see if user parameters such as type of disability, age, sex, and computer experience affect the degree of satisfaction. Further to be evaluated, is if FASTY reduces the number of keystrokes and the effort required to type.

The fundamental steps of the test procedure were set at the FASTY Consortium meeting that was held at Multitel ASBL, Copernic Avenue 1, B-7000 Mons, Belgium, on the 21st of January 2003. Representatives of each participating contractor were taking part.

The evaluation will be based on a *questionnaire* and a large number of *log files* generated by the test users, who have used FASTY to enter text. It might be difficult to compare the results in this study with those in word prediction literature, because there exists no agreed standard of measurement for this type of evaluation. The results will, however, lead to recommendations about what to focus on when improving the system before taking it to market.

1.2 Outline of the thesis

This thesis is organised into six main parts. Part 2 introduces the field of study concerning assistive technology and augmentative and alternative communication for disabled persons. It also gives a description of word prediction and who may benefit from it. Part 3 describes the FASTY project and the FASTY word predictor. In part 4 the concepts of user evaluation and usability are clarified and in part 5 a description of the data and the methods used in this user evaluation is given. Part 6, which is the head part, consists of a presentation and a discussion of the test results. Finally, part 7 provides a conclusion of the study described in this thesis.

2 Assistive technology and word prediction

This part introduces the concepts of assistive technologies and augmentative and alternative communication. It further provides a short introduction to the specific assistive technology that is focused on in this thesis, namely word prediction. Section 2.3 gives a detailed account of word prediction in general and section 2.3.1 discusses what kind of persons may benefit from it.

2.1 Assistive technology

As quoted in (Raskind and Shaw 1999), the Technology Related Assistance for Persons With Disabilities Act of 1988 defines assistive technology as: “any piece of equipment or product system, whether acquired commercially of the shelf, modified or customised, that is used to increase, maintain or improve functional capabilities of individuals with disabilities”. Thus, this definition does not imply that assistive technology must include computers, be expensive or be prescribed by certain medical professionals. Assistive technology should be seen as a strategy to compensate for areas of difficulty and only imagination and creativity restrict what it can be (Rizer, Cirlot-New and Ethridge 1999).

Assistive technologies benefit many people with motor, sensory, communicative, and cognitive impairments and help them to complete an everyday task. For individuals with severe disabilities, assistive technology is the key to successful social participation, i.e. to establish and maintain contact with the world around them. Assistive technology includes common devices like a remote control for the TV or a pair of glasses, as well as specialised devices like typing telephones, motorised wheelchairs, word processors, word prediction, speech recognition, spell checkers etc. (Laine and Bristow 1999).

2.2 Augmentative and Alternative communication

Augmentative and Alternative Communication (AAC) is the field of study concerned with providing devices or techniques to augment the communication ability of a person whose disability makes it difficult to communicate in an understandable manner. It refers to ways, other than speech, that are used to send a message from one person to another. Such devices allow a person who cannot speak, or whose speech is not understood by others, to communicate. (McCoy 1998). Many of the AAC devices existing today, are aimed at maximising specific language concepts and strategies to enable persons to interact with their environment (Rizer et al. 1999). AAC tools include such things as sign language, speech prostheses, symbolic languages and letter charts (Kronlid 2001). A person, who has lost the ability to speak or move being an adult, may however prefer to continue to use his or her original language. Therefore prediction techniques have been extensively used in AAC (Copestake 1997).

The use of computer based AAC generally has many advantages. Disabled

persons are enabled to communicate and exhibit intelligence that was previously impossible because of their disabilities. But there are also some challenges of AAC. One of the biggest is that the group of users is so diverse. A computer based AAC system must therefore be tailored to the user depending on his or her physical and cognitive circumstances and the task they want to perform (McCoy 1998).

2.3 Word prediction

Word prediction is a technique frequently used in the AAC field of writing support devices with the purpose of improving the keyboard text input speed, and the quality of spelling and syntax.

Word prediction is about predicting which word tokens or word completions are most likely to follow a given segment of text. This means that a few keystrokes produce complete words or word sequences and the number of keystrokes necessary to generate texts will be reduced.

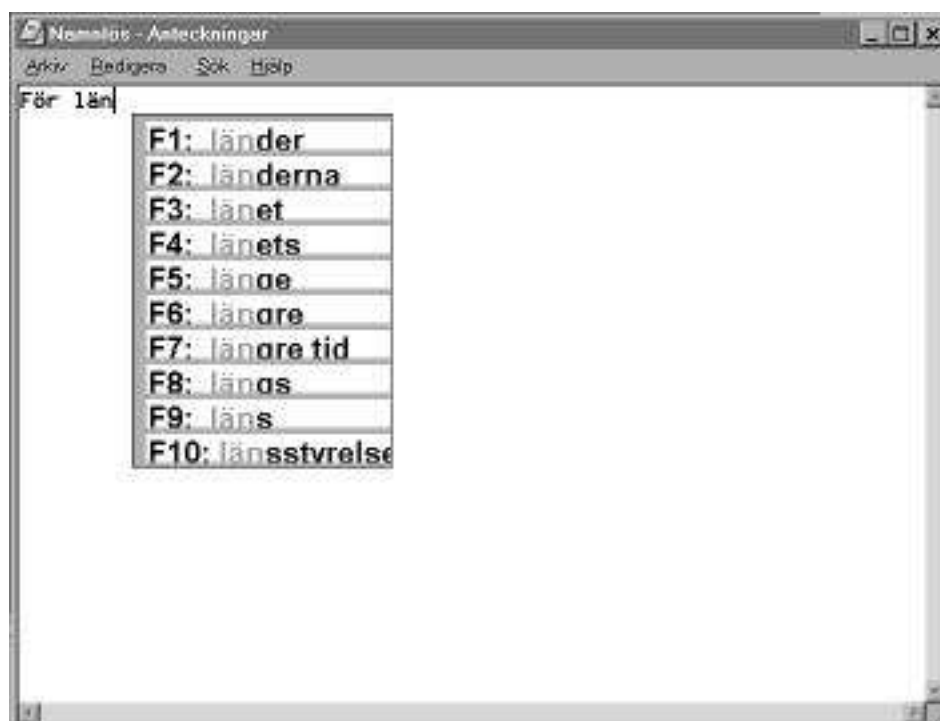


Figure 1: Completion - predictions based on the typed substring

A computer program performing word prediction, usually called a *word predictor*, operates by taking the character or characters the user has entered via keyboard or other input device, and generating a list of suggestions of possible words or word completions. The user either chooses one of the suggestions, by pressing the associated function key, or continues to feed new characters into the interface until the intended word appears. A selected suggestion is automatically inserted into the text

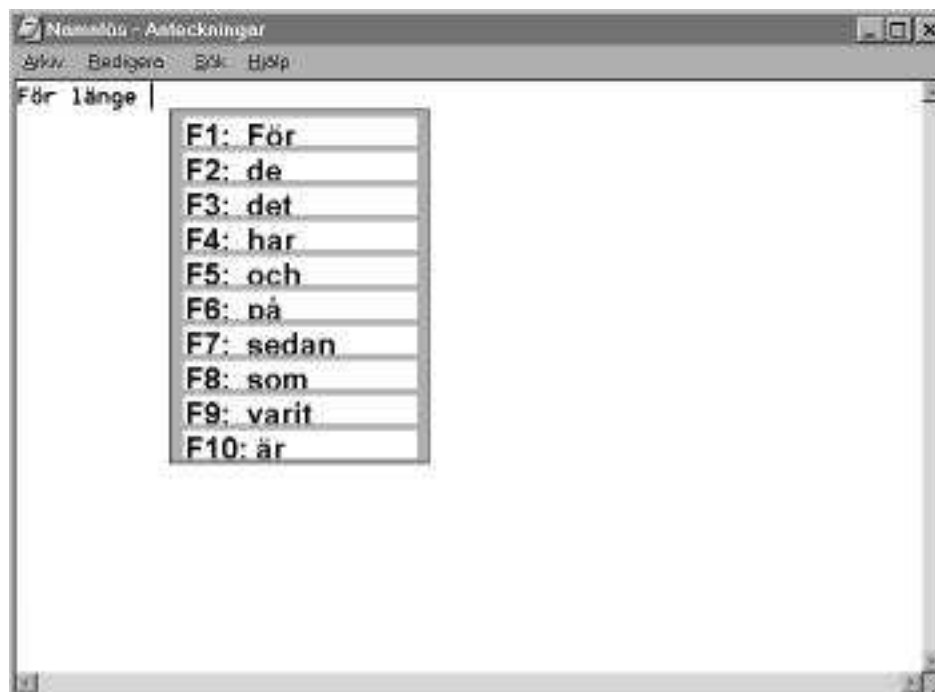


Figure 2: Proper word prediction - prediction based on the context

(Carlberger 1997). A word predictor can either predict whole words or complete a typed substring of a word. The former is usually called proper word prediction and the latter completion. *Completion* (see figure 1) focuses on the next character of the word, i.e. to complete the current word with respect only to the already typed characters of a certain string. Completion of predictions always require at least one character to be able to predict the intended word.

Proper word prediction (see figure 2) takes the previous word(s), as well as any initial substring of the intended word into consideration when predicting the next word. Hence, completion may be a sub-function of proper word prediction. If there are no preceding words, as is the case at the beginning of a sentence (see figure 3), words that are most likely to initiate a sentence are suggested. The method for considering the context is normally based on word form statistics and/or part-of-speech tags (Klund and Novak 1995).

To be able to predict words or word sequences in a satisfactory manner, every word predictor needs a collection of language data of the language in question. It is used to extract statistics or pattern representations, that can make predictions given previously unseen language fragments (Kronlid and Nilsson 2000). Word predictors for AAC usually make use of statistical language modelling techniques that rely on the assumption that the majority of the words to be predicted have also occurred in the training data of the model. Sometimes that modelling technique is augmented by a language model with language awareness, a so called knowledge-

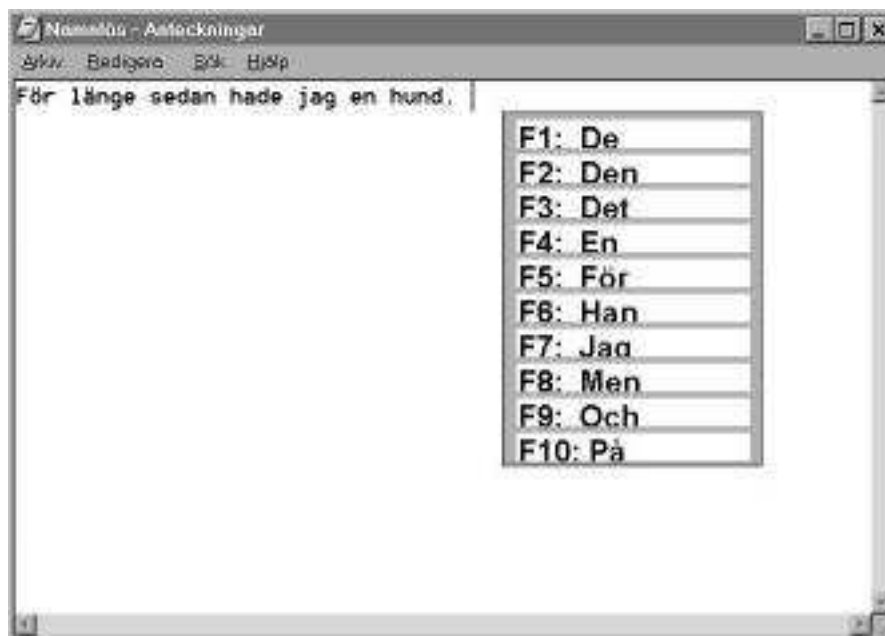


Figure 3: Prediction based on the most likely word to begin a sentence

based language model. Further, the implemented language model requires certain information about each word form, such as possible part-of-speech tag or morpho-syntactic description. This information is provided in a dictionary (Gustavii and Pettersson 2003).

Additionally, several techniques to make the system more suitable for the specific users can be added to the system. Two frequently used short term learning techniques are recency of use and domain specific word prediction. *Recency of use* means that a word that already has occurred in the current text will get a higher probability of occurrence and will therefore appear higher in the prediction list than it would otherwise (Carlberger and Hunnicutt 2001). Using *domain specific word prediction* means adapting the predictions to the subject matter of the current text. This technique requires the standard dictionary to be supplemented with dictionaries containing domain specific words, that almost only occur within certain domains (Leshner and Rinkus 2001).

Opposed to the short term learning techniques described above, it is also common to make use of long term learning techniques. *Long term learning* means that the system adapts to the user by taking the current text and/or previous texts produced by the user into consideration. It usually has to do with adding new words to the existing dictionary. It may also mean creating user specific dictionaries that are combined with the standard dictionary. These techniques have the drawback that misspelled and ungrammatical words may be added to the dictionary. This can be avoided solely by adding words that have been typed more than once, or by using spell checkers and grammar checkers. Another problem, that is more difficult

to solve, is that new words have to be provided with the correct morpho-syntactic information (Gustavii and Pettersson 2003).

2.3.1 Who benefits from a word predictor?

Word predictors play, as noted above, an important role in facilitating the text production process. There are two main groups of users who are said to benefit from using a word predictor. Traditionally, the primary users of word prediction have been physically disabled persons with typing difficulties. By using a word predictor both time and effort needed for producing texts may be reduced (Carlberger 1997). Only having to type a reduced number of keystrokes, the user may also be able to write for a longer time (Klund and Novak 1995).

The other group consists of persons with dyslexia or other impaired language skills, who have difficulties in spelling correctly and writing grammatically well-formed sentences (Carlberger 1997). The success of word prediction for spelling assistance depends strongly upon the dictionary and the prediction algorithm (Klund and Novak 1995).

The usage of word prediction does, however, impose a high degree of perceptual and cognitive load since it requires the user to shift attention from the text in progress to the prediction suggestions. Especially for persons with learning disabilities but with good keyboard skills it may be distracting to have to interrupt the flow. Sometimes the cognitive load costs more in terms of text generation rate (Fazly 2002).

Characteristics of a word predictor that may influence the cognitive load are length, orientation, placement and sorting of the suggestion list (Klund and Novak 1995).

Obviously there are a number of constraints on how powerful a word predictor can be, and the appropriate one to work on will depend on the user. A person that has severe motor impairments but is cognitively strong, might need a different type of word predictor than a person who has good motor control, but who gets frustrated by a large cognitive load (Willis 2001).

3 FASTY

This part gives an introduction to FASTY. Section 3.1 presents the FASTY project and its main purpose. Section 3.2 describes how the FASTY word predictor works.

3.1 The FASTY project

European research activities are usually structured around consecutive four year framework programs. FASTY is an EU-funded project within Information Society Technologies (IST) that is part of the Fifth Research and Technological Development Framework Program¹. It is scheduled for the period from January 2001 to December 2003.

The FASTY project has four participating countries: Austria, Belgium, Germany and Sweden and the Consortium consists of nine contractors, specified in table 1.

Participants	Status	Country
fortec	Coordinator	Austria
ÖFAI	Principal Contractor	Austria
FTB	Principal Contractor	Germany
IGEL	Principal Contractor	Germany
Uppsala University	Principal Contractor	Sweden
Multitel ASBL	Principal Contractor	Belgium
ELI	Assistant Contractor	Austria
IKuT	Assistant Contractor	Germany
FUNDP	Assistant Contractor	Belgium

Table 1: Contractors of FASTY Consortium

FASTY aims at developing a communication support system to assist motor, speech, learning and language impaired persons to produce texts faster, with less physical and cognitive load and with better spelling and syntax. The concrete goal is to create a system for increasing the text generation rate of disabled persons by so called predictive typing. The system is being developed for German, French, Dutch and Swedish. These languages differ from English in that they are highly inflecting².

FASTY is an intelligent system, that uses methods of Natural Language Processing (NLP), Artificial Intelligence (AI), a self adaptive user interface and linguistic resources such as dictionaries and grammars (Baroni, Matiasek and Trost 2002a).

¹For further information on IST, please visit: <http://www.cordis.lu/ist/ist-fp5.html>.

²For further information on the FASTY project please visit FASTY website: <http://www.fortec.tuwien.ac.at/reha.e/projects/fasty/fasty.html>

3.1.1 Main parts of FASTY

The main components of the FASTY system can be divided into three parts: *runtime system*, *adjustment tool* and *developer tools*, which will be described in this section. An overview of the parts is to be seen in figure 4.

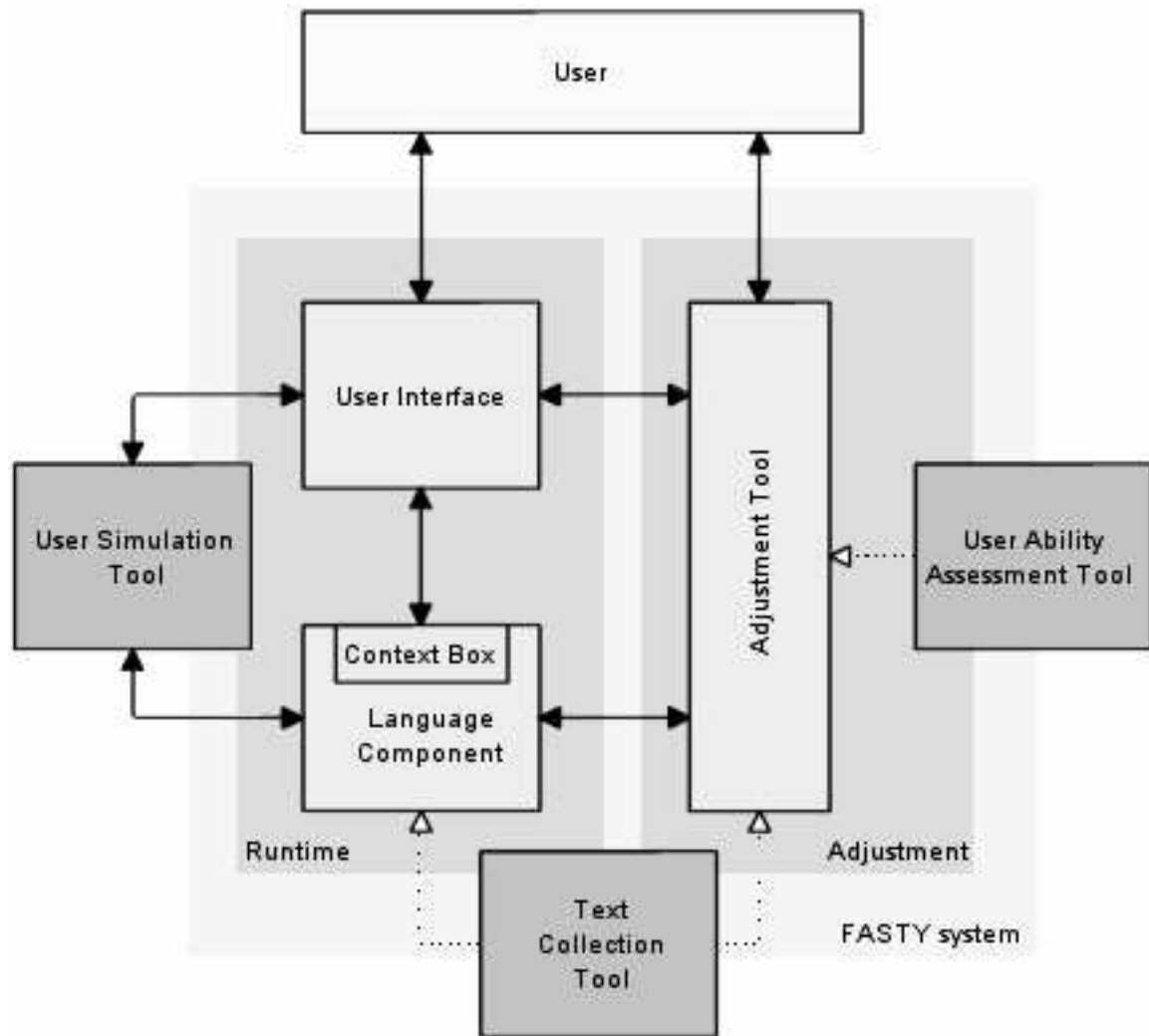


Figure 4: The main parts of the FASTY system

1. The runtime system consists of two components:

- User interface (UI)

The user interface in FASTY is proposed to be innovative and ergonomic and to facilitate the interaction between the users and the system. It is developed to fulfil the usability of the system by offering a variety of options concerning colour, font size, position, sorting etc.

- Language component (LC)
The language component consists of linguistic resources of the four involved languages. The FASTY language model is based on both basic and innovative resources. The basic module relies on a statistic language model of word form bigrams and trigrams of part-of-speech tags. The innovative modules consist of grammar checking and compound prediction (Gustavii and Pettersson 2003).
2. The adjustment tool (AT) supports the process of adjusting the system's options to the user's specific needs and situation. The user runs a set of tests and answers some questions, which are intended to provide a picture of his abilities. The adjustment tool provides three modes: a simple guided one for beginners, a more sophisticated one for advanced users and an expert mode with all options.
 3. The third part consists of four developer tools, in order to support the developers in the ongoing design and implementation process:
 - User ability assessment tool (UAAT)
The user ability assessment tool was developed to collect data about basic user performance, such as typing speed, or reaction time. The collected information is used to determine the applicability of further tests to the situation of the user. It is also used to get an impression of the user's current hardware and software status and the way text input is written.
 - Text collection tool (TCT)
Since the language component requires big amounts of user generated texts for building user-adapted dictionaries, a text collection tool was needed. The texts are collected from the potential users in order to fit the type of language the users use. An important feature of the text collector is an anonymity module, which codes those items in the text, that would make it possible for someone to identify the writer of the text.
 - User simulation tool (UST)
The user simulation tool aims at helping the developers of the user interface and the language component to evaluate different algorithms and settings. Later commercial versions may help to find the best settings for each user in an iterative way without burdening the users with the testing of all possible settings and options.
 - A simple word predictor (SWP)
SWP is a tool that has been developed to support the continuous improvement of the language component. As new functionality is added

to the SWP, it is possible to study how the predictive power and precision increases with the integration of new functionality³.

3.2 FASTY Word Predictor

Predictive typing systems for English have proved to be useful and efficient for a long time, but for other European languages there are only a few such systems powerful enough to improve the communication rate for disabled persons. Adapting the English programs to highly inflecting languages, like for instance Swedish, by replacing the English dictionaries, often leads to a significant reduction of the keystroke saving rate. Therefore FASTY is designed by taking into account the specific properties of each language for which it is to be used. To be able to develop a system which is applicable to several European languages, the system has to separate the predictor from the language-specific resources (Baroni et al. 2002a).

3.2.1 Prediction technique

As with all word predictors, FASTY is designed to assist and help users with data entry. The predictive typing in FASTY works in resemblance to proper word prediction described in section 2.3. As a user types, the software monitors the input character-by-character, takes the previous word form and the two previous part-of-speech tags into consideration, and produces a list of words beginning with the string recorded, (see figure 5). Each time a new character is added, the list is revised taking the extended context into account. When the intended word appears in the list, it can be chosen and inserted into the ongoing text with a single keystroke. FASTY predicts words even if no character has been typed, i.e. after white space or punctuation. In these cases, the prediction list is based on words that are most likely to follow the previous word form and the part-of-speech tags of the two previous words. At the beginning of a sentence, the prediction list consists of word suggestions that are most likely to initiate a sentence.

3.2.2 Data

The Swedish word form statistics are extracted from a training corpus of approximately 19 million word tokens. The corpus constitutes 90% of the UNT-corpus, compiled at the Uppsala University. The UNT-corpus contains newspaper articles, that were published in Upsala Nya Tidning during 1995-1996. The remaining 10% were left aside to form potential test material⁴. The part-of-speech statistics is extracted from the hand-tagged, balanced Stockholm Umeå Corpus (SUC) comprising one million word tokens (Gustavii and Pettersson 2003).

³For further information on the FASTY project, please visit FASTY website: <http://www.fortec.tuwien.ac.at/reha.e/projects/fasty/fasty.html>

⁴This has not yet been done.

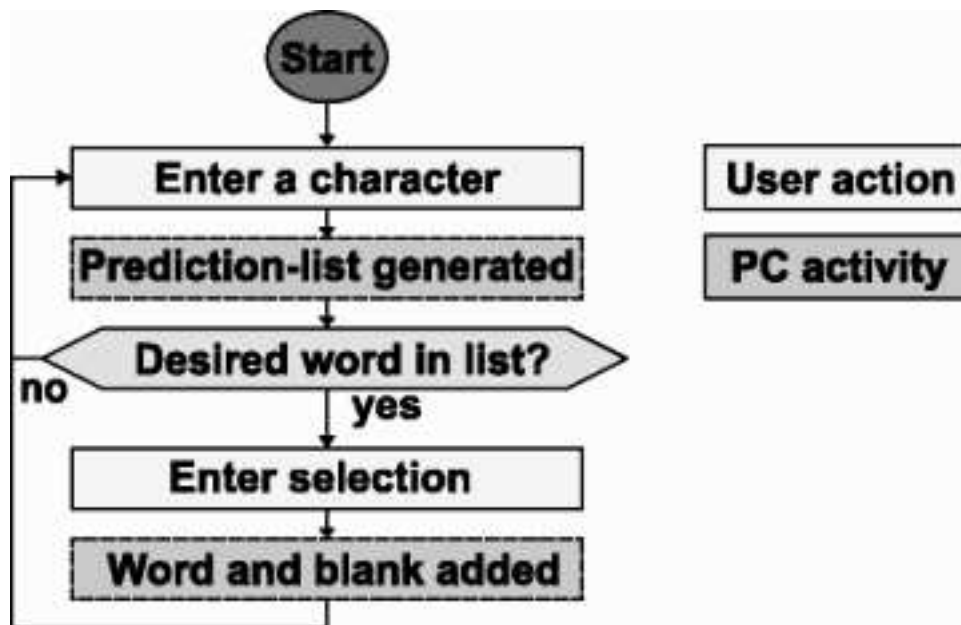


Figure 5: The predictive typing in FASTY

3.2.3 The dictionary

The Swedish dictionary in FASTY is extracted from the Scarrie lexical database, which is based on the UNT- and SVD-corpora and contains approximately 200,000 word forms annotated with morpho-syntactic features. The dictionary is static, but has the ability of recency of use to the current text. This means that words, recently typed by the user will get a higher probability and will thus be suggested sooner than other words (Gustavii and Pettersson 2003).

It is also possible for the user to create a user specific dictionary to be combined with the standard dictionary. These words will, however, not be automatically provided with the morpho-syntactic information required by the language model. There is neither a spell-check, that filters misspelled words.

A further step would be to dynamically add new words to the standard dictionary. This may however, as in the case of the user specific dictionary, cause addition of misspelled and incorrect words to the dictionary.

3.2.4 Compound prediction

In FASTY, compound prediction is a configurable setting that is either active or inactive. Compound prediction differs from ordinary prediction of full, non-compound words in that it treats the two parts of a two-word compound separately. The first part is predicted by modifier prediction, and the last by head prediction.

If compound prediction is active, modifier prediction is integrated with the ordinary word prediction. If the user selects as next word a word that is more frequent

as a modifier in a compound than as a full word, then the system predicts a possible head of that compound, instead of inserting a white space. When predicting the head, the system makes use of the word preceding the modifier, as if the modifier was absent. Active compound prediction can suggest compounds that have not occurred in the training corpora, as well as compounds that have occurred. If deactivated, FASTY will not predict compounds that have not occurred in the training corpora (Gustavii and Pettersson 2003).

3.2.5 Grammar checking

The grammar checker in FASTY filters the suggestions made by the statistical language model, based on grammar rules. Input to the grammar is a ranked list of the most probable words according to the statistical language model. The grammar rules either accept, reject or leave the predictions undecided⁵. As a basis for the syntactic component, a modified version of the Uppsala Chart Parser (FastyUCP) is used. FastyUCP is based on incremental parsing, which means that it does not need the whole input sentences to interpret a constituent.

It is possible for the users to choose whether the grammar rules are active or not when using FASTY. If the grammar rules are active the grammar checker reranks the predictions as follows: the rejected ones are given the lowest probability while the undecided and the accepted words are given the same probability. It is further possible to promote the accepted word. If this option is activated, the accepted words are given the highest probability and the rejected the lowest. The undecided words are thus being placed somewhere between the accepted and the rejected words (Gustavii and Pettersson 2003).

3.2.6 Settings

In order to meet different needs, the user interface of FASTY offers a lot of individual settings. The user can for instance choose how many suggestions that are shown in the prediction list. The different settings are presented in table 2 below. Images of the interface are also to be seen in Appendix A. The option *predictions start with n number of characters* means that there will not occur any predictions with a lower number of characters than the number a user has chosen. It is possible to choose a number between 1-99, but a selection over 20 characters will probably not make any sense. Therefore only 20 options, i.e. predictions start with 1-20 characters, have been taken into consideration when calculating the number of possible system settings.

All the configurable options produce a total of 9,600,000 potential system settings⁶. It is quite obvious that this provides a problem for the evaluation, since it is impossible, or at least very time-consuming to separately evaluate each possible configuration.

⁵For a detailed descriptions of the grammar rules please see (Gustavii and Pettersson 2003).

⁶Colour, font and font size options are also user configurable, but not included here.

Settings	Options
Kind of input	Function keys Numpad
Position	Near caret Random Top of application Bottom of application Left of application Right of application Top of screen Bottom of screen Left side of screen Right side of screen
Sorting	Alphabetic Probability Length - increasing Length - decreasing
Number of prediction	1-10
Smart punctuation	activated inactivated
Add space	activated inactivated
Compound prediction	activated inactivated
Grammar based prediction	activated inactivated
Grammar based promote accepted (<i>can only be active if grammar is active</i>)	activated inactivated
Predictions start with <i>n</i> number of characters	1-99
Unigram Match Type (<i>regarding case</i>)	0-4
Bigram Match Type (<i>regarding case</i>)	0-4
Font	17 options
Font size	1-10
Colour	Predictions window Colour of prefix Colour of numbering Background colour Frame colour

Table 2: Settings of FASTY

4 Usability and user evaluation

In this part the concepts of usability and user friendliness are clarified. In section 4.2 dimensions and terminology used in literature on system evaluations are clarified and in section 4.3 the different types of user evaluation methods are described.

4.1 Usability

The point of using computers is to make it easier to perform the tasks we want to perform. Computers aim at raising our productivity. Productivity depends on functionality as well as on different aspects of usability, and also on how well the user's information need is satisfied (Allwood 1998). Developing a product with good functionality and usability means a high level of productivity and makes sure that the software works satisfactorily and that it acts as expected. In comparing two or more systems that have similar functions, it is more likely that the users will buy the system that requires the least effort to learn and that is easiest to use. Thus, it is hardly acceptable developing a system in which the usability has to give way to technical priorities.

Usability and user friendliness have become popular buzzwords when talking about system development. High usability and user friendliness can reduce the users' mental effort and hence increase the productivity substantially. Usability and user friendliness are related in many ways since they both aim at higher user satisfaction, but they are yet not identical. User friendliness focuses more on the *perceived ease* in using the system, while usability concentrates on the *objectively measurable efficiency* in using the system. Disregarding this difference, both usability and user friendliness will henceforth be referred to by the term usability.

Each interaction between user and computer has to be natural for the user and the system should preferably be seen as observant, reliable, competent and trustworthy. Usability depends to a great extent on how well the components interact in a work situation. A system can be very usable for specific users doing specific tasks but useless for other users or other tasks. The usability can also change over time as the users' experiences change (Min-Yang Wang 1992). Usability is therefore said to be the quality of use in a specific context and when specifying the usability it is important that the context selected is representative of the actual or intended context of use (Bevan and Macleod 1994). The two most important aspects to consider when evaluating usability are the users' tasks and their individual characteristics and differences. It is important to know about the users' computer experience, system experience, and knowledge about the domain. That is why usability usually is measured by using real users performing some predefined tasks (Nielsen and Mack 1994).

4.2 User evaluation

Evaluations play a very important role in system development. To better understand, and be able to satisfy, the users' needs and requirements it is necessary that the developers maintain a continuous interaction with the users. To create a well-functioning human-computer interaction, user participation during the entire development process is required.

Evaluations consist of gathering data about the usability of a design or a product, by a specified group of users, for a particular activity within a specified environment or work context (Preece, Rogers, Sharp, Benyon, Holland and Carey 1994). Here, it may be useful to distinguish process data from bottom-line data. *Process data* are observations of what the test users are doing and thinking as they work with the system. These observations tell what is happening step-by-step and why it is happening. *Bottom-line data* give a summary of what happened: how long did users take, were they successful, how many errors did they make etc. (Lewis and Rieman 1994).

The properties evaluated are first and foremost the technical and functional ability to adjust to the user's need. Another property is the intuitiveness of the user interface. The user interface should act the way the users expect it to act. Hence, it is of great importance that the developers are aware of these expectations (Kristof and Satran 1995). The general reason for evaluation is to find out what the users need or want. They also provide ways of answering questions about how well a product meets the user's need. The more understanding that developers have of users' needs, the better the product will be (Preece et al. 1994). Evaluations can also be helpful when motivating design decisions and measuring the quality of the system (Min-Yang Wang 1992).

A successful evaluation will lead to the introduced product most likely being accepted. Without an evaluation the product reaches the market untested and will therefore only be based on the developers' preconceived notion about what the users want. This may result in the product not meeting all expectations or not living up to the users' demands (Preece et al. 1994). No matter how many analyses have been done in developing the system, experience has shown that there are problems that only appear when the design is tested with users (Lewis and Rieman 1994). (Preece et al. 1994) summarises the reasons for doing evaluations into four main points:

- Understanding the real world
- Comparing designs
- Checking conformance to the predefined target
- Checking conformance to a standard

Since evaluations constitute an aspect of such great importance it would be desirable to keep them present during the entire development process. Unfortunately,

that is not possible in most cases. The primary reason is cost. Therefore the developers have to decide when the need of an evaluation is the greatest. An evaluation can be made at different stages in development, depending on the purpose and the nature of the different components. It is sometimes necessary to compare the product with other products on the market. Evaluations of this kind take place after the product has been developed and are therefore called *summative*. Sometimes the developers need answers to questions in order to check that their ideas are what the users need or want. This kind of evaluation is called *formative* and provides feedback that helps to create a product that will be usable as well as useful. An advantage with a formative evaluation is that it is easier to make improvements of the system early in the development process. Thus, which type of evaluation to choose depends to a great extent on how far the product has been developed (Preece et al. 1994).

4.3 Evaluation methods

Evaluations can be performed by two main types of methods: empirical and formal. In the case of *empirical* evaluation methods, the system is tested by the proposed users. This type of method often leads to good results, but can be very costly and time-consuming. When considering the number of test users, available time and money are often limiting factors. However, it is always better to do an evaluation with a small group of users than not doing any empirical evaluation at all. There are a lot of different types of empirical evaluation methods: scales, questionnaires, interviews, think-aloud protocols etc. (Allwood 1998).

In the case of *formal* evaluation methods the developers test their product themselves, on the basis of guidelines and heuristics. This is often done by an expert using pen and paper. Formal evaluations can be valuable since they do not require real users and are thus faster and cheaper. They do not constitute a guarantee for high usability, though, and should preferably be seen only as a complement to the empirical evaluations (Allwood 1998).

Each evaluation method can further be divided into two different types of studies: qualitative or quantitative. *Qualitative* evaluations are distinguished by simple and straightforward questions, constituted by either interviews or questionnaires, that invite complex and comprehensive answers (Trost 1997). A qualitative evaluation of a word predictor can be about making a detailed study of the text generated by the user and comparing the quality, e.g. the spelling, of the text generated with and without the predictor. It can also be about the users' subjective opinions: if they like the system, if the system makes it easier for them to write, if they feel confident using the system etc. (Palazuelos Cagigas 2001). *Quantitative* evaluations usually involve numbers and statistics (Trost 1997). A quantitative evaluation of a word predictor usually consists of a study of the prediction results according to time and effort saved by the user or the number of predicted words. Word predictors are usually evaluated both qualitatively and quantitatively (Palazuelos Cagigas 2001).

When selecting an evaluation method it is always important to take into consid-

eration concepts like reliability and validity. *Reliability* deals with the demands that the measure should be stable and carried out in a reliable manner. *Validity* means that the evaluation should be valid and measure what was intended and nothing else (Trost 1997).

The difficulty in evaluating assistive technologies is that the users differ as much personally as they do functionally. Each user brings to the evaluation and selection process a unique set of needs and expectations as well as attraction to assistive technology use and readiness for use. The fact that the group of communication disabled persons is very heterogeneous makes it very hard to define a specific group of end-users. When evaluating prediction systems particularly, another problem occurs. Very different results may be obtained depending upon the test conditions and the actual implementation, because several factors, not directly related to word prediction quality, may influence them. (Palazuelos Cagigas 2001) mentions some factors that may influence the result and that makes it difficult to compare different systems:

- Differences among the languages, such as different behaviour of the verbs, different treatment of compounds and other language-specific features.
- System specific features, such as automatic inclusion of white space, auto-capitalisation and suffix prediction.
- Differences in the training and test information, such as to test the performance of the word prediction methods in particular conditions.
- Different measurements, such as time generation rate or keystroke saving rate.
- The method used to obtain the data necessary to calculate the measurements.

To sum up, it is always of great importance that an evaluation is well prepared. If there are users involved, they should be asked to perform the kind of tasks that the system is supposed to support. It is also important that the testing is done with persons whose background approximate those of the system's real users. If the evaluation method is bad or incorrect when testing a prototype, it may result in the prototype fitting the test method, but the real system failing in the real world.

5 User evaluation of FASTY

In this part the data gathered and the evaluation methods used during this study are described. As mentioned in section 4.3 word predictors usually are evaluated both qualitatively and quantitatively. In this evaluation not only the prediction of the words is evaluated, but also the interface of the system into which it is integrated. The evaluation is partly based on a questionnaire and partly on log files. Hence, the questionnaire constitutes the qualitative study and the analyses of the log files constitute the quantitative study.

5.1 Test procedure

As noted in the introduction part, the fundamental steps of the evaluation was set by the FASTY Consortium. The decision was taken, to find at least five test users per language, either adults or children. The fact that this is an evaluation of a rather new type of technology, makes it difficult to use traditional evaluation methods. There does not exist any standard measurements, even though there have been attempts to define such measurements⁷.

After introducing the users to the usage of the system and the purpose and methodology of the verification procedure, individual user dictionaries for each user were created. This was important in order to adapt the predictions individually to the users. The dictionary was created using *make_dictionary*, a program implemented by Mikael Wiberg. This program takes previous texts produced by each single user as input and creates a user specific dictionary consisting of words frequently used by the individual user.

Then each test user received the first prototype of FASTY for eight weeks, starting in the middle of April 2003. During this period, the users were to use FASTY for their every-day typing and communication jobs. The test period was divided into three phases⁸. Each Friday, a report was sent to the technical team concerning possible bugs, the users' behaviour when using the system, software which does not work with FASTY etc.⁹ The technical team proposed a new version of the prototype for the beginning of each new phase. Each version brought solutions to problems discovered during the previous phase.

5.2 Test users

Since this is an empirical study, FASTY was to be tested by the potential users. The prototype PT1 was tested by seven voluntary Swedish test users. Unfortunately, one of the users, UU1, decided to drop off before the end of the test period, so the result presented in the next part will be based on logged data and filled-in questionnaires from six Swedish users. The reason why UU1 still is presented here, is because she

⁷For further information see (Sparck Jones and Galliers 1995).

⁸For the whole schedule please see appendix C.

⁹For the test report template, please see appendix D.

has a specific type of purpose in using word prediction that is interesting to focus on. UU1 also gave well motivated suggestions regarding the redesign phase as long as she took part in the test.

As noted above, the point of testing is to anticipate what will happen when real users start using the system. As stated in section 3, FASTY is intended mainly for persons with motor impairments, but also for linguistically disabled persons. In order to receive a result as realistic as possible, users from both of these groups were represented in the test. (See table 3)

UserID	Age	Gender	Disability	PC usage	Operating system
UU1	30-35	Female	Physical	Average	Windows XP
UU2	35-40	Female	Linguistic	Average	Windows XP
UU3	70-75	Male	Physical	High	Windows XP
UU4	35-40	Male	Physical	High	Windows XP
UU5	35-40	Female	Physical	High	Windows 98
UU6	20-25	Female	Physical	Average	Windows XP
UU7	50-55	Male	Linguistic/physical	Low	Windows NT

Table 3: Description of the test users

They were all adults within a range of 20 to 75 years of age. One user has dyslexia, one user has both aphasia and impaired ability to move, the others have different degrees of motor disabilities. All users use FASTY mainly with Microsoft Word and Excel. Some of them also uses FASTY for Internet applications, such as writing e-mails and chatting. Their PC usage were mainly average. One of the users have operating system Windows 98, another Windows NT, and the rest Windows XP. Two of the users use HeadMouse®¹⁰ and SofType¹¹. One user uses a plastic stick to type on a regular keyboard, and touch pad instead of the mouse. The others use standard keyboards and ergonomic mouses. Only one user has used a word predictor before. The word predictor that was used was the Profet system described in (Carlberger 1997). The users will henceforth be referred to by their userID.

5.2.1 Ethical aspects

It is difficult to predict every situation in which an ethical problem might arise during a user evaluation. Each user who participates in a test has to be informed about the purpose of the test and who is performing the tests. The users must also be informed about how, and in which form, the information will be used and disclosed. Each user must finally be informed about who is going to collect the information and what kind of information is to be collected. Any data that can be linked to an identified or identifiable person counts as personal data. In general personal data will be subject to the user's consent.

¹⁰Head-Controlled Computer Access

¹¹On-Screen Keyboard for Windows 95 through Windows XP that can be accessed using a mouse or mouse emulator such as the HeadMouse®

In the FASTY project, some actions and efforts are stated to be taken, to insure an appropriate ethical standard:

- All proper names will be removed from any material, no matter if it is the author's name or names of persons addressed in the paper. This will be done before processing the data in any way.
- The text collection tool will extract sufficient data instead of using the plain text. Sufficient data could be word lists reflecting word recency, word frequency, or other relevant attributes.
- The data will be stored, whenever possible, in an encrypted format.

5.3 Qualitative study

As mentioned in section 4.3, a qualitative evaluation of a word predictor usually has to do with measuring usability and user satisfaction or to compare the quality of the produced texts.

The ideal way to measure the usability of a product would be to specify the features and attributes required to make it usable. This is usually the approach taken with other software qualities such as functionality, efficiency and portability. The problem with usability is that it is very difficult to specify these features and attributes, mainly because the nature of the criteria required depends on the context in which the product is used (Bevan and Macleod 1994). Which criteria one should give priority to depends foremost, as noted above, on the potential users and the tasks they are supposed to solve by using the system (Min-Yang Wang 1992).

5.3.1 Usability measures

There have been many attempts to describe the features and attributes required to improve usability, including guidelines, checklists, and dialogue principles. High level principles for user interface design are contained in ISO 9241-10¹² The principles are:

- Suitability for the task
- Suitability for learning
- Suitability for individualisation
- Conformity with user expectations
- Self descriptiveness
- Controllability

¹²ISO 9241-10:1995 Ergonomic requirements for office work with visual display terminals: Dialogue principles (ISO:9241-10 1995).

- Error tolerance

These general principles have broad application. It has, however, proved to be a complicated task to formally assess compliance to them, since it is difficult to decide whether and to what extent they apply in certain cases. It is therefore not possible to use them as a basis for measurement.

A description of the quality of use should consist of appropriate measures of user performance. Effectiveness and efficiency are often the primary measures for this, but satisfaction can also be very important, especially when the usage is not restricted to a certain task or environment.

Measures of *effectiveness* concern the goal of the system and measure the accuracy and completeness with which this goal can be achieved. Measures of *efficiency* relate the level of effectiveness achieved to the consumption of resources. The resources may be constituted by mental or physical effort, time, or financial cost. From the users' perspective, the time they spend carrying out the task, or the effort required to complete the task are resources they consume. From the developers' perspective the resource consumed is the cost.

Measures of *satisfaction* describe the perceived usability of the overall system by its users as well as the users' acceptance of the system. Satisfaction may be measured by attitude rating scales, the ratio of positive or negative comments during the use, rate of absenteeism or health problem reports.

There are also usability objectives, such as learnability and flexibility that can be assessed by measuring effectiveness, efficiency and satisfaction across a range of contexts. *Learnability* can be measured by comparing the quality of use for users over time, or comparing the usability of a product for experienced and inexperienced users. *Flexibility* of use by different users for different tasks can be assessed by measuring usability in a number of different contexts. (Bevan and Macleod 1994).

5.3.2 Questionnaire

A qualitative study is, as mentioned, usually based on interviews or questionnaires. What to use mainly depends on how many persons that are to answer the questions. Questionnaires are often used for a larger number of persons. This means that it is of great importance that the questions are clear and unambiguous. To avoid misunderstanding it may be a good idea to provide a pilot study (Preece et al. 1994).

A questionnaire may consist of two different kinds of questions, depending on what type of answers that is required. Questions that have a number of fixed answer suggestions are called *closed questions*. *Open questions*, on the other hand, have no eligible answers. The users are asked to give their comment about the system using their own words. Open questions should be used with caution though, since it may be very time-consuming and cumbersome to analyse them (Trost 2001).

After the end of the test period, the test users of FASTY were asked to fill in a questionnaire about the system's functionality and usability. The questions

were related to the users' experience with the FASTY system during the writing tasks. Questioned were the users' personal opinion, experience and impression. It aimed at comparing the users' experiences with the proposed expectations and needs. The questions were put together by the FASTY Consortium after taking into consideration the international ergonomics norm ISO 9241-10 described above, and after consultation with the test users.

A few weeks earlier a draft questionnaire had been presented to the users, in order to guarantee that the questions were understandable and unambiguous and that the questionnaire met the reliability and validity requirements¹³.

The final questionnaire consisted of 20 closed questions and 5 open questions. The closed questions were constituted by statements, which the users were to agree or disagree with. Each question had four alternative answers for the user to choose among: *strongly disagree*, *rather disagree*, *rather agree* and *strongly agree*. The questions were further subdivided into four subgroups depending on the main subject of them. These subgroups were:

- Use of prediction and other functions during writing
- Adaption of the system
- Documentation, online-help, and tutorial
- General impression

The open questions were mainly about user satisfaction and the users were among others asked to give concrete examples of how to improve the system.

5.4 Quantitative study

For people with physical disabilities a quantitative measure of effort saved by the users when typing is most important. For people with linguistic impairments a quantitative measure of the number of correctly predicted words when typing is a very important aspect. Since users with physical as well as linguistic impairments are to benefit from FASTY, quantitative information concerning both categories is taken into consideration in this evaluation.

5.4.1 Log files

In order to obtain precise measurements during the user tests, a logging functionality was added to the prototype. All relevant events, such as keystrokes made by the user, selection events, changes in parameter setting, etc., were written to a log file. Thus, the log files consists of bottom-line data, i.e. the total record of what the users have typed, and when.

¹³Following this procedure may result in the answers being affected by the fact that the users already have seen the questions once and that they have had a chance to influence on them. It might result in the users getting a more positive attitude to the questions.

An experiment carried out by (Copestake and Flickinger 1998) shows that the use of logged data to evaluate word prediction has both advantages and disadvantages. The main advantage is that logging allows the collection of more realistic data than can be obtained from existing corpora since it is more representative of the type of texts the users wish to write. Further, compared with collecting data in an experimental setting, data logging involves very low overhead for quantity of data obtain. The same logged data can be used in batch mode, i.e. non-interactively, for multiple experiments in order to tune prediction algorithms. Logged conversations must, however, be treated as confidential, so they cannot be widely distributed. There is also a danger of developing systems that are over-adapted. Additionally, simulating the performance of prediction techniques by running in batch mode does not allow for effects of the algorithm on ease of choice of menu items.

The log files that contain sensitive and confidential material were filtered through an anonymity module in order to protect the users identity. This anonymizer, *log-tool0.8.exe* was implemented by Mikael Wiberg. For reasons of respect for the users privacy, logging of events that could be used to reveal the text that has been typed by the user, may be switched off by the user. The logging is turned on or off by pressing the key combination CTRL + l (lower-case L). If turned off, the logger will not record any keystrokes or selections, however, changes of dictionary or parameter settings will still be recorded. The status of logging is indicated in a small window, shown in figure 6 and figure 7, that is placed initially at the right upper



Figure 6: Logging activated



Figure 7: Logging not activated

area of the desktop and can be moved to another location if necessary. Logged data is collected in the separate subdirectory Log. Each session creates a new file named `logn.dat` where $n = 0, 1, \dots, m$. The logs are stored as plain text files. Every line in the log files starts with a numeric record identifier, explained below. Every log file contains only one start record as its first line. After the start line all parameter settings are logged as parameter setting records. User specific dictionary and user

abbreviation file names are logged as well as all relevant settings of the language component. Unfortunately, no settings of the user interface, such as sorting options or position options, are logged. Fields within a line are separated by a tab character. The log record types are:

0: start record

The start record only contains one integer representing the number of seconds elapsed since 00:00:00 on January 1, 1970, Universal Time (UTC)¹⁴. Example:

```
0 1043339095
```

1: keystroke record

The first field of the keystroke record line constitutes the milliseconds elapsed since start time. The second field consists of the decimal ASCII code of the character typed. Thus, the example:

```
1 3129 115
```

represents the typing of character s 3129 milliseconds after the start time.

2: prediction record

The first field constitutes the milliseconds elapsed since start time. The next fields constitute the predictions as plain text strings. The example:

```
2 4600 har han handlar här hade hon hur hos håller hör
```

represents the delivery of 10 predictions in the presented order 4600 milliseconds after the start time.

3: selection record

The first field constitutes the milliseconds elapsed since start time. The second field represents the number of selected predictions in the last prediction record. The number is zero-based which means that 1 constitutes the 2nd word in the prediction list. Hence, the example:

```
3 5600 1
```

represents the selection of han, assuming the previous example to be the last prediction record, 5600 milliseconds after the start time.

4: parameter setting record

The first field constitutes the milliseconds elapsed since start time. The second field consists of the name of a specific parameter. The third field constitutes the on(1) or off(0) status of the parameter as a string. This means that the example:

¹⁴For further information on UTC, Universal Time Conversion, please visit: <http://setiathome.ssl.berkeley.edu/utc.html>

```
4 0034 use_grammar_based_prediction 1
```

represents the setting of parameter `use_grammar_based_prediction` to be activated 34 milliseconds after the start time.

5: current ABEX in use

The first field constitutes the milliseconds elapsed since start time. The second field consists of the name of the used abbreviations file. The example:

```
5 2914 abbreviation2.abr
```

represents the activation of abbreviation file `abbreviation2.abr` 2915 milliseconds after the start time.

6: current dictionary in use

The first field constitutes the milliseconds elapsed since start time. The second field consists of the name of the user dictionary file. Thus, the example:

```
6 2914 new_dict.wd2
```

represents the activation of user dictionary `new_dict.abr` 2914 milliseconds after the start time.

7: number of predictions

The first field constitutes the milliseconds elapsed since start time. The second field constitutes the number of predictions that are presented to the user. It can be either number between 1 and 10. The example:

```
7 2914 5
```

represents the setting of the number of predictions to five, 2914 milliseconds after the start time.

8: Logger state

The first field constitutes the milliseconds elapsed since start time. The second field constitutes on(1) or off(0) status of the logger. The example:

```
8 59916 0
```

represents the deactivation of logging 59916 milliseconds after the start time.

5.4.2 Keystroke savings rate

Logged data can be used to compare the range of word prediction on realistic dataset. It can also be used to compare the actual performance of a word predictor with its theoretical performance (Copestake and Flickinger 1998). The performance of a

word predictor can be measured in several different ways depending on the purpose of the measure.

The success of a commercial word predictor like FASTY is normally expressed in terms of saved keystrokes, i.e. to calculate how many keystrokes the users do not need to type. Thus, the *keystroke savings rate (KSR)* is the percentage number of keystrokes, that a user saves by using a word predictor to type a certain text, compared to the total number of keystrokes that are required to generate the same text without using a predictor. This measure is closely related to reduction of the users' physical effort (Carlberger 1997). (Baroni, Matiasek and Trost 2002b) defines KSR as follows:

$$\text{KSR} = 100 \left(1 - \frac{k_i + k_s}{k_n} \right),$$

where k_i is the number of characters actually typed, k_s is the number of keystrokes required to select among the suggestions in the prediction list and k_n is the number of keystrokes that would be needed if the text was typed without any help from a word predictor. The KSR is usually influenced by the quality of the prediction system and the number of suggestions in the prediction list (Baroni et al. 2002b). Apart from these parameters the size of the vocabulary, variation of subject matter, and complexity of the language in question, such as high inflection, might play a significant role (Willis 2001). KSR cannot be said to be a function of perplexity, but KSR and perplexity are inversely correlated. Thus, if the perplexity is low the KSR is high and vice versa (Carlberger 1997). Neither does saving keystrokes always mean, that the *text generation rate (TGR)* increases, i.e. the time needed to construct the text.

KSR is, however, only an approximation to the ultimate criteria and does not correlate well with saving of time and effort on the part of the user (Copestake and Flickinger 1998). Using a word predictor requires cognitive and perceptual cost, such as time for reading the suggestions in the prediction list and making decisions. A user may also miss a predicted choice, or select the wrong suggestion by mistake. For every user there is a value of KSR below which the use of a word predictor will not increase the TGR (Baroni et al. 2002a). Using KSR as evaluation measure also has the drawback that an exact computation of the KSR is possible only by running a simulation of the prediction process. However, it is the measure that best reflects the benefits a disabled person has when using a word predictor (Baroni and Matiasek 2003).

An ideal word predictor would always suggest the intended word in every context. Thus, that would require one letter keystroke per word. With an average word length of 6.7 characters the KSR would be 85% (Kronlid 2001). State-of-the-art programs for predictive typing in English claim KSR up to 75% (Baroni et al. 2002a). Several studies cited in the literature support the use of word prediction to enhance keystroke savings with experimentally determined keystroke savings' ranges of 37-47% and clinical data to support 23-58% keystroke savings (Klund and Novak 1995). The Swedish word predictor Profet has a KSR of 35%

(Carlberger, Carlberger, Magnusson, Hunnicutt, Palazuelos Cagigas and Aguilera Navarro 1997). Traditional predictive methods such as Antic, Anticipator, PAL, and Predict have reported maximal KSR of 20 to 50% (Shieber and Baker 2003). FASTY aims at a KSR above 50%. That means that a prototypic user using FASTY to type a specific text would only have to use half the number of keystrokes than he/she must use to type the same text without a word predictor.

5.4.3 Collection and calculation

In order to maintain privacy, the users were offered to use the anonymizer before they handed over the log files. After collecting all the produced log files, consisting of the information explained in section 5.4.1, the log files were calculated. This was done by using the program `statistics.pl`, implemented by ÖFAI. This program takes a log file as input and generates start record, the number of keystrokes, the actual length of the text¹⁵, the number of selections from the prediction list, the number of backspace keystrokes, the reached KSR and the average time for keystrokes, selections and predictions. It also generates the written text. An example of a calculated log file may be seen below.

```
O: 0 1052905890
Keys pressed = 146+40=186
Text length = 293
KSR= 36.518771331058 %
Backspaces=10

Avg. Keystroke time: 2226.57534246575 ms
Avg. Selection time: 2423.25 ms
Avg. Prediction time: 700.248756218905 ms
```

```
-----
FASTY är ett projekt med svenskt deltagande som pågår från 2001 till 2003.
Vi skapar ett system som ökar skrivhastigheten genom att förutsäga ord.
Det betyder att personer med funktionshinder kan producera text snabbare,
med mindre ansträngning och bättre stavning och grammatik.
-----
```

`statistics.pl` does, however, not distinguish between the different sorting alternatives of the predictions. Each log file is computed as if it has probability sorting, and hence the out-coming data is incorrect for those with another type of sorting. None of the log files were sorted by word length order so there was only a need for a program that could handle alphabetic sorting. A modified program, `statistics_alpha.pl` implemented by Mikael Wiberg, was used in order to calculate the alphabetically sorted log files. The summation was then done manually.

¹⁵Where text length is the number of characters, not the number of words.

6 Test result

The test results in this study are based on a number of analyses of the log files and on the filled-in questionnaires.

6.1 Questionnaire analysis

As already mentioned, the questionnaire consisted of 20 closed and 5 open questions. The closed questions were further subdivided into the four subgroups described in section 5.3.2. Some explanations of the phrases used in the questionnaire can be found in the glossary in appendix B.

The result of the questionnaire is presented in the figures below. In order to examine differences between different parameters of the users, each question is presented with four different diagrams. The purpose was to be able to compare questions like: Does FASTY satisfy linguistically disabled persons as well as physical disabled persons? Do more experienced computer users have different opinions than less experienced users? Do men have more fun when using the system than women?

One thing to bear in mind when analysing the questionnaire, is that it is not clear if the users have answered the questions with the current prototype or the general system in a complete version in mind.

6.1.1 Use of prediction and other functions during writing

The questions asked under this subgroup have to do with the users' experience of the functionality of FASTY. The general opinion is that the system is easy to use and that the functions are well integrated.

Question 1: It is easy to use the system

As can be seen in figure 8 the majority of the users experienced the system to be easy or rather easy to use. There is no clear difference between the parameter groups.

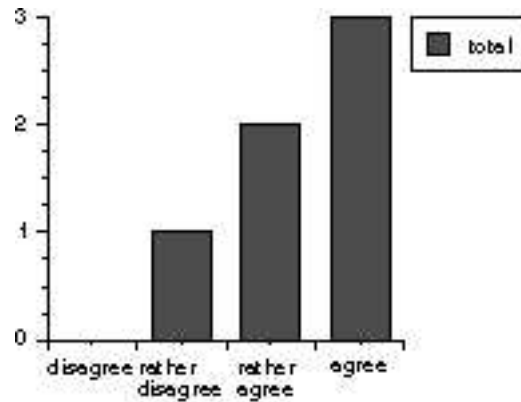


Figure 8: Question 1: Summary

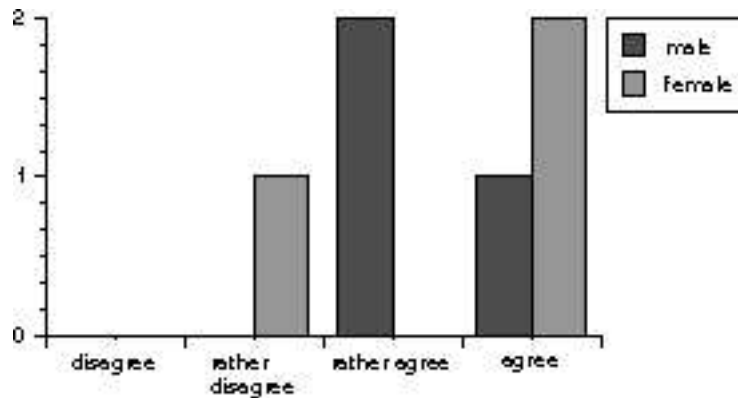


Figure 9: Question 1: Male vs. female test users

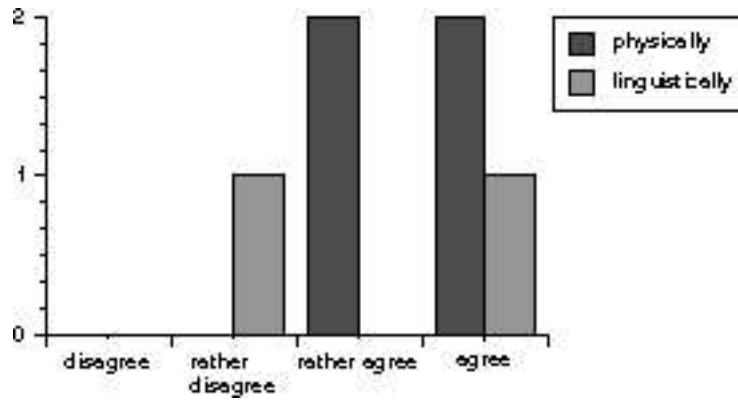


Figure 10: Question 1: Physically vs. Linguistically disabled users

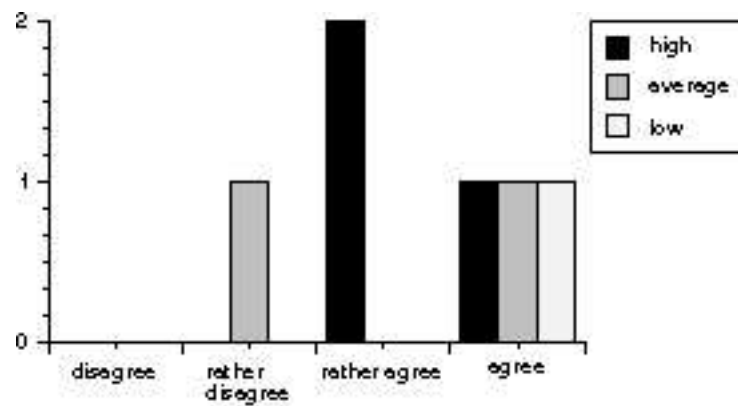


Figure 11: Question 1: Experienced vs. inexperienced users

Question 2: The various functions in this system are well integrated

On this question the users' opinions seem to be rather divided. Figure 13 does however show that the female users have a more positive attitude to the integration of the functions.

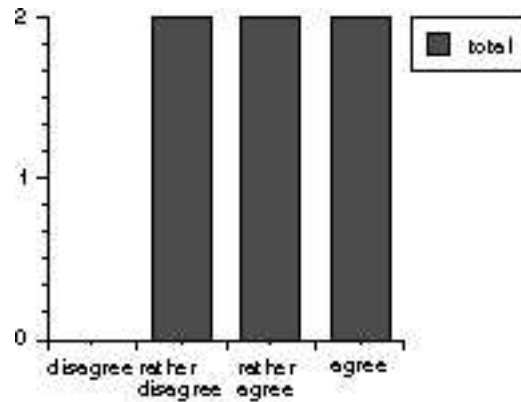


Figure 12: Question 2: Summary

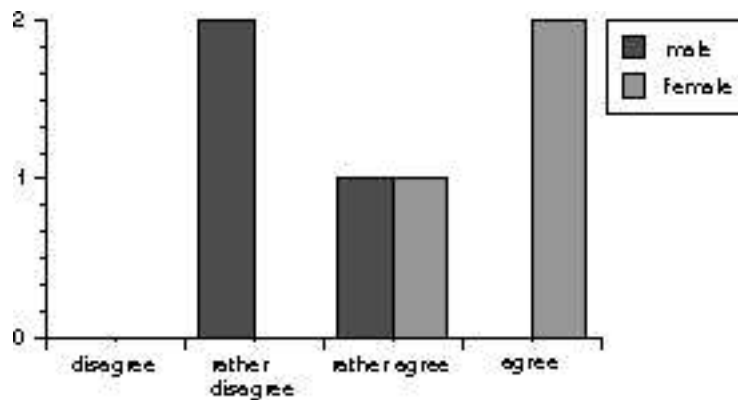


Figure 13: Question 2: Male vs. female test users

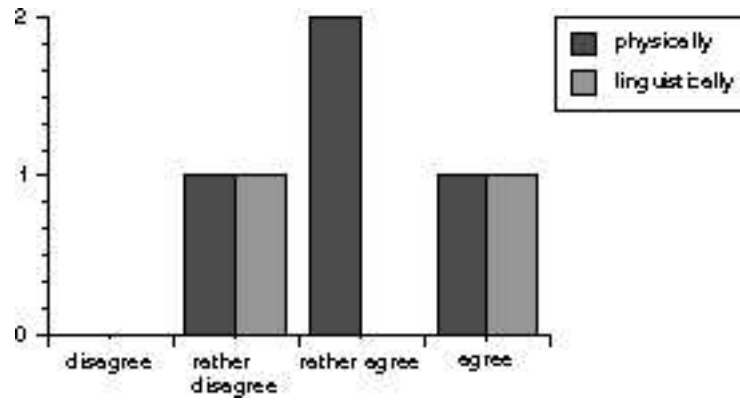


Figure 14: Question 2: Physically vs. linguistically disabled test users

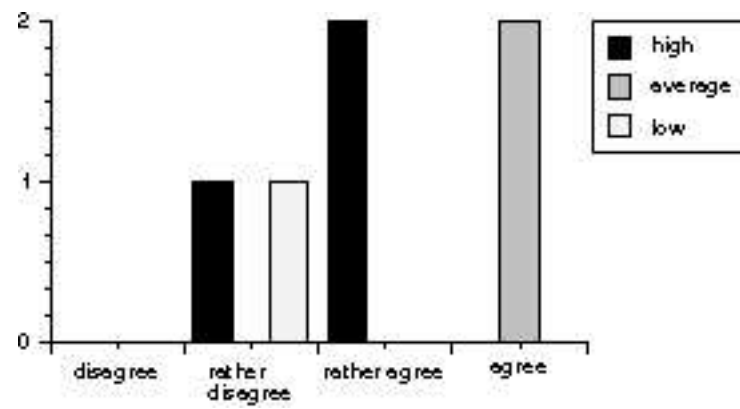


Figure 15: Question 2: Experienced vs. inexperienced users

Question 3: I would like to use the system every day (if writing is demanded)

The users' opinions on this question are rather divided. What can be said is that the linguistically disabled users are slightly more positive towards an every day use of the system.

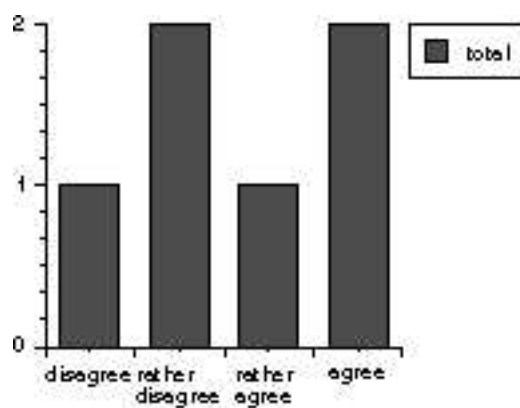


Figure 16: Question 3: Summary

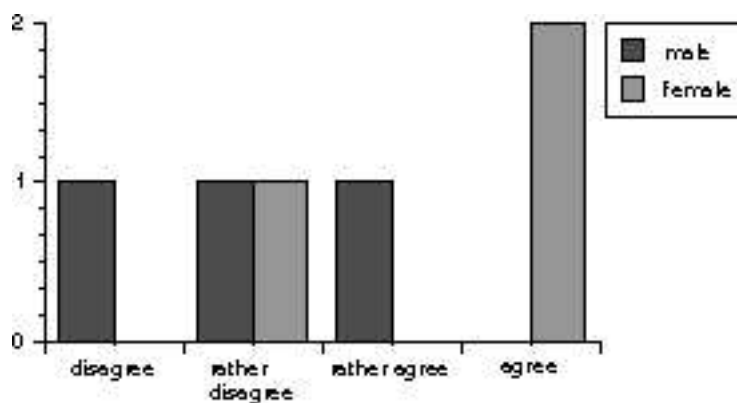


Figure 17: Question 3: Male vs. female test users

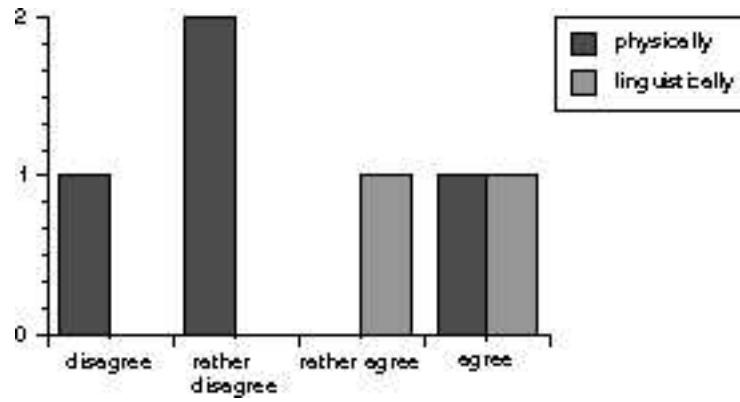


Figure 18: Question 3: Physically vs. linguistically disabled test users

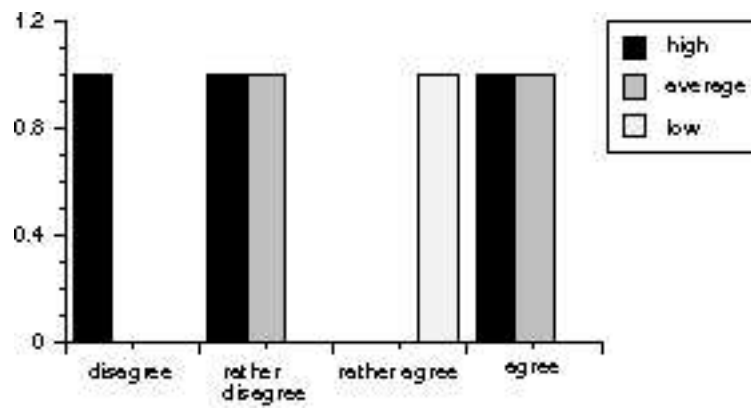


Figure 19: Question 3: Experienced vs. inexperienced users

6.1.2 Adaption of the system

The questions under this subgroup concern the adaptivity and complexity of the system.

Question 4: The system is easy to adjust

Most of the users did not think that the system was easy to adjust. Figure 21 shows that the linguistically disabled are more positive to the adjustments of the system. This may be explained by the fact that they do not need as many adjustments as the physically disabled users do.

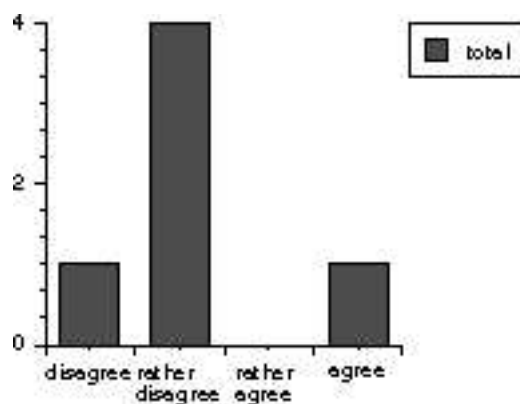


Figure 20: Question 4: Summary

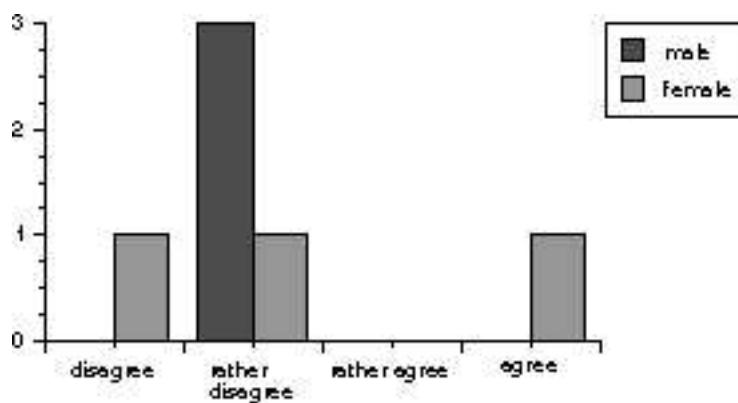


Figure 21: Question 4: Male vs. female test users

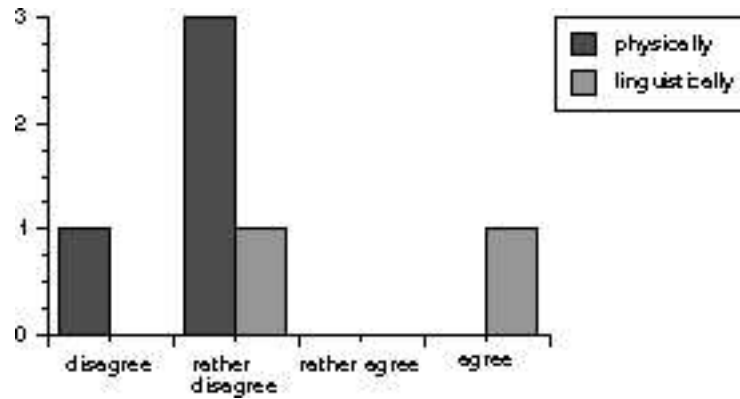


Figure 22: Question 4: Physically vs. linguistically disabled test users

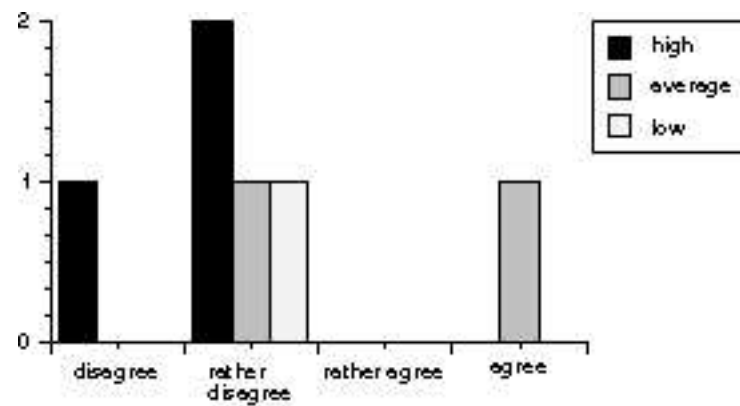


Figure 23: Question 4: Experienced vs. inexperienced users

Question 5: The personal adaptation is well supported by the system

None of the users did fully agree on this question. It is, however, obvious that the users with the most severe physical disability require a better possibility of personal adaptation, while the linguistically impaired users are more, but not fully, satisfied with the current suggest for individual adaptation.

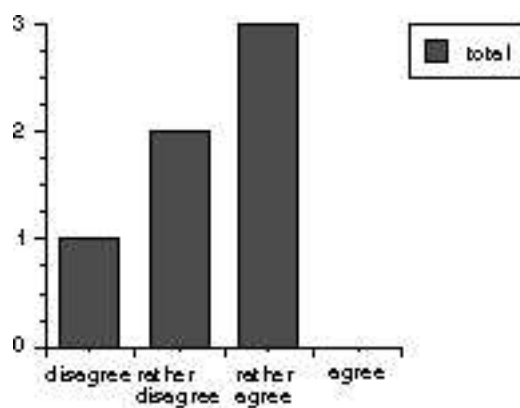


Figure 24: Question 5: Summary

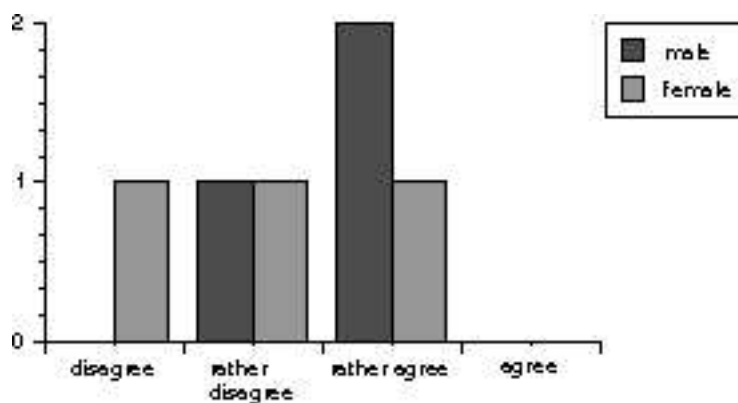


Figure 25: Question 5: Male vs. female test users

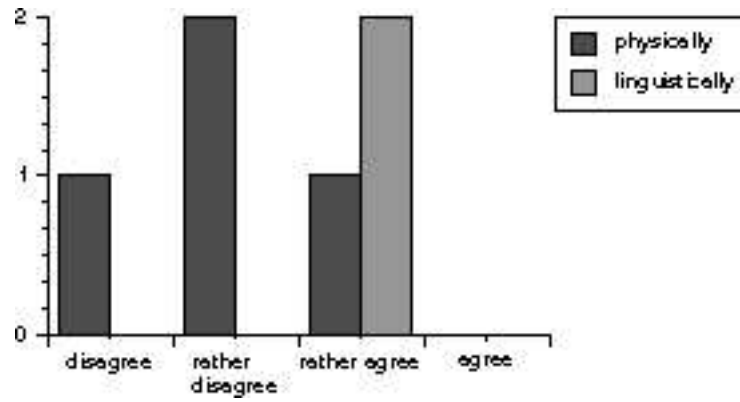


Figure 26: Question 5: Physically vs. linguistically disabled test users

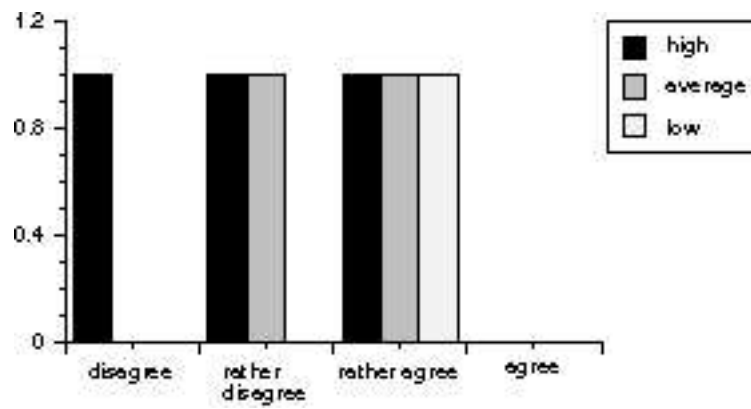


Figure 27: Question 5: Experienced vs. inexperienced users

Question 6: The different functions and settings are clearly arranged in the menus

A majority of the users was satisfied with the arrangement of the menus and settings. The female users were slightly more positive than the male users.

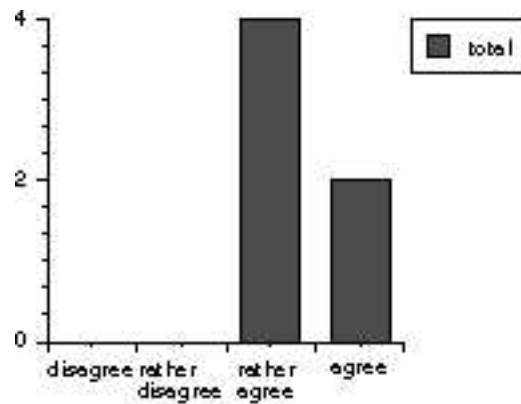


Figure 28: Question 6: Summary

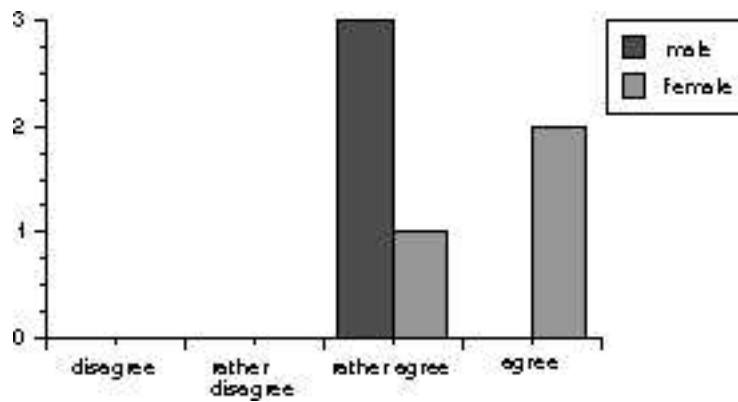


Figure 29: Question 6: Male vs. female test users

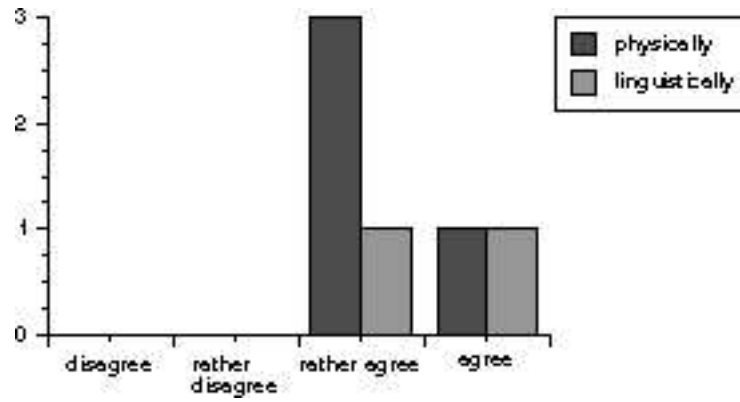


Figure 30: Question 6: Physically vs. linguistically disabled test users

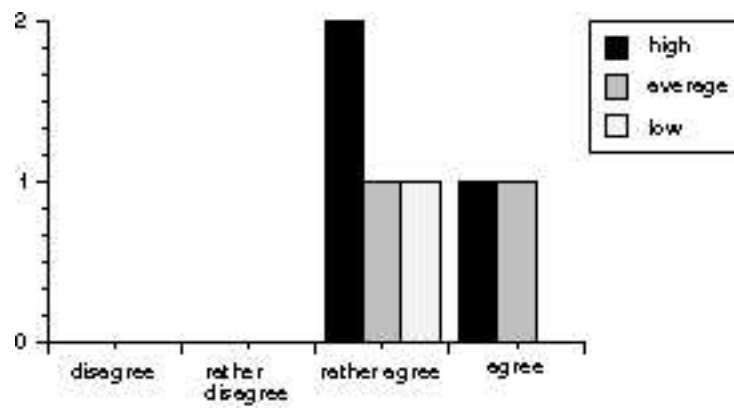


Figure 31: Question 6: Experienced vs. inexperienced users

Question 7: Terms used in the menu are easy to understand

The opinions on this questions were very divided and it is difficult to draw any clear conclusions concerning the different parameters. No user did disagree but only two did fully agree and this may indicate that the terms are not satisfactorily understandable.

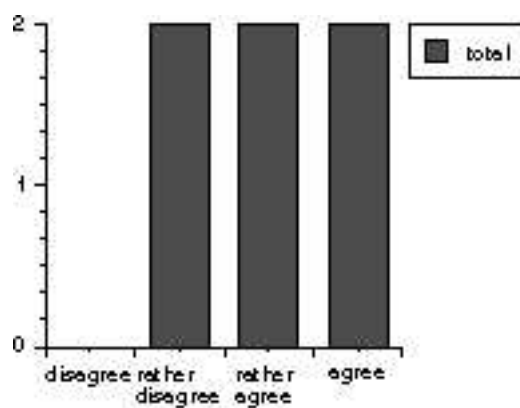


Figure 32: Question 7: Summary

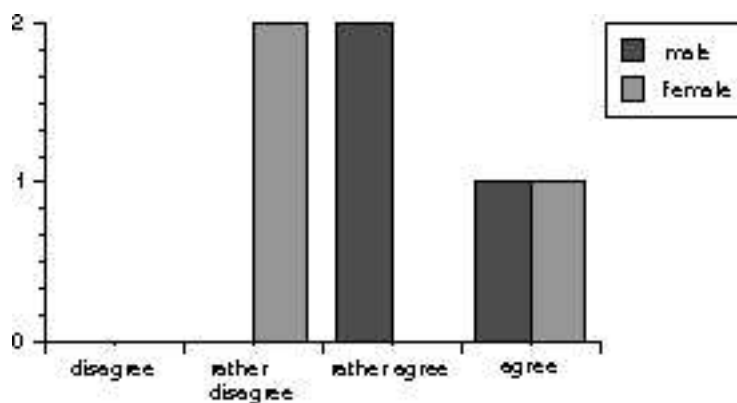


Figure 33: Question 7: Male vs. female test users

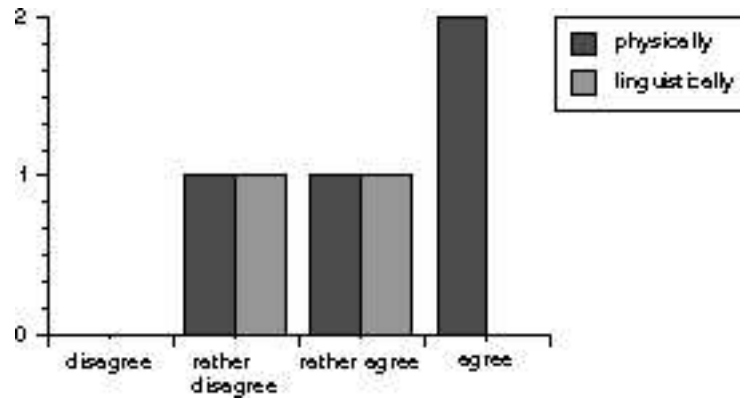


Figure 34: Question 7: Physically vs. linguistically disabled test users

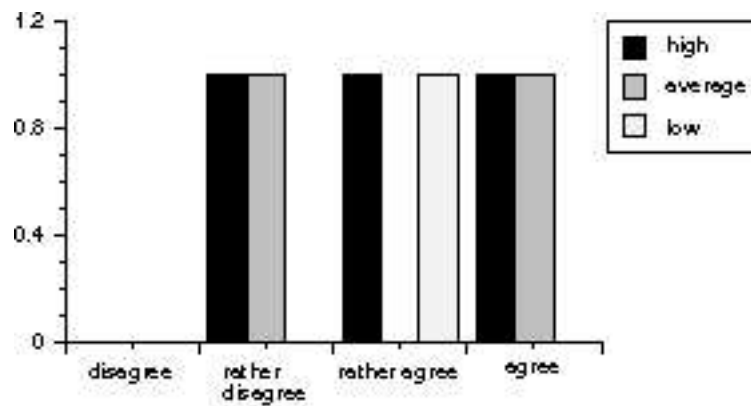


Figure 35: Question 7: Experienced vs. inexperienced users

Question 8: The system is unnecessarily complex

A majority of the users did not think that the system was unnecessarily complex. Figure 39 does, however, show that the inexperienced user finds the system rather complex.

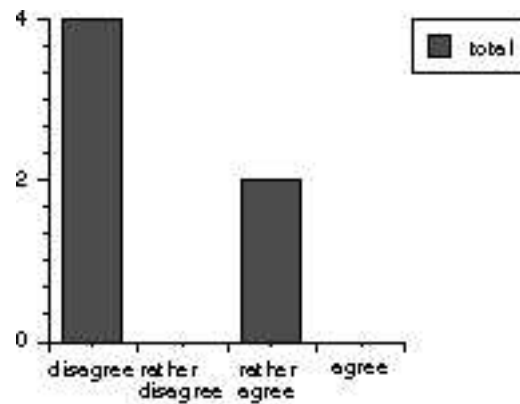


Figure 36: Question 8: Summary

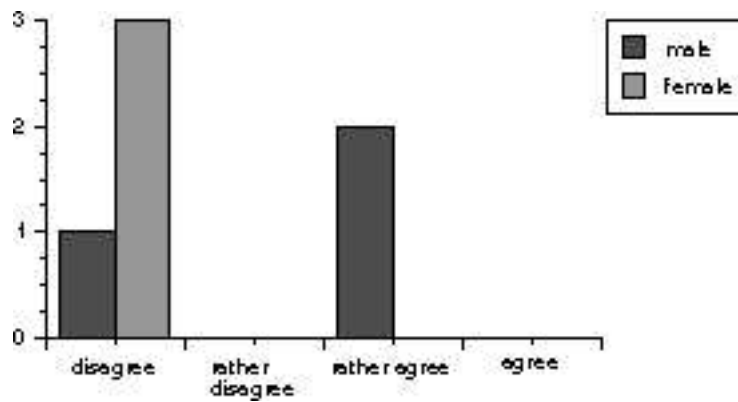


Figure 37: Question 8: Male vs. female test users

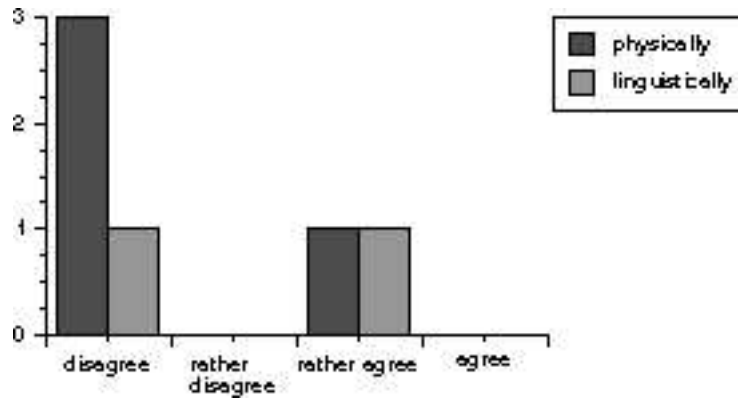


Figure 38: Question 8: Physically vs. linguistically disabled test users

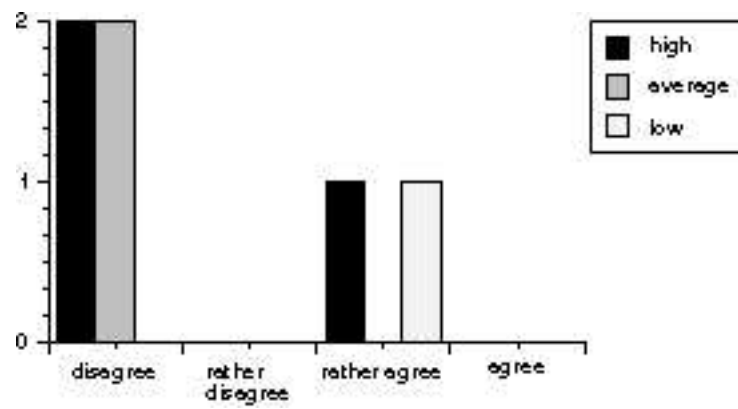


Figure 39: Question 8: Experienced vs. inexperienced users

Question 9: I would need the support of a technical person to be able to use this system

The majority of the users did not need support of a technical person. UU7 does, however, claim that he needs support of a technical person to be able to use the system in an appropriate way. This may indicate that the system has to be adjusted to better fit inexperienced computer users.

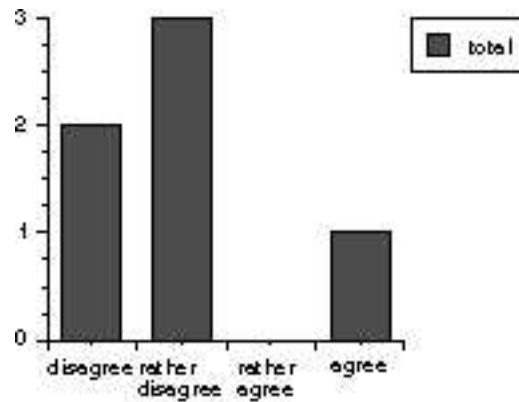


Figure 40: Question 9: Summary

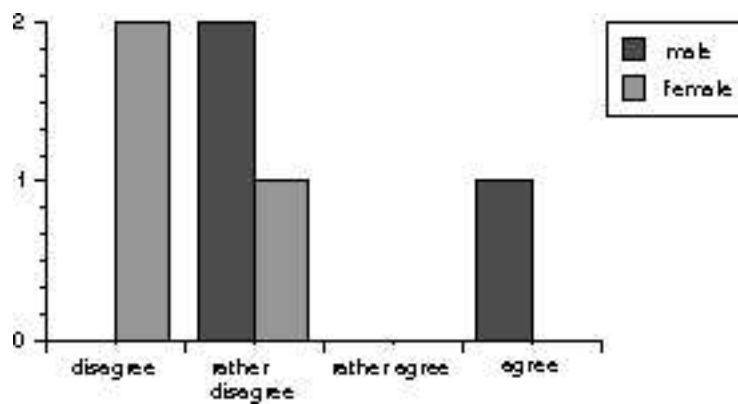


Figure 41: Question 9: Male vs. female test users

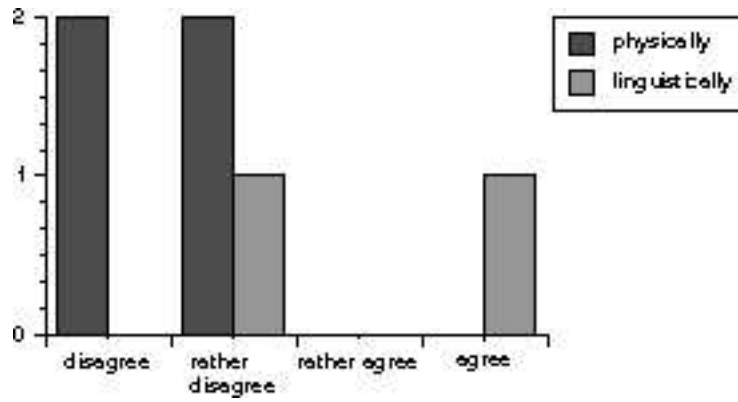


Figure 42: Question 9: Physically vs. linguistically disabled test users

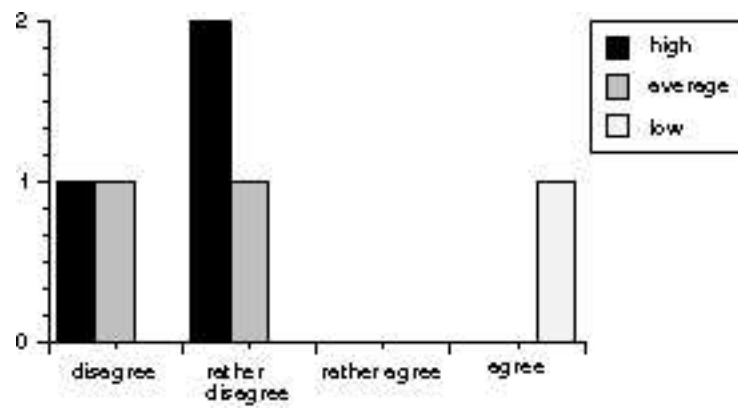


Figure 43: Question 9: Experienced vs. inexperienced users

6.1.3 Documentation, online-help, and tutorial

The results of the questions on this topic show that the users for the most part are very satisfied with the documentation, online-help and tutorial. It is, however, important to mention that their opinions concerning manual and online-help are partly based on a very abbreviated version of the manual made by the author of this thesis and partly on the information available on the FASTY user site¹⁶. The reason for this, is that the real manual and tutorial had not been fully developed for the first test period.

Question 10: Onscreen-explanations in the menu are easy to understand

The opinions on this question were rather divided. Figure 46 shows that the linguistically disabled users found the explanations more easy to understand than the physically disabled users. This was a rather surprising result. Figure 47 does further show that the more inexperienced users were more positive than the experienced users.

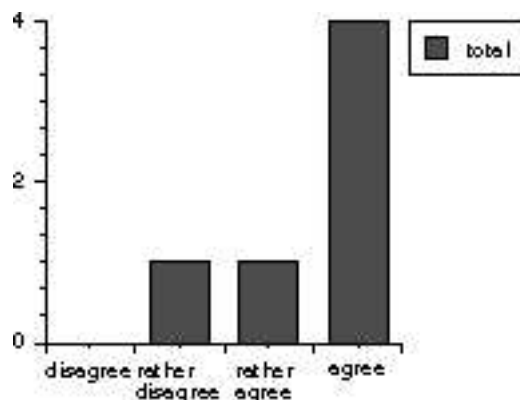


Figure 44: Question 10: Summary

¹⁶The FASTY user site is only available for the test users at <http://reha01lin.iemw.tuwien.ac.at/fastypt1/>

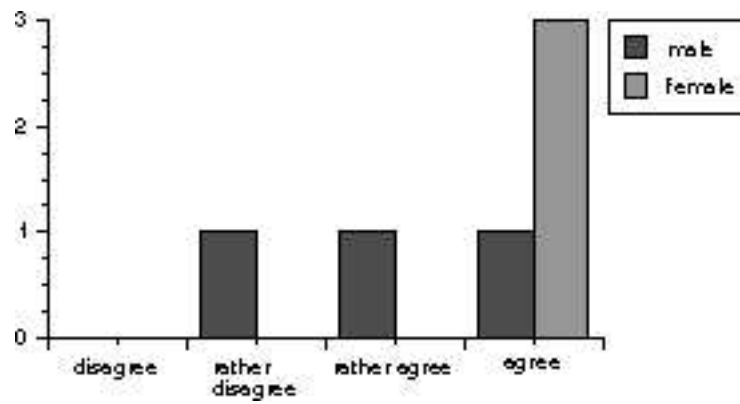


Figure 45: Question 10: Male vs. female test users

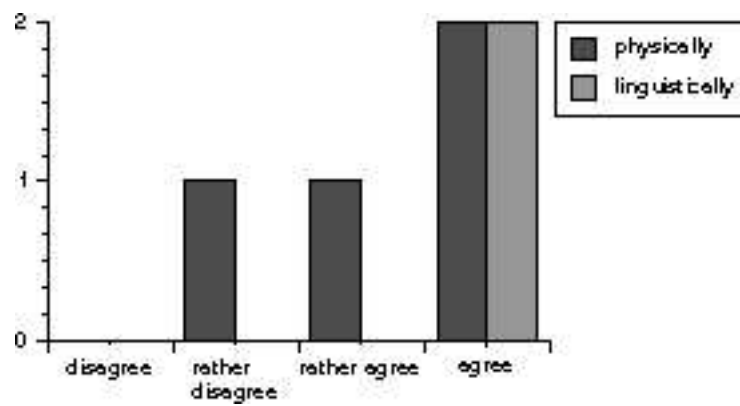


Figure 46: Question 10: Physically vs. linguistically disabled test users

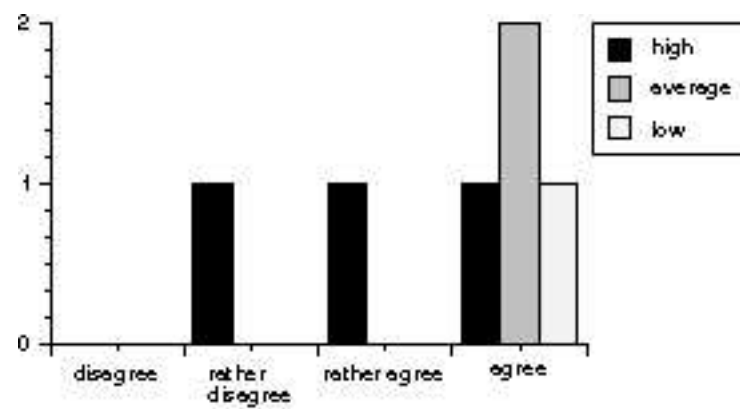


Figure 47: Question 10: Experienced vs. inexperienced users

Question 11: The manual/online-help is easy to understand

A majority of the users agreed with this question.

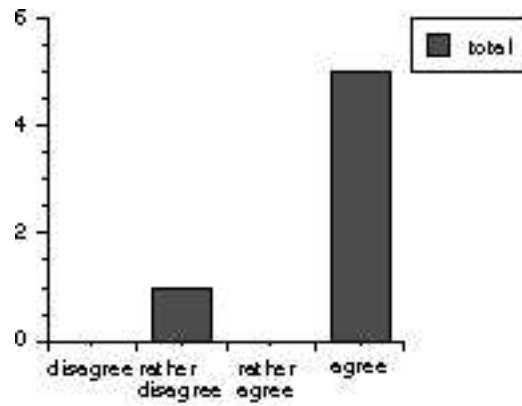


Figure 48: Question 11: Summary

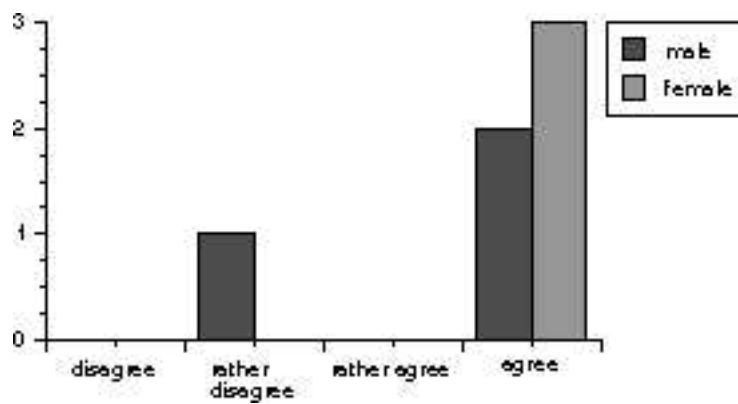


Figure 49: Question 11: Male vs. female test users

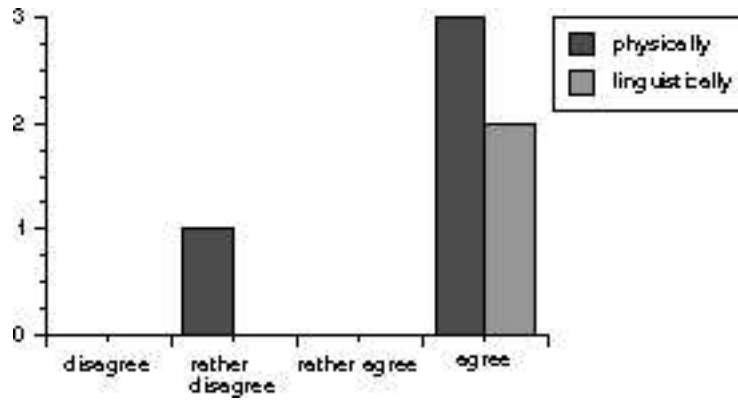


Figure 50: Question 11: Physically vs. linguistically disabled test users

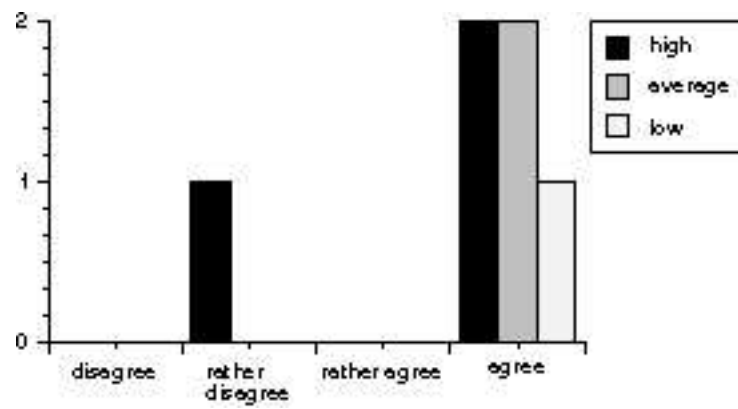


Figure 51: Question 11: Experienced vs. inexperienced users

6.1.4 General impression

These questions concern user satisfaction and user performance. Thus, they are the most important questions in an evaluation of this kind.

Question 12: It is fun to use the system

The result of this question show that the opinions on this question were very divided. Figure 53 does, however, show a clear difference between men and women. Thus, the female users had more fun using the system than the male users.

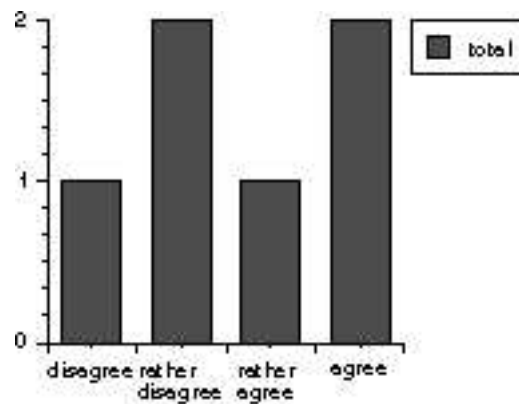


Figure 52: Question 12: Summary

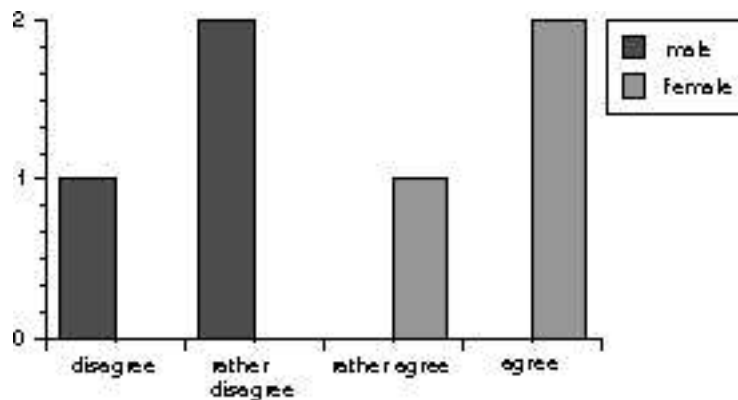


Figure 53: Question 12: Male vs. female test users

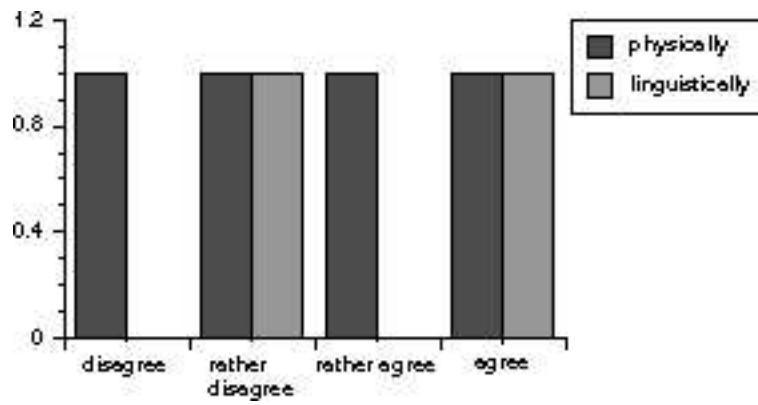


Figure 54: Question 12: Physically vs. linguistically disabled test users

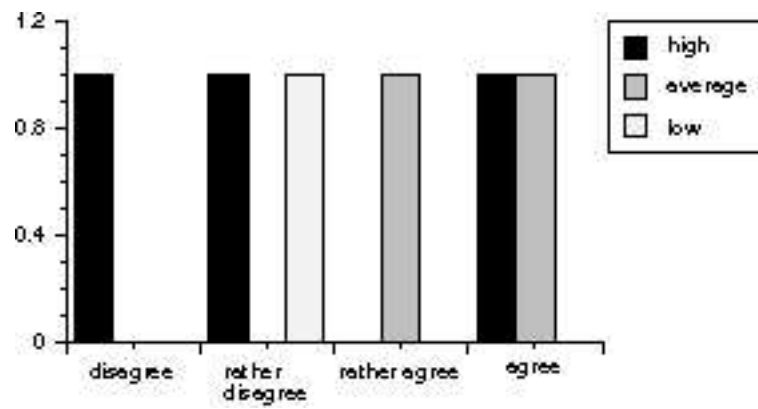


Figure 55: Question 12: Experienced vs. inexperienced users

Question 13: I felt very confident using the system

The users with the highest computer experience did not feel confident using FASTY. This may be explained by the fact that they may be able to understand the problems in an other way than inexperienced users.

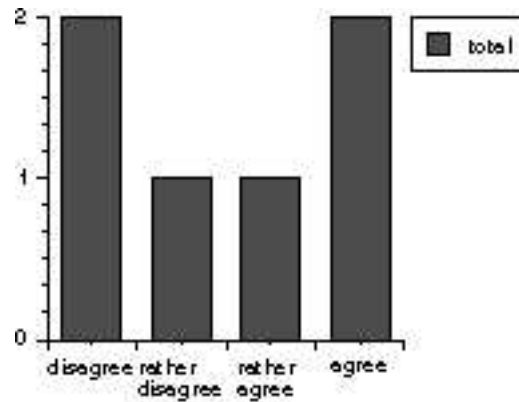


Figure 56: Question 13: Summary

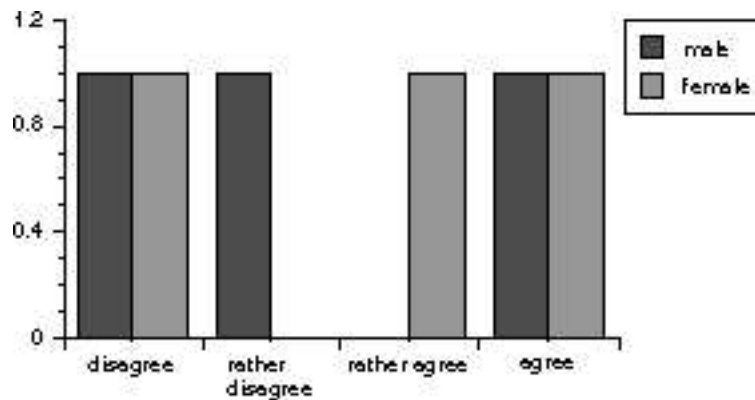


Figure 57: Question 13: Male vs. female test users

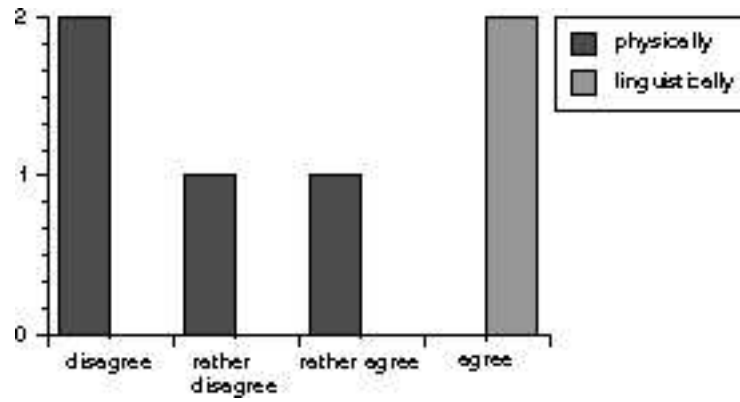


Figure 58: Question 13: Physically vs. linguistically disabled test users

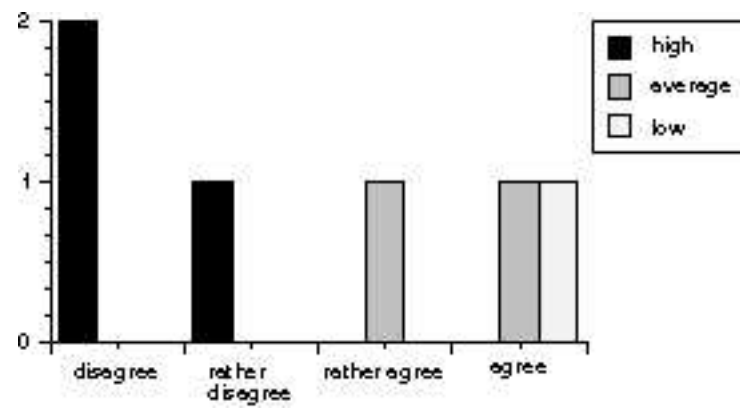


Figure 59: Question 13: Experienced vs. inexperienced users

Question 14: I needed to learn a lot of things before I could get going with this system

Almost every user stated that they did not have to learn many new things about computers in order to be able to use the system. The exception is UU7, who has a low computer experience.

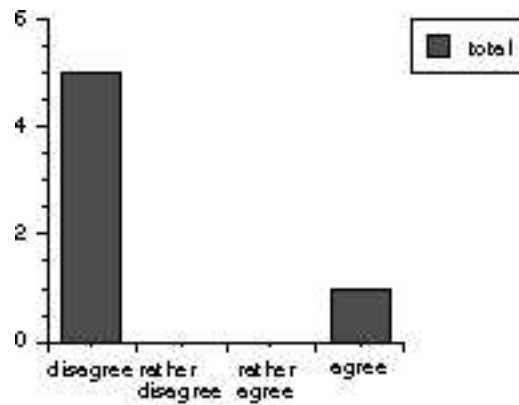


Figure 60: Question 14: Summary

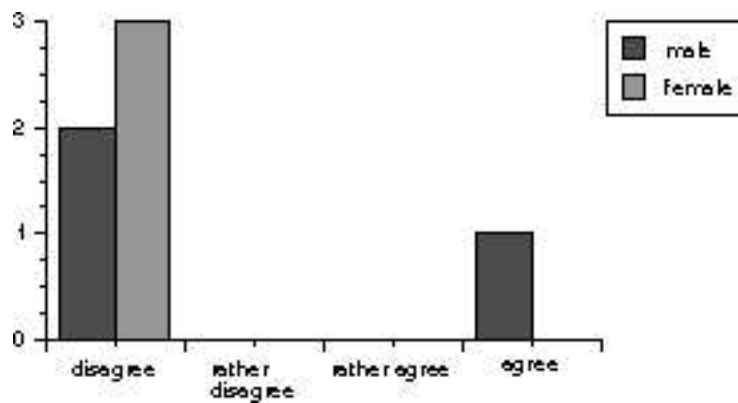


Figure 61: Question 14: Male vs. female test users

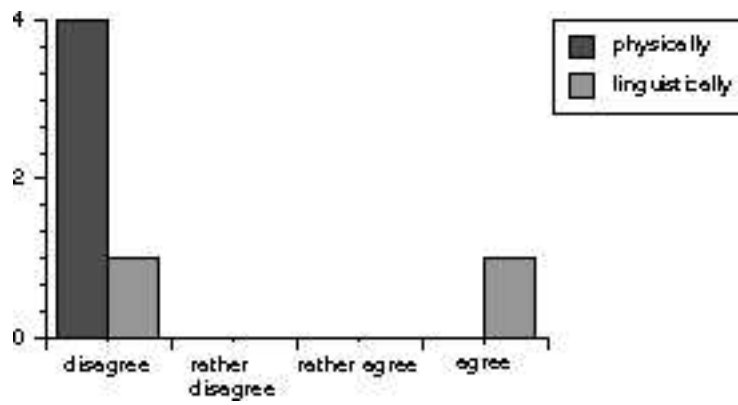


Figure 62: Question 14: Physically vs. linguistically disabled test users

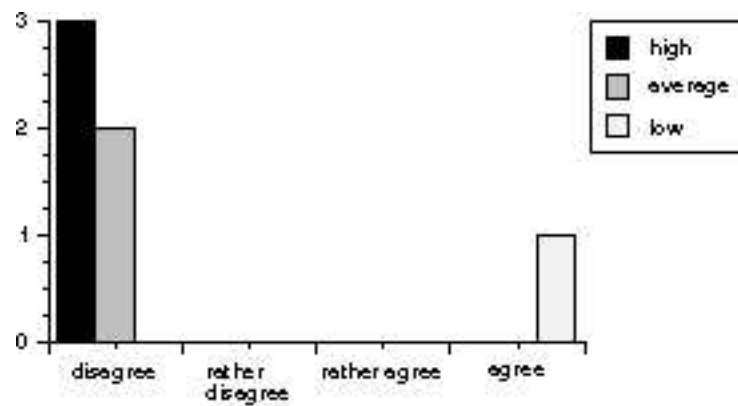


Figure 63: Question 14: Experienced vs. inexperienced users

Question 15: Most people would learn to use this system quickly

Most of the users approve that a lot of people would learn to use the system quickly. As can be seen in figure 67 the exception is also here the inexperienced user.

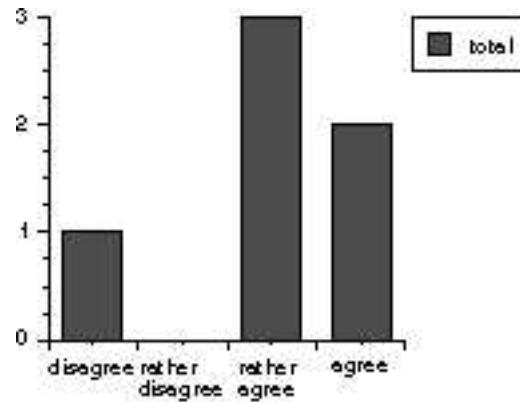


Figure 64: Question 15: Summary

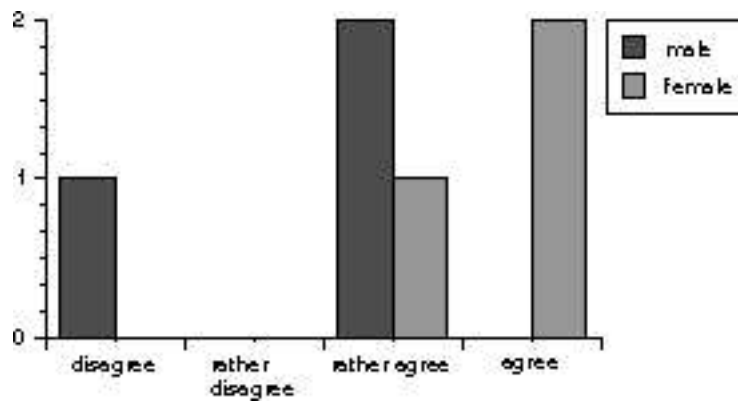


Figure 65: Question 15: Male vs. female test users

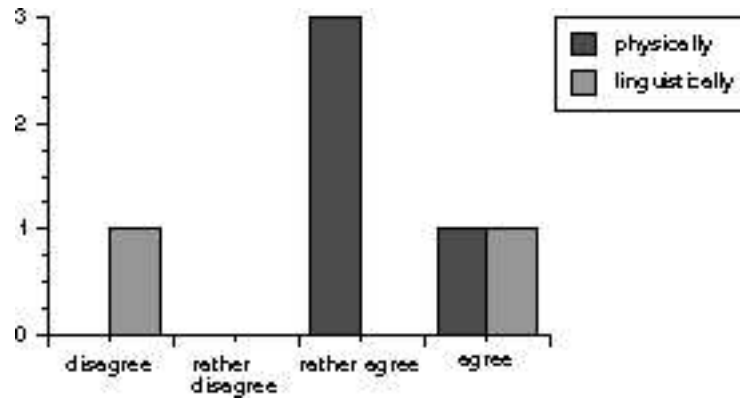


Figure 66: Question 15: Physically vs. linguistically disabled test users

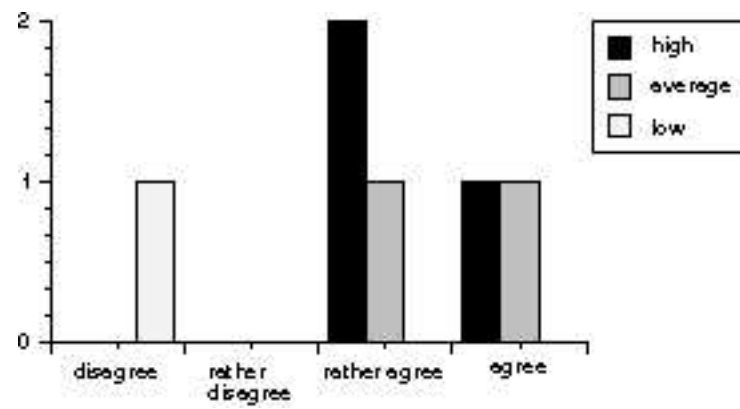


Figure 67: Question 15: Experienced vs. inexperienced users

Question 16: There is too much inconsistency in this system

The opinions on this question were divided. The more inexperienced users were more negative towards this question, though, and experienced that similar or related aspects of the system use different techniques or vocabulary to achieve a result, or that different parts of the system require the user to do the same task in different ways.

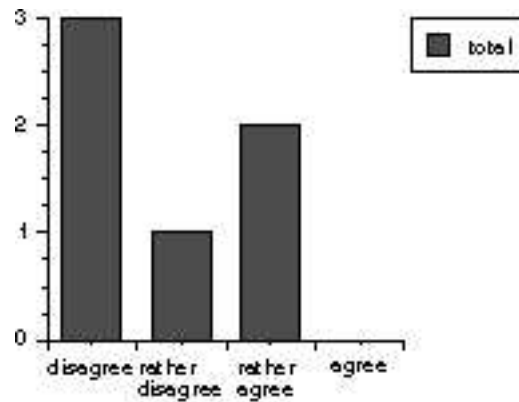


Figure 68: Question 16: Summary

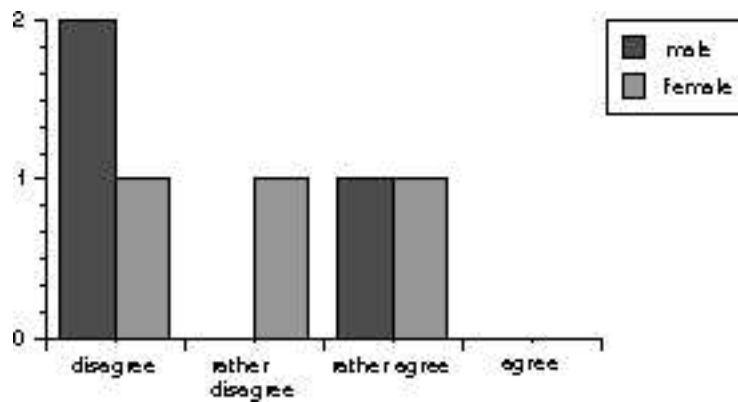


Figure 69: Question 16: Male vs. female test users

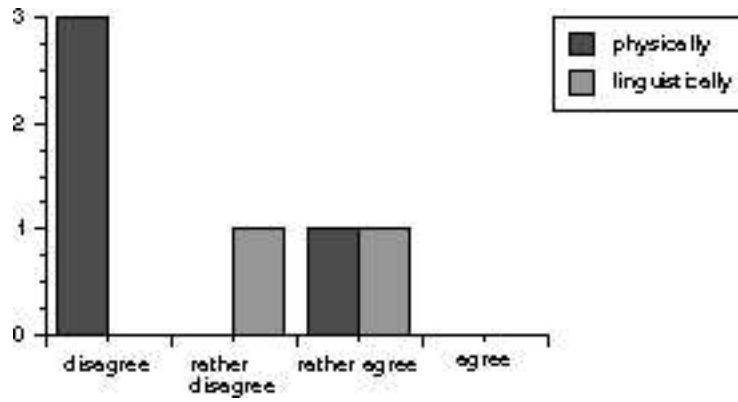


Figure 70: Question 16: Physically vs. linguistically disabled test users

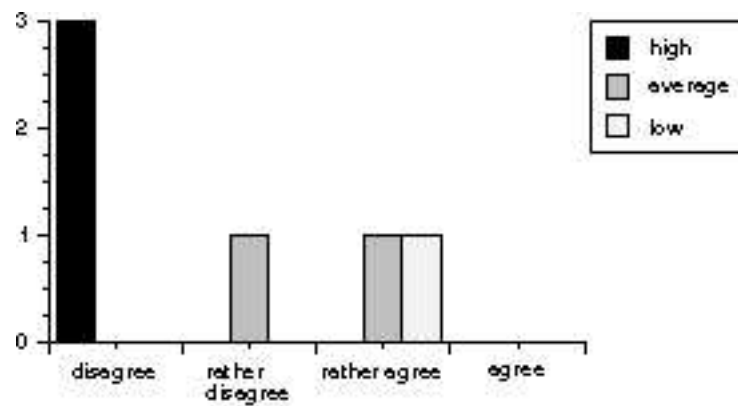


Figure 71: Question 16: Experienced vs. inexperienced users

Question 17: The system is cumbersome to use

A majority of the users did not think that the system was cumbersome to use. Figure 75 shows a clear division between the grade of experience. UU7, the user that had to learn a lot of new things to be able to use FASTY did rather agree with the statement.

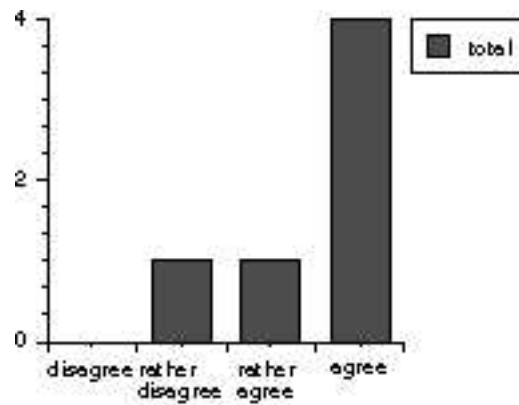


Figure 72: Question 17: Summary

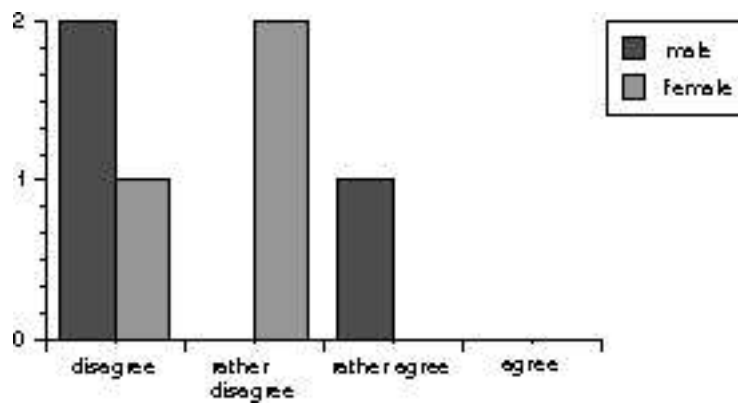


Figure 73: Question 17: Male vs. female test users

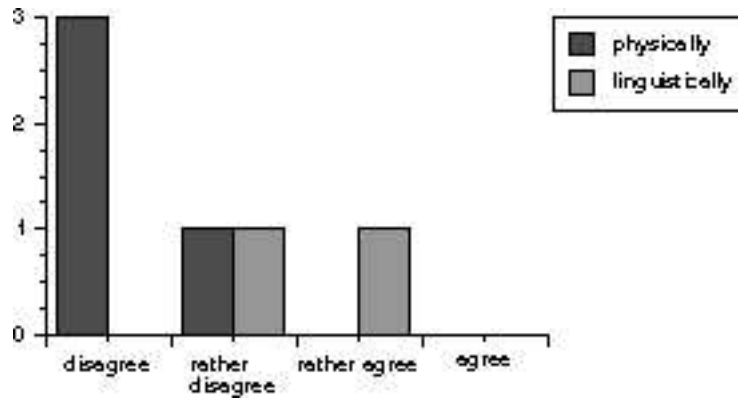


Figure 74: Question 17: Physically vs. linguistically disabled test users

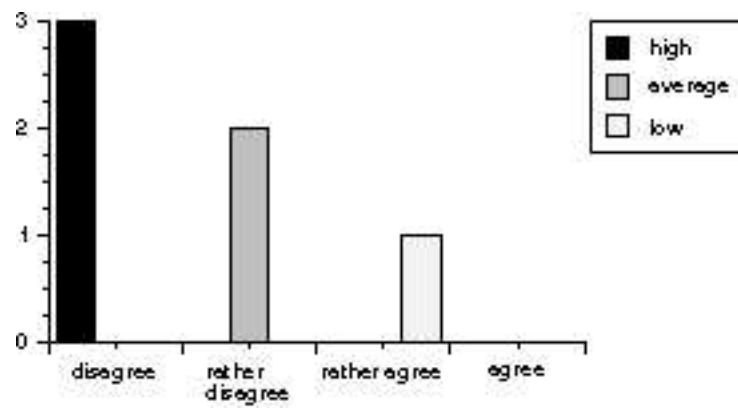


Figure 75: Question 17: Experienced vs. inexperienced users

Question 18: Using the system reduces my necessary efforts in writing tasks

A majority of the users thought that the system reduced their effort in writing. This is a very positive result. What can further be said, is that the linguistically disabled users were more positive than the physically disabled users.

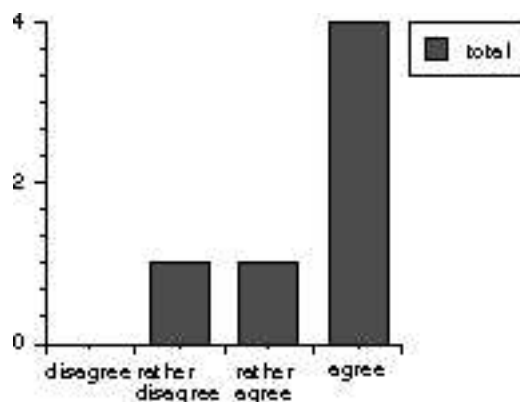


Figure 76: Question 18: Summary

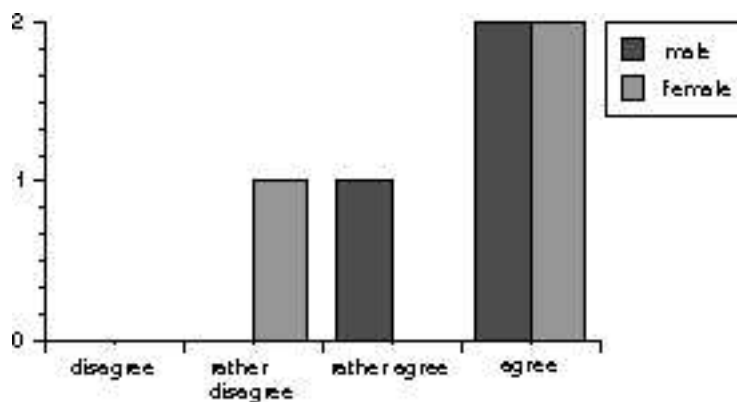


Figure 77: Question 18: Male vs. female test users

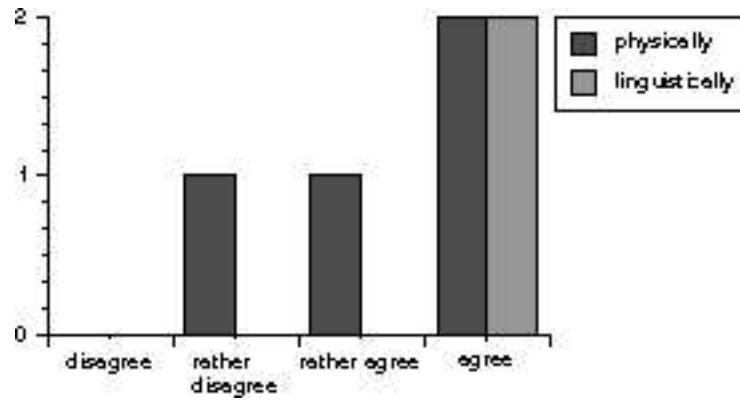


Figure 78: Question 18: Physically vs. linguistically disabled test users

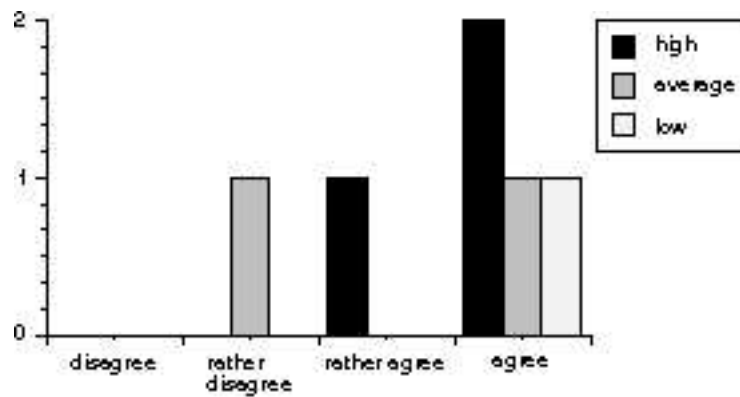


Figure 79: Question 18: Experienced vs. inexperienced users

Question 19: The system improves the quality of my text

This question is mostly intended for the linguistically impaired users, who have troubles in spelling correct and produce grammatically well-formed sentences. UU2 experienced that the quality of the texts gets better when she uses FASTY to type. The system suggests well-motivated words that UU2 herself never had thought about. This is also the case with UU7.

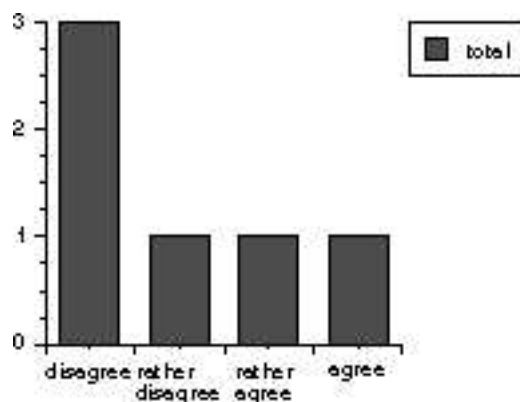


Figure 80: Question 19: Summary

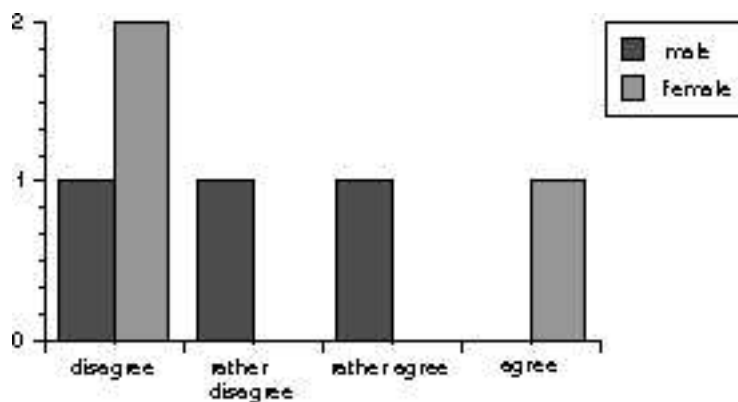


Figure 81: Question 19: Male vs. female test users

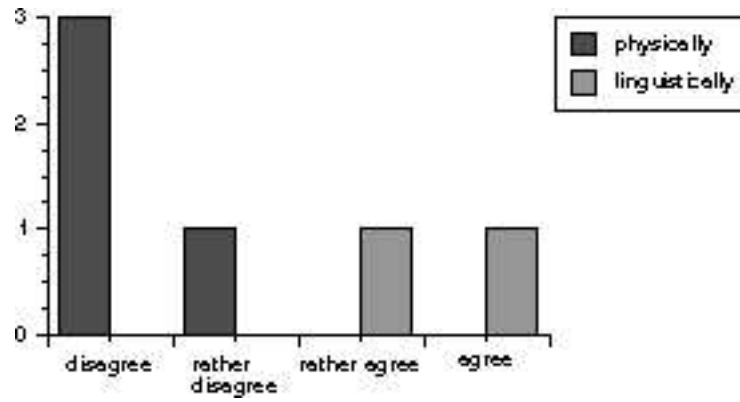


Figure 82: Question 19: Physically vs. linguistically disabled test users

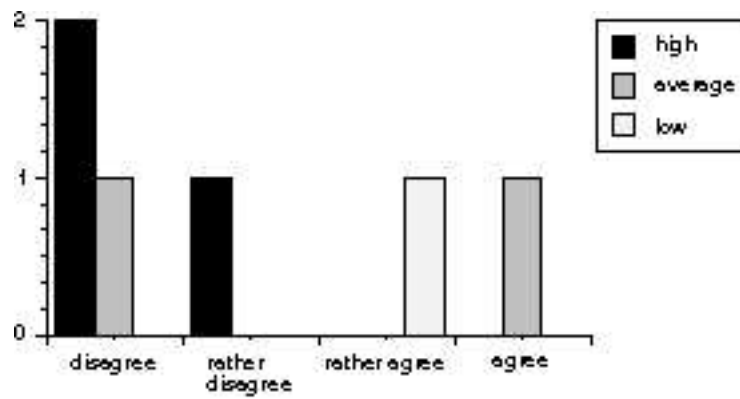


Figure 83: Question 19: Experienced vs. inexperienced users

Question 20: The system improves my text generation rate

The users that have been using FASTY frequently give the most positive response to this question. There is also a difference between the users with physical impairment and those with linguistic impairment. UU2 has the ability to type faster than the other and sometimes the system reacts slower than is satisfactory for her.

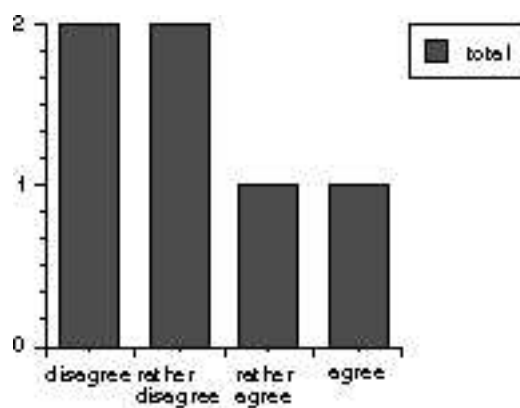


Figure 84: Question 20: Summary

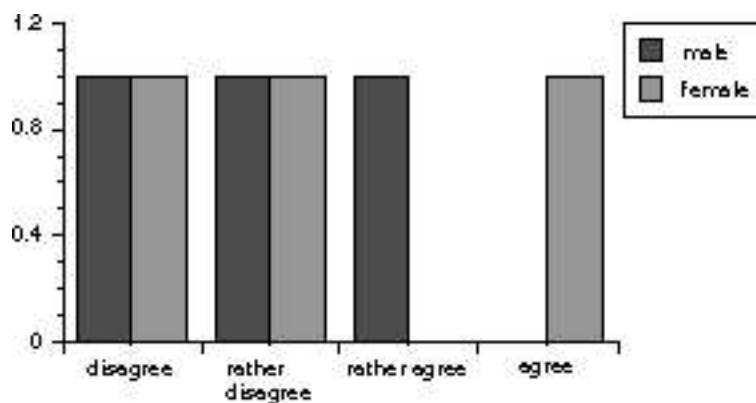


Figure 85: Question 20: Male vs. female test users

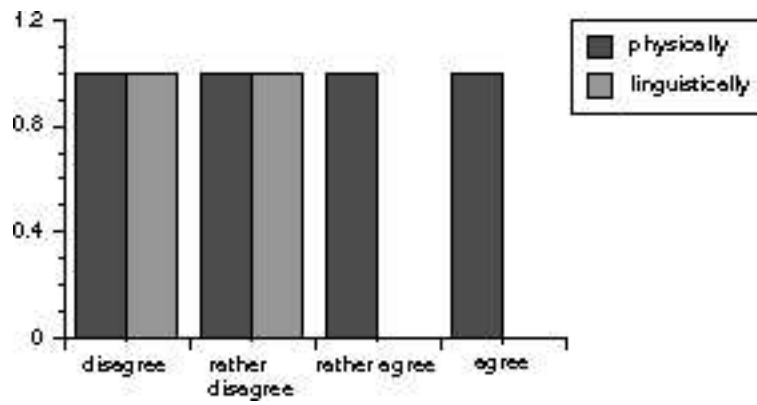


Figure 86: Question 20: Physically vs. linguistically disabled test users

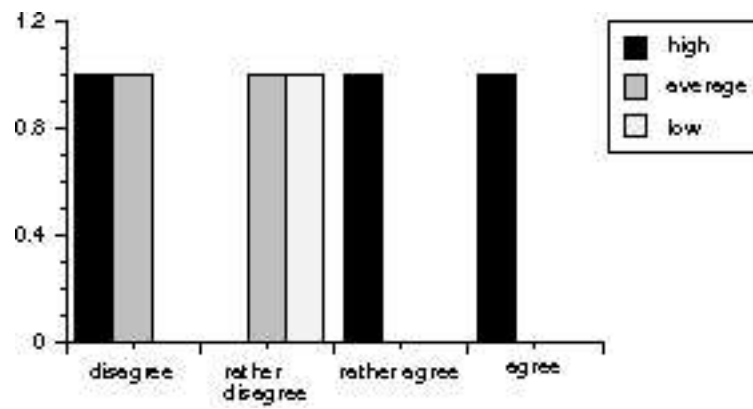


Figure 87: Question 20: Experienced vs. inexperienced users

6.1.5 Open questions

The result of the open questions is presented as lists of statements from the user. The opinions were very divided. Not all users have given an answer to every question. Nearly all users did, however, think that improvements should be done to the current prototype and they also propose some possible improvements. None of the users thought that some features were not useful.

Do you think that the system satisfies your needs as a user?

- Yes.
- Basically, yes.
- No, it does not work properly with SofType.
- It has many bugs in Outlook Express.
- It does not work with Ebookreader.
- Not completely. It is strenuous for me to reach the F-keys and the numpad. It would be better for me to have the ability to choose the keys with which to confirm the predictions.
- Both yes and no.

Do you think that some improvements can be done on the system?

- Yes, I would like the system to start automatically as the computer is turned on.
- Yes, it should work better in combination with other computer-based assistive technologies.
- Yes, it takes a long time to start up the system and it requires a lot of work memory. It is therefore not possible to have several programs running at the same time as FASTY, because the computer locks itself. This should be fixed.
- Yes, the system locks itself when entering Internet. This should be fixed.
- Yes, I would like a more adaptive function, that is trained to remember what the user has typed from time to time and that adaptively learns to recognise the users' own type of language, i.e. dynamically adds new words to the standard dictionary or creates specific user dictionaries as the users produce text.
- Yes, I would like a function that does not add words that only have been typed once to the recency of use function.
- Yes, when using cut and paste the system locks itself. It should be fixed.
- Yes, it is not possible to use short commands with the ctrl-key.

Does the system lack some usable function or feature?

- Yes, I would like a larger liberty of choosing the orientation of the prediction window.
- Yes, I would like an easier way to shift the kind of input between the F-keys and the numpad.
- Yes, I would like to be able to select the prediction with the voice instead of any keys.
- Yes, I would like editable dictionaries or spell checker for the words in the user specific dictionary.
- Yes, I would like to be able to click with the mouse directly on the prediction instead of using the F-keys or the numpad to confirm a prediction.
- Yes, I would like auto capitalisation at the beginning of a sentence.

Do you think that some features are not useful?

- No.
- I am not able to judge.
- No.

Do you have further remarks?

- I would like instructions and on-screen explanations to be written in an easier type of language.
- It was fun taking part in the test.
- I would like to have more support and information at the beginning of the test.
- It is important to carefully specify the end-users that are to benefit from a system like this.
- Changing from alphabetic sorting to probability sorting really increased my text generation rate.

6.1.6 Summary of the qualitative study

To sum up, the questionnaire shows that the users are generally confident that the system is useful. They think that the system is easy to understand, the adjustment possibilities are well structured, and most of them did not need special support in order to use the system. The many bugs of the interface of the tested prototype did however restrict the usability of the system, and it did therefore not convince the users to be well developed.

The questionnaire does also show a significant difference between the users with physical impairment and the users with linguistic impairment. The main purpose of a word predictor for a person with physical impairment is to increase the

text generating speed, i.e. the quantity of text, while the purpose for the linguistically impaired persons is to increase the quality of the text.

Further, there is a difference between UU2 and UU7. For UU2, who can write the first letters of a word with relative accuracy, it can be very helpful in predicting longer, more difficult words.

6.2 Log file analysis

The log files were computed with respect to the keystroke savings measure described in section 5.4.2.

Some of the log files were damaged and incorrect due to bugs of the logging functionality. These log files were filtered. So were also very short log files, with less than 10 typed characters. The evaluation was based on 141 log files but as can be seen in table 4, the number of log files was not uniformly distributed across the participating test users.

UserID	Number of logs	Keys pressed	Text length	KSR	opt.KSR
UU2	13	3613	3974	8.81	16.55
UU3	8	4679	5348	15.68	19.10
UU4	4	1236	1107	-21.15	-7.34
UU5	85	54208	61908	14.17	18.39
UU6	15	12414	13918	5.49	13.36
UU7	16	3645	4187	15.73	23.75

Table 4: Number of log files and average number of characters per log file

The average KSR of the Swedish log files was 12.01%¹⁷. This may seem like a very negative result. When inspecting the log files, it turned out that the users used a lot of backspace keystrokes to erase typing errors. This may to a certain extent explain the low KSR. Sometimes the actual KSR based on text length and keys pressed even had a negative value due to user errors. In order to compensate for this negative impact on the KSR, attempts have been made to calculate the optimum KSR, i.e. the KSR without such unnecessary keystrokes that typing errors produce. The number of keystrokes due to user errors has been computed in the following way:

- If a backspace is typed after a keystroke event, i.e. the last character typed by the user is erased, then the error key count is incremented by 2.
- If a backspace is typed at the beginning of the text, the error key count is incremented by 1.
- Sequences of backspaces after selection events, i. e. correcting or erasing accepted predictions, are not counted as errors.

¹⁷The results are shown in detail in appendix E.

The optimum KSR is thus based on the length of the text and the number of actual keystrokes reduced by the number of erroneously typed keys. The average optimum KSR is 17.6%, which is much better than the actual KSR. The optimum KSR may not reflect the correct value of the KSR, but is still more reliable than the actual KSR.

Studying the log files more in detail concerning situations when backspace is used, it turned out that backspace quite often is used to erase a typed white space after a selection. This may be explained by the user not paying attention to the automatically inserted white space after selection and thus typing another one just by habit. It is however clear that the users that used FASTY more frequently during the test period learned to handle this function better.

Yet another explanation for the poor average result is the data of UU4. Inspecting the log files of UU4 specifically, it turned out that this user had almost not made use of the predictions in the list. (See table 5) Thus, it is rather meaningless

Keys pressed	Text length	Selections	Backspace	KSR	opt.KSR
1236	1107	8	74	-21.15	-7,34

Table 5: Summings-up of the UU4 data

to pay attention to these figures, since the system has not been used the way it was supposed to. When excluding the results of UU4, the actual KSR is 12.98% and the optimum KSR is 18.33%, which is slightly better. However, it is still a much lower KSR than was intended.

6.2.1 Change over time

It may be interesting to see if there was any improvement of the KSR over time. This study was done by calculating the average KSR for each week and comparing the results. It was however only done separately with the log files of UU5. The other users have not used the system frequently enough and have not produced enough log files to make it possible to generate a reliable result. Figure 88 shows the average KSR of UU5 week by week. The large increase of the KSR from week 4 to week 5 can be explained by the change from alphabetic order to probability order. Further, the figure shows that the last three weeks of use, the KSR has been rather stable. This may show that after a certain time of frequent use, the user has found the most suitable setting for him- or herself.

Figure 89 demonstrates the change of the KSR of all users. It is however difficult to draw any conclusions from the result, mainly because of the reasons mentioned above. This figure does however also show that the KSR becomes stable after a certain time of use.

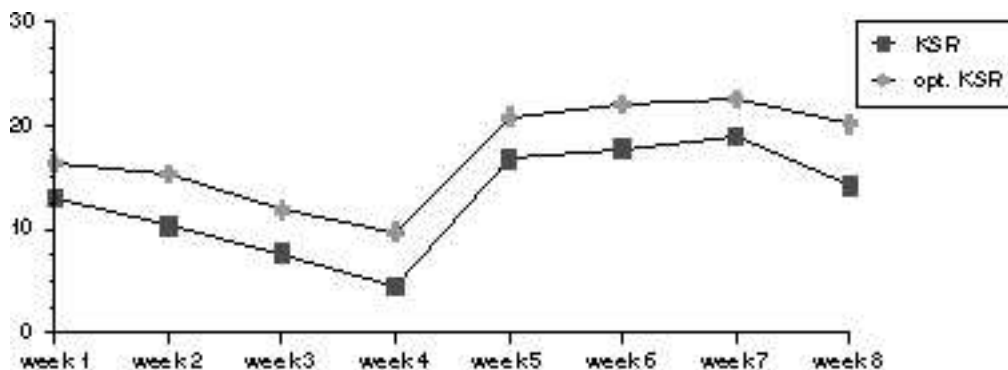


Figure 88: Weekly KSR of UU5

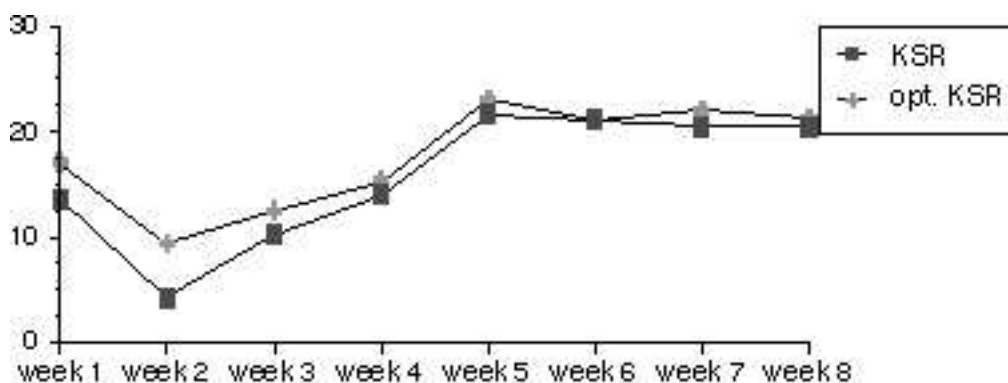


Figure 89: Weekly KSR of all users

6.2.2 Number of predictions

One influential factor of the KSR is the number of predicted words shown to the user. It is obvious that with more predicted words offered, the intended word will appear sooner and less keystrokes are required. This does, however, not always mean that the TGR increases, because of the cognitive load required. Studies have shown that the ideal number of suggestions presented in the prediction window is a list of five words (Klund and Novak 1995).

6.2.3 Sorting

As could be seen in figure 88, the change from alphabetical to probability sorting made a significant improvement of the KSR for UU5. The average KSR for UU5 was almost twice as good with probability sorting as with alphabetic sorting.

It has also turned out that different types of users may benefit from different types of sorting methods. UU2, for instance, was able to better distinguish the suggestions with a list sorted by probability than with an alphabetically ordered list.

6.2.4 Placement and orientation

As explained in section 3.2.6 it is possible to change the orientation of the prediction window in FASTY. A vertical prediction list may be easier to scan visually than a horizontal list. Since there is no information in the log files on what orientation of the prediction window the users have made use of, it is unfortunately not possible to say anything about this.

6.2.5 Grammar checking

The grammar checker in FASTY has been activated during the evaluation phase, except during the last week, when it was deactivated. The reason was to compare the KSR with and without grammar based prediction. Unfortunately, the data from the test was insufficient to use as a base for any conclusions. A formal evaluation made by (Gustavii and Pettersson 2003) using the test platform SWP, does however show a small improvement when grammar rules were activated and the accepted words were promoted. Table 6 and table 7 below show their results when the grammar checker was evaluated in relation to three different kinds of test texts. The test was run with one as well as five suggestions and the improvement was slightly larger with one suggestion in the prediction list than with five suggestions. This shows that the grammar checker reranks the already displayed suggestions rather than adds new suggestions to the list.

Test text	Grammar inactive	Grammar active	Diff (point)	Diff (percent)
Review	34.55	34.89	+0.34	+0.98
Article	31.09	31.82	+0.73	+2.35
Short story	28.17	28.35	+0.18	+0.64
Average	31.27	31.69	+0.42	+1.34

Table 6: Evaluation results for a list length of one

Test text	Grammar inactive	Grammar active	Diff (point)	Diff (percent)
Review	47.78	48.04	+0.26	+0.54
Article	44.31	44.33	+0.02	+0.05
Short story	41.60	41.87	+0.27	+0.65
Average	44.56	44.75	+0.19	+0.43

Table 7: Evaluation results for a list length of five

Despite the poor figures of the test described above, some of the test users in this evaluation claimed a noticeable negative change after deactivation of the grammar based prediction. This holds especially for the linguistically disabled test users who are more dependent on grammatically correct suggestions.

7 Conclusion

As stated in the introduction, the main purpose of this work was to evaluate the first prototype of FASTY and make some recommendation in order to modify the system. FASTY is supposed to reduce the number of keystrokes, enhance the text generation rate and give support for linguistic skills, such as spelling and syntax. This evaluation shows that these goals are not fully achieved in the current prototype of FASTY. This may lead to a low grade of user satisfaction and acceptance of the system.

The effect of a word predictor is unique to each individual user. It is dependent on the characteristics of the user, the cost versus benefits of using the word predictor, and of course the characteristics of the word predictor itself. It is most likely the case that if a person does not see how the assistive technology can help him or her achieve desired goals and dreams, the use of assistive technology has less appeal. People with no physical or linguistic disabilities will probably not have any use of FASTY, since typing on a normal keyboard is an efficient way of entering text into a computer. FASTY is most useful when text input is very slow and difficult.

It may, on the other hand, have positive side-effects, such as reduction of misspellings and typographic errors. Sometimes it is more valuable to let the aspect of typing speed give priority to creation of qualitatively better texts. In order to make people use FASTY it has to be clear how and to what extent the system may help the specific users.

The questionnaire analysis did show that FASTY is not as suitable for inexperienced computer users as would have been desirable. The interface of the current prototype neither fulfils the intended goal of intuitiveness nor consistency. This conflicts with the dialogue principles explained in section 5.3.1. A system has to be easy to learn, so that the users can start the work as soon as possible. Inconsistency makes the system difficult to use and less powerful. Compared to other predictive products on the market, the prediction power of FASTY is almost as good as the competitive products, but the current user interface is far from as good as the competitive ones. Therefore efforts have to be made to improve the learnability and make the system more consistent and intuitive. This may, as a first step, be done by using a simpler language in the onscreen-explanations and providing educational tutorials.

Another way to improve the usability, may be to decrease the number of configurable settings. As mentioned above, earlier studies have shown that certain settings, concerning for instance the the number of predictions, are cognitively better than others. Considering the results of these studies, the scientifically best settings can be prioritised. This would decrease the cognitive load of the users, and hence increase the productivity and satisfaction of use.

Now, the cost of correction against the severity of each problem has to be balanced. Then the system has to be revised and tested again in order to get it in use. A final user test will be carried out in the end of 2003 and then the system will be marketed.

References

- Allwood, M. C. (1998). *Människa-datorinteraktion: Ett psykologiskt perspektiv*, Lund: Studentlitteratur.
- Baroni, M. and Matiasek, J. (2003). Exploiting long distance collocational relations in predictive typing, *Proceedings of the EACL Workshop on Language Modeling for Text Entry Methods*, Budapest, Hungary, pp. 1–8. Available online at <http://www.oefai.at/john/papers/eachws03.pdf> (2003-05-28).
- Baroni, M., Matiasek, J. and Trost, H. (2002a). Fasty - a multi-lingual approach to text prediction, in J. Klaus, K. Miesenberger and W. Zagler (eds), *Computers Helping People with Special Needs: 8th International Conference, ICCHP 2002, Linz, Austria*, Vol. 2398, Springer-Verlag Berlin-Heidelberg-New York, pp. 243–250. Available online at <http://www.oefai.at/~john/papers/ICCHP-02.pdf> (2003-02-28).
- Baroni, M., Matiasek, J. and Trost, H. (2002b). Predicting the components of german nominal compounds, *Proceedings of the 15th European Conference on Artificial Intelligence, ECAI 2002*, IOS Press, Amsterdam, pp. 470–474. Available online at <http://www.oefai.at/~john/papers/ecai02.pdf> (2003-02-28).
- Bevan, N. and Macleod, M. (1994). Usability measurement in context, *Behaviour and Information Technology* **13**: 132–145. Available online at <http://www.usability.serco.com/papers/music94.pdf> (2003-04-23).
- Carlberger, A., Carlberger, J., Magnusson, T., Hunnicutt, S., Palazuelos Cagigas, S. E. and Aguilera Navarro, S. (1997). Profet, a new generation of word prediction: An evaluation study. Available online at <http://acl.ldc.upenn.edu/W/W97/W97-0504.pdf> (2003-08-14).
- Carlberger, J. (1997). *Wordpredict: Design and implementation of a probabilistic word prediction program*, Master's thesis, Royal Institute of Technology, Stockholm.
- Carlberger, J. and Hunnicutt, S. (2001). Improving word prediction using markov models and heuristic methods, *Augmentative and Alternative Communication* **17**: 255–264.
- Copestake, A. (1997). Augmented and alternative nlp techniques for augmentative and alternative communication, *Proceedings of the ACL workshop on Natural Language Processing for Communication Aid*. Available online at <http://acl.ldc.upenn.edu/W/W97/W97-0506.pdf> (2003-03-18).
- Copestake, A. and Flickinger, D. (1998). Evaluation of nlp technology for aac using logged data, in J. C. Filip Loncke and L. Lloyd (eds), *Proceedings of*

the 1998 ISAAC Research Symposium, Dublin, Ireland, Whurr Publishers, London. Available online at <http://www-csli.stanford.edu/aac/evaluation-dsp1.htm> (2003-06-17).

Fazly, A. (2002). *The use of syntax in word completion utilities*, Master's thesis, University of Toronto, Toronto.

Gustavii, E. and Pettersson, E. (2003). *A swedish grammar for word prediction*, Master's thesis, Department of Linguistics at Uppsala University, Uppsala. Available online at http://stp.ling.uu.se/matsd/thesis/arch/2003_gustavii_pettersson.pdf (2003-07-01).

ISO:9241-10 (1995). Ergonomic requirements for office work with display terminals (vdts): Dialogue principles.

Klund, J. and Novak, M. (1995). If word prediction can help, which program do you choose? Available online at <http://trace.wisc.edu/docs/wordprediction2001/> (2003-03-03).

Kristof, R. and Satran, A. (1995). *Interactivity by design*, Mountain View, CA: Adobe Press.

Kronlid, F. (2001). Prediction and nlp - term paper for the course nlp 1. Available online at http://www.ling.gu.se/~kronlid/term_paper/nlp_paper.pdf (2003-03-14).

Kronlid, F. and Nilsson, V. (2000). *Treepredict*, Master's thesis, Gothenburg University, Gothenburg.

Laine, C. J. and Bristow, T. (1999). Using manual word-prediction technology to cue student's writing: Does it really help? Available online at <http://www.csun.edu/cod/conf/1999/proceedings/session0067.htm> (2003-03-03).

Lesh, G. W. and Rinkus, G. J. (2001). Domain-specific word prediction for augmentative communication. Available online at <http://www.enkidu.net/downloads/papers/LeRi01.pdf> (2003-03-03).

Lewis, C. and Rieman, J. (1994). *Task-centered User Interface design. A practical introduction*, Available at <ftp.cs.colorado.edu> (2003-02-07).

McCoy, K. F. (1998). Interface and language issues in intelligent systems for people with disabilities. Available online at <http://www.cis.udel.edu/~mccoy/publications/1998/McCoy98-ATBook.pdf> (2003-03-03).

Min-Yang Wang, E. (1992). *Usability Evaluation for Human Computer interaction*, Phd dissertation, Luleå Universitet, Luleå.

- Nielsen, J. and Mack, R. L. (1994). *Usability Inspections Methods*, Boston.
- Palazuelos Cagigas, S. E. (2001). *Contribution to word prediction in Spanish and its integration in technical aids for people with physical disabilities*, Phd dissertation, Madrid University, Madrid.
- Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. (1994). *Human-Computer Interaction*, Addison-Wesley Publishers Ltd., Wokingham, England.
- Raskind, M. H. and Shaw, T. (1999). Assistive technology for individuals with learning disabilities, *Center On Disabilities, Technology And Persons With Disabilities Conference 1999*. Available online at <http://www.csun.edu/cod/conf/1999/proceedings/csun99.htm> (2003-03-03).
- Rizer, B., Cirlot-New, J. and Ethridge, J. (1999). Overview of assistive technology, *Center On Disabilities, Technology And Persons With Disabilities Conference 1999*. Available online at <http://www.csun.edu/cod/conf/1999/proceedings/session1017.htm> (2003-03-03).
- Shieber, S. M. and Baker, E. (2003). Abbreviated text input, *International conference on Intelligent user interfaces: 2003, Miami, Florida*, pp. 293–296. Available online at <http://www.eecs.harvard.edu/shieber/papers/abbrev-iii.pdf> (2003-08-14).
- Sparck Jones, K. and Galliers, J. R. (1995). *Evaluating Natural Language Processing Systems: An analysis and Review*, Springer-Verlag.
- Trost, J. (1997). *Kvalitativa intervjuer*, 2nd edn, Lund: Studentlitteratur.
- Trost, J. (2001). *Enkätboken*, 2nd edn, Lund: Studentlitteratur.
- Willis, T. (2001). Research proposal - data-intensive linguistics applied to word prediction for users with motor disabilities. Available online at http://www.cogsci.ed.ac.uk/~twillis/ALL_TOGETHER_6.doc (2003-03-18).

A User interface images



Figure 90: FASTY settings

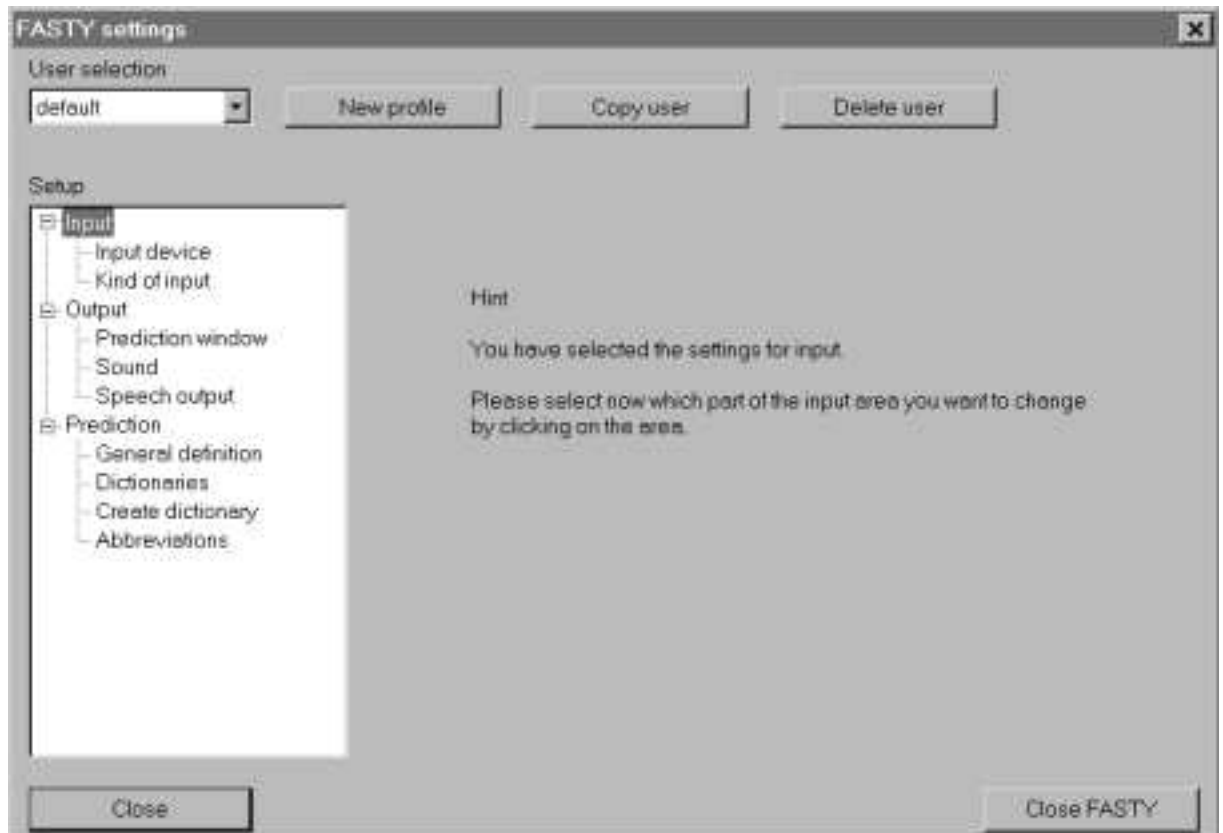


Figure 91: The main settings

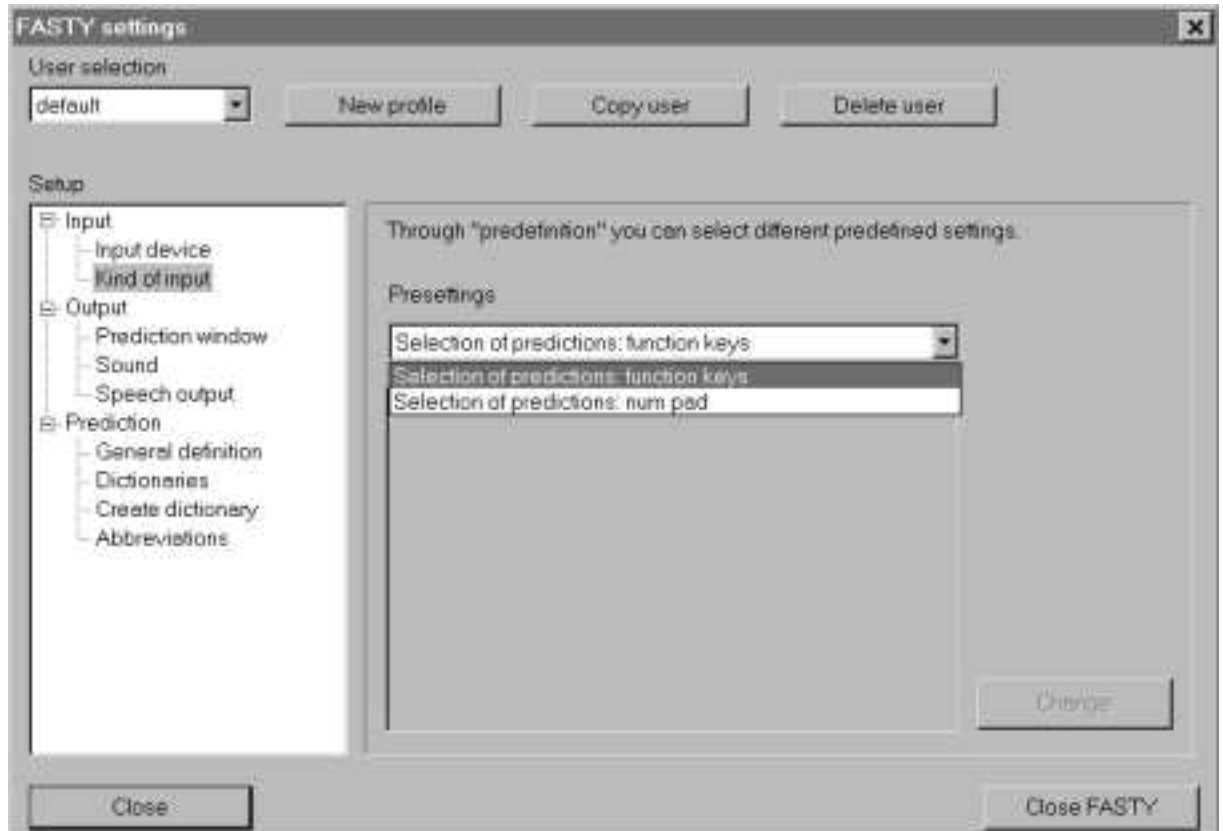


Figure 92: Kind of input

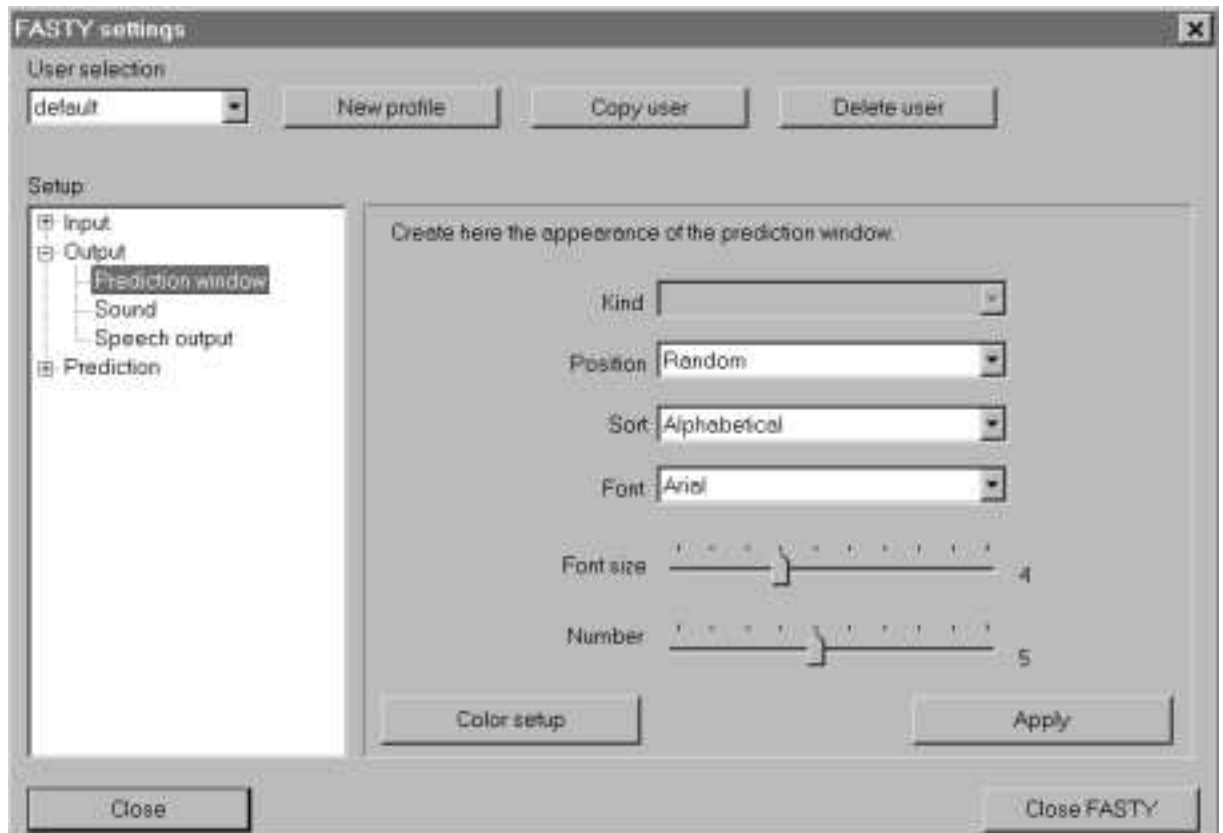


Figure 93: Settings for the prediction window

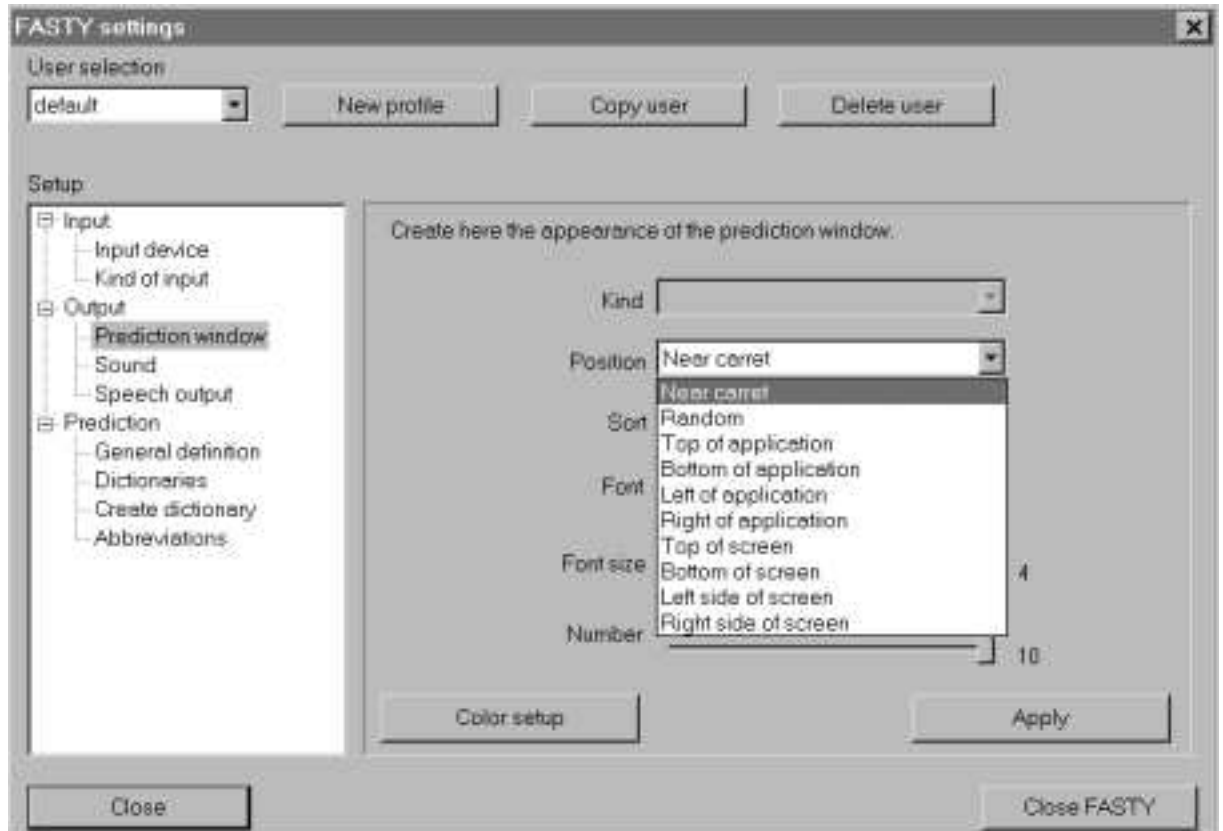


Figure 94: Position options of the prediction window

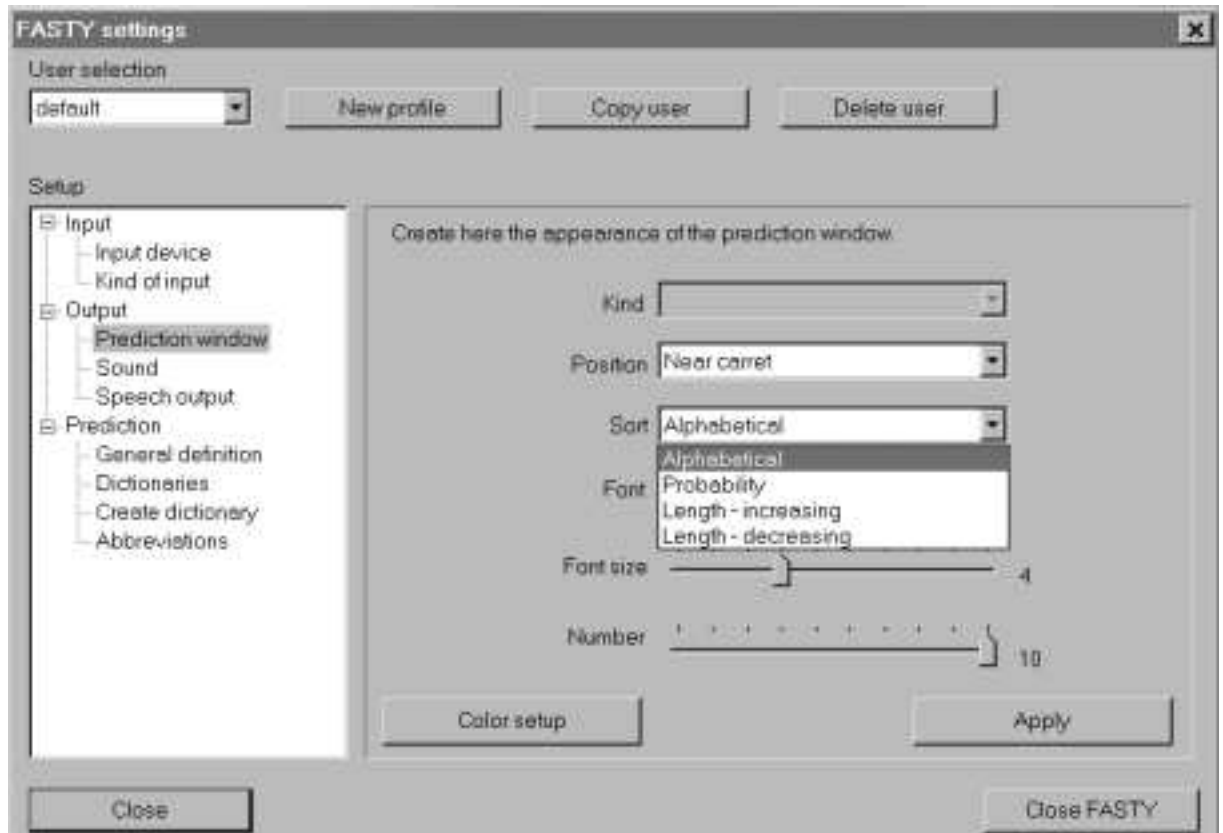


Figure 95: Sorting options of the prediction window

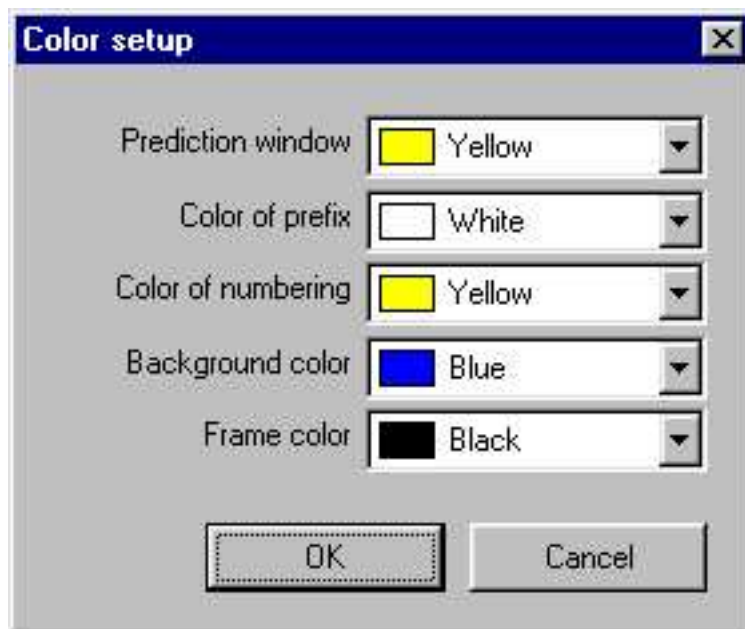


Figure 96: Colour options for the prediction window

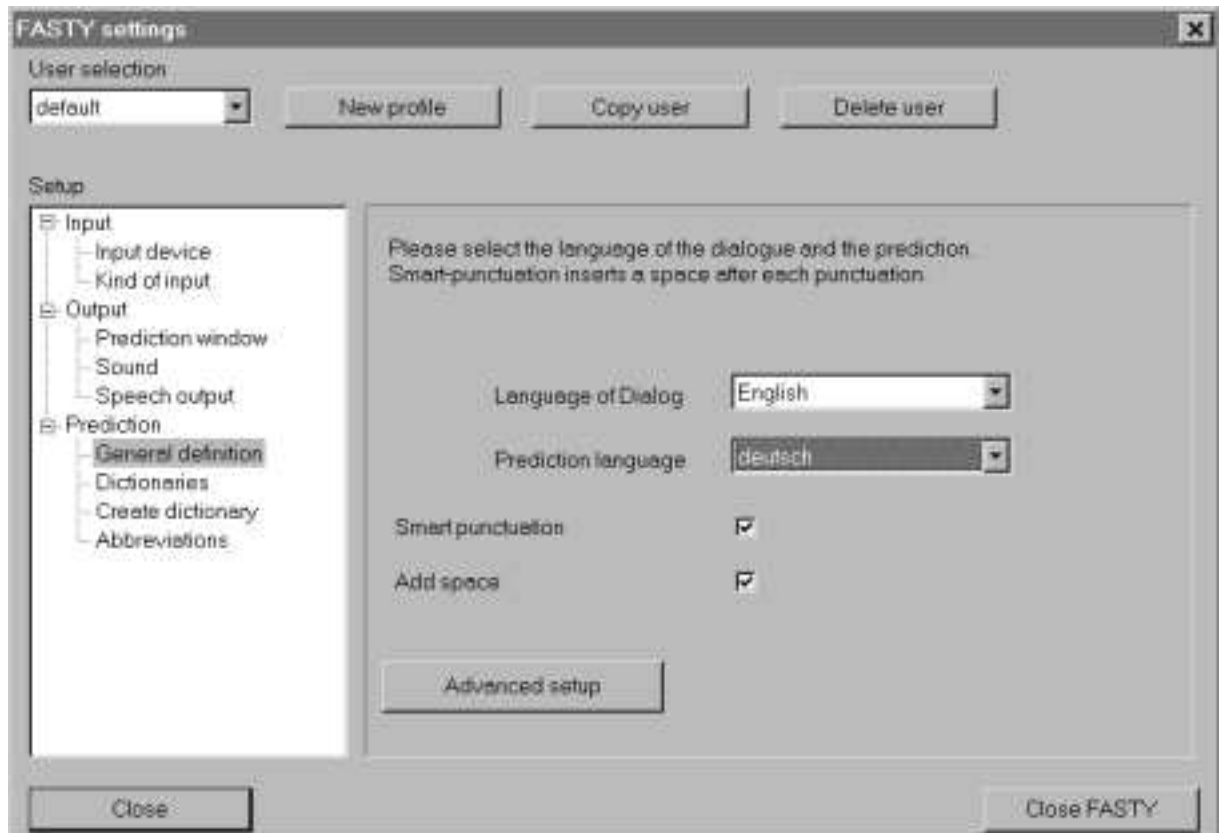


Figure 97: General definition

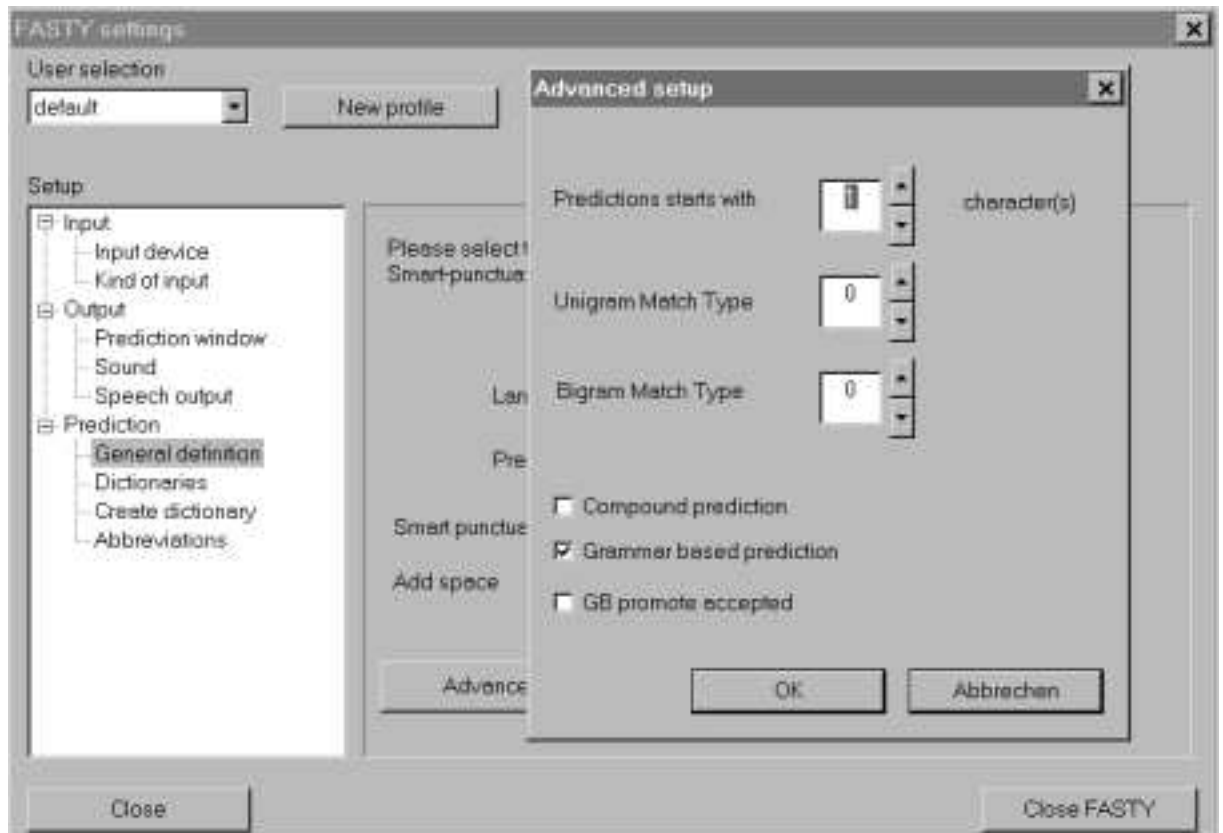


Figure 98: Advanced setup

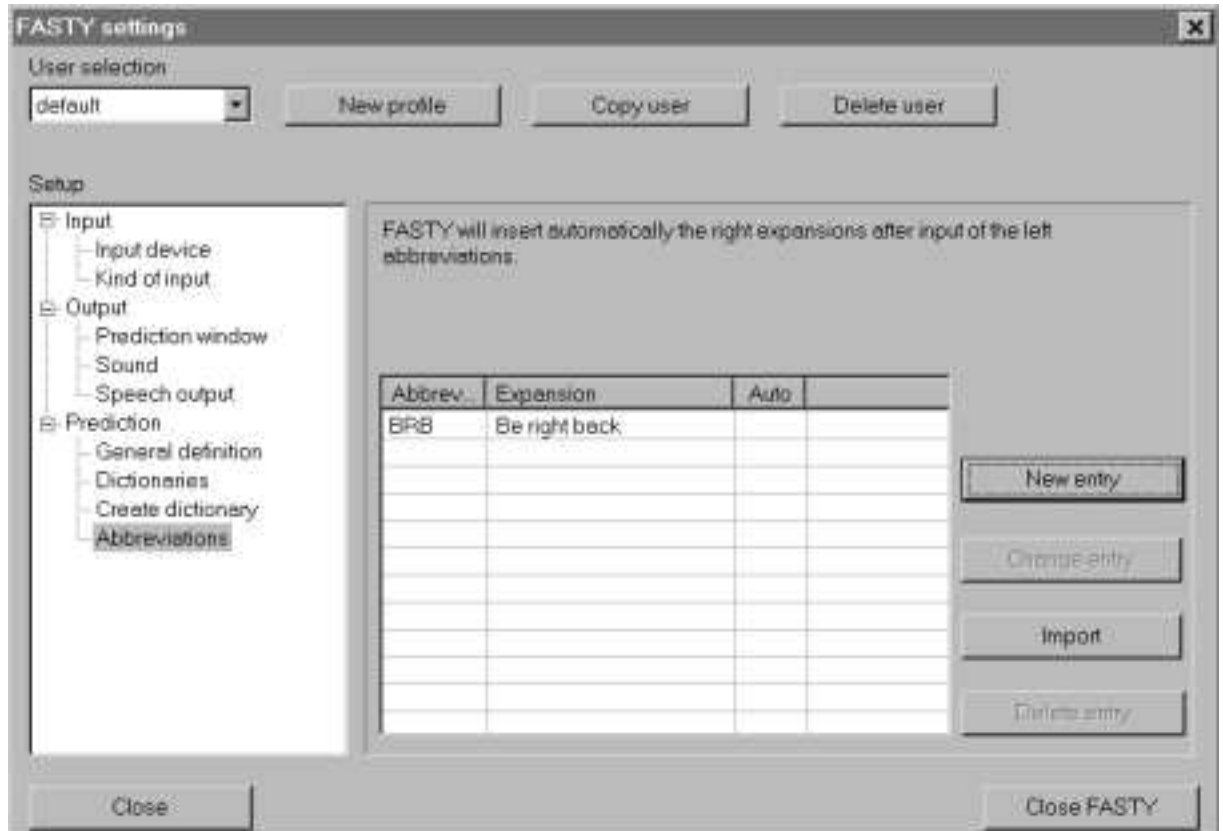


Figure 99: Abbreviation

B Questionnaire glossary

Here you will find explanations of some of the phrases used in each question in the questionnaire, in case any of them are unclear.

Easy to use = not requiring repeated reference to manuals, or stopping to think hard about how to perform actions, or making frequent mistakes.

Well integrated = form a sensible and intuitive whole, supporting a well-defined set of tasks.

Need the support of a technical person = Not be able to work the system and achieve one's goal relying solely on the one-line help and tutorial provided.

Confident = Sure that one could use it on others tasks and that one would know how to navigate through its facilities.

A lot of things = Many new facts about computers, new skills with keyboards, etc., new areas of application domains.

Most people = Any potential users of the system with similar experience and training to yourself.

Inconsistency = Similar or related aspects of the system using different techniques or vocabulary to achieve a result, or different parts of the system requiring the user to do the same task in different ways.

Cumbersome = Making actions or procedures unnecessarily awkward or difficult.

C Time table

Phase I (April)	1	Introduction				
		31	1	2	3	4
	2	Calibration + Report on Friday				
		7	8	9	10	11
	3	Report on Friday				
	14	15	16	17	18	
4	Draft Questionnaire + Report on Friday					
	21	22	23	24	25	
Phase II (April, May)	5	Report on Friday				
		28	29	30	1	2
	6	Report on Friday				
		5	6	7	8	9
	7	Report on Friday				
	12	13	14	15	16	
Phase III (May, June)	8	Calibration + Report on Friday				
		19	20	21	22	23
	9	Report on Friday				
		26	27	28	29	30
	10	Final Questionnaire + Report on Friday				
	2	3	4	5	6	

Table 8: Time table - validation period I

D Test report template

Behaviours

Please write something you noticed according the behaviour of your users.

Bugs

- Major
Please report encountered bugs which make the system crash.
- Minor
Please report cosmetic bugs which make the system non trivial.

Software which does not work with FASTY

Please write here the list of known software which does not work with FASTY or which have some strange behaviour with FASTY. (if possible, write a description of the problem)

Open questions

Please put here open questions from the end users.

E Keystroke saving rate and typing errors

Date	Week	Chars	Keys	Sel	Bsp	Errorkeys	err%	KSR	opt.KSR
04/08	1	37	31	5	6	2	15.50	2.70	8.11
04/08	1	81	68	12	14	12	5.66	1.23	16.05
04/08	1	26	2	6	0	0	0	69.23	69,23
04/08	1	289	144	39	13	6	24.00	36.68	38.75
04/08	1	204	123	34	31	16	7.69	23.04	30.88
04/09	1	486	554	72	161	78	7.10	-28.80	-12.75
04/10	1	423	287	29	1	3	95.67	25.30	26.00
04/11	1	259	145	42	25	13	11.15	27.80	32.82
04/12	1	324	258	35	51	33	7.82	9.57	19.75
04/13	1	159	162	27	54	18	9.00	-18.87	-7.54
04/14	1	368	387	9	19	19	20.37	-7.61	-2.45
04/14	1	699	569	61	47	0	0	9.87	9.87
04/14	1	841	823	13	12	12	68.58	0.59	2.02
04/14	1	280	164	31	17	6	27.33	30.36	32.50
04/14	1	247	132	26	19	4	33	36.03	37.65
04/15	1	295	177	40	27	17	10.41	26.44	32.20
04/15	1	114	115	20	47	25	4.60	-18.42	3.51
04/16	1	232	192	20	24	10	19.20	8.62	12.93
04/17	1	45	28	5	2	2	14.00	26.67	31.11
04/17	1	278	164	35	13	13	12.61	28.42	33.09
04/17	1	273	161	38	7	6	26.83	27.11	29.30
04/17	1	852	648	71	26	30	21.60	15.61	19.13
04/17	1	764	691	43	40	34	20.32	3.93	8.38
04/19	1	2052	1257	332	304	129	9.74	22.56	28.85
04/19	1	312	422	67	204	37	11.41	-56.73	-44.87
04/19	1	695	644	123	228	74	8.70	-10.36	0.29
04/19	1	49	53	3	8	8	6.23	-14.28	2.04
04/20	1	3150	2173	503	612	216	10.06	15.05	21.90
04/21	2	890	900	21	36	32	28.13	-3.48	0.11
04/21	2	1444	1323	105	117	99	13.36	1.11	7.96
04/22	2	1071	895	96	73	61	14.67	7.47	13.17
04/22	2	741	789	2	28	28	28.18	-6.75	-2.97

Date	Week	Chars	Keys	Sel	Bsp	Errorkeys	err%	KSR	opt.KSR
04/22	2	961	962	24	51	55	17.49	-2.60	3.12
04/22	2	1345	721	186	77	66	10.92	32.56	37.47
04/22	2	2423	1755	314	297	145	12.10	14.61	20.59
04/23	2	1071	1184	239	567	104	11.38	-32.87	-23.16
04/24	2	376	203	66	39	23	8.82	28.46	34.57
04/24	2	331	412	20	102	71	5.80	-30.51	-9.06
04/24	2	727	559	61	35	30	18.63	14.72	18.84
04/24	2	792	768	32	51	48	16.00	-1.01	5.05
04/25	2	2804	2328	190	124	111	20.97	10.20	14.16
04/26	2	775	618	71	42	38	16.26	11.10	16.00
04/26	2	704	426	93	34	31	13.74	26.28	30.68
04/26	2	47	66	2	12	12	5.50	-44.68	-19.15
04/27	2	464	468	29	71	55	8.51	-7.11	4.74
04/27	2	605	474	55	28	23	20.61	12.56	16.36
04/28	3	216	183	12	10	9	20.33	9.72	13.89
04/28	3	627	579	29	33	28	20.68	3.03	7.50
04/28	3	150	119	10	6	6	19.83	14.00	18.00
04/29	3	212	91	41	21	14	6.50	37.74	44.34
04/29	3	426	308	36	28	17	18.12	19.25	23.24
05/01	3	425	278	53	30	19	14.63	22.12	26.59
05/01	3	921	773	75	65	55	14.05	7.93	13.90
05/01	3	604	495	49	31	25	19.80	9.93	14.07
05/02	3	544	446	41	25	17	26.24	10.48	13.60
05/02	3	428	381	26	28	26	14.65	4.91	10.98
05/02	3	62	60	1	1	1	60.00	1.61	3.23
05/02	3	324	246	27	7	6	41.00	15.74	17.59
05/02	3	337	303	17	19	19	15.95	5.04	10.68
05/02	3	861	813	31	44	41	19.83	1.97	6.74
05/03	3	675	485	83	77	26	18.65	15.85	19.70
05/03	3	560	544	68	102	37	14.70	-9.28	-2.68
05/03	3	179	260	16	71	47	5.53	-54.19	-27.93
05/04	3	854	648	68	35	28	23.14	16.16	19.44
05/04	3	651	632	20	34	34	18.59	-0.15	5.07
05/04	3	712	629	38	37	33	19.06	6.32	10.96
05/05	4	260	176	38	27	20	8.80	17.69	25.38
05/06	4	259	143	44	19	5	28.60	27.80	29.73
05/06	4	626	584	23	21	22	26.55	3.04	6.55

Date	Week	Chars	Keys	Sel	Bsp	Errorkeys	err%	KSR	opt.KSR
05/06	4	379	306	25	24	15	20.40	12.66	16.62
05/07	4	305	291	5	8	8	36.38	2.95	5.57
05/07	4	1072	837	87	57	50	16.74	13.81	18.47
05/08	4	120	118	6	10	10	11.80	-3.33	5.00
05/13	5	588	292	85	66	34	8.59	35.88	41.67
05/13	5	581	493	45	56	25	19.72	7.40	11.70
05/14	5	4178	2265	617	211	111	20.41	31.02	33.68
05/14	5	618	282	105	23	14	20.14	37.38	39.64
05/14	5	293	146	40	10	6	24.33	36.52	38.57
05/14	5	1521	1150	121	52	43	26.74	16.44	19.26
05/14	5	788	637	83	51	40	15.93	8.63	13.70
05/15	5	223	207	29	19	15	13.80	-5.83	0.90
05/15	5	276	168	31	29	2	84.00	27.89	28.62
05/15	5	164	134	15	21	10	13.40	9.15	15.24
05/16	5	1405	944	167	58	39	24.21	20.93	23.70
05/16	5	1051	785	105	64	57	13.77	15.32	20.74
05/16	5	677	544	60	52	33	16.48	10.78	15.66
05/16	5	2951	2982	82	180	177	16.85	-3.83	2.17
05/17	5	279	139	41	24	16	8.69	35.48	41.22
05/19	6	1182	954	96	48	44	21.68	11.17	14.89
05/19	6	350	277	37	12	7	39.57	10.29	12.29
05/19	6	697	420	93	30	23	18.26	26.40	29.70
05/20	6	690	457	86	61	57	8.02	21.30	29.57
05/20	6	1053	736	128	61	49	15.02	17.95	22.60
05/21	6	579	345	93	35	24	14.38	24.35	28.50
05/21	6	2121	2197	80	168	157	13.99	-7.36	0.05
05/21	6	796	561	95	36	21	26.71	17.59	20.23
05/21	6	1443	873	194	51	40	21.83	26.06	28.83
05/21	6	897	686	91	53	47	14.60	13.38	18.62
05/22	6	176	121	16	14	14	8.64	22.16	30.11
05/22	6	308	241	30	36	21	11.48	12.01	18.83
05/22	6	306	172	47	10	7	24.57	28.43	30.72
05/23	6	122	123	7	11	10	12.30	-6.56	1.64
05/23	6	691	459	80	36	28	16.39	22.00	26.05
05/23	6	334	243	36	17	15	16.20	16.47	20.96
05/23	6	473	284	64	21	18	15.78	26.43	30.23
05/24	6	25	16	3	0	0	0	24.00	24.00

Date	Week	Chars	Keys	Sel	Bsp	Errorkeys	err%	KSR	opt.KSR
05/24	6	84	66	9	4	4	16.50	10.71	15.48
05/24	6	706	487	83	20	18	27.06	19.26	21.81
05/24	6	293	138	51	10	10	13.80	35.49	38.91
05/25	6	408	300	40	13	13	23.08	16.67	19.85
05/25	6	324	308	44	67	33	9.33	-8.64	1.54
05/26	7	1428	1235	114	78	69	17.90	5.53	10.36
05/26	7	185	87	31	8	5	17.40	36.22	38.92
05/27	7	1198	992	98	44	41	24.19	9.02	12.44
05/27	7	212	133	28	11	10	13.30	24.06	28.77
05/27	7	945	746	80	34	32	23.31	12.59	15.98
05/29	7	176	100	21	9	6	16.67	31.25	34.66
05/29	7	269	195	36	37	24	8.13	14.13	23.05
05/30	7	343	257	35	45	20	12.85	14.87	20.70
05/30	7	137	117	14	25	17	6.88	4.38	16.79
05/30	7	856	631	84	33	25	25.24	16.47	19.39
06/01	7	1103	735	144	52	46	15.98	20.31	24.48
06/01	7	107	81	10	3	3	27.00	14.95	17.76
06/02	8	1047	950	70	44	36	26.39	2.58	6.02
06/02	8	1005	652	133	42	38	17.16	21.89	25.67
06/02	8	270	320	1	26	26	12.31	-18.88	-9.26
06/03	8	420	256	76	56	22	11.64	20.95	26.19
06/03	8	185	109	25	6	5	21.80	27.57	30.27
06/04	8	360	251	44	20	18	13.94	18.06	23.06
06/04	8	565	442	63	62	38	11.63	10.62	17.35
06/05	8	551	387	61	31	26	14.88	18.69	23.41
06/05	8	672	468	66	24	22	21.27	20.54	23.81
06/06	8	546	414	53	30	28	14.78	14.47	19.60
06/06	8	701	502	61	19	20	25.10	19.69	22.54
06/07	8	455	207	85	36	20	10.35	35.82	40.22
06/08	8	594	503	41	20	20	25.15	8.42	11.78
06/09	8	646	458	72	35	31	14.77	17.96	22.76
06/09	8	258	206	24	18	17	12.12	10.85	17.44
06/09	8	900	626	96	28	30	20.87	19.77	23.11
06/09	8	258	216	48	58	39	5.54	-2.33	12.79
06/13	8	205	72	38	9	1	72.00	46.34	46.83

Table 9: Keystroke saving rate and typing errors