Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Anna Kantosalo

Usability Evaluation with Children
Case: Poetry Machine

Master's Thesis
Espoo, November 14, 2014

Supervisors: Professor Marko Nieminen, Aalto University
Advisor: Sirpa Riihiaho Lic. Sc. (Tech.), Aalto University
Mari Tyllinen M.Sc. (Tech), Aalto University

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author: Anna Kantosalo

Title:
Usability Evaluation with Children Case: Poetry Machine

Date: November 14, 2014 **Pages:** 115

Major: Usability and User Interfaces **Code:** T-121

Supervisors: Professor Marko Nieminen

Advisor: Sirpa Riihiaho Lic. Sc. (Tech.)
Mari Tyllinen M.Sc. (Tech)

This thesis presents a case study of a usability evaluation of a creative poetry writing tool conducted with Peer Tutoring and Group Testing methods in a Finnish 3rd year class. The Peer Tutoring method has previously been used for usability testing with children, but the Group Testing method is new. Additionally a new, experimental feedback gathering method, called the Feedback Game was designed to be used as a part of the Group Testing. The goals of the evaluation were to produce a list of usability errors, evaluate the enjoyability and usefulness of the concept as well as produce descriptive feedback and improvement ideas for the program concept. This thesis focuses on evaluating the testing methods for their performance against these goals, as well as their reliability, coverage, quality of analysis, and suitability for testing with children.

The Peer Tutoring method was found to work well for practical usability evaluation with children. This approach, involving pairs of children, instead of individual children was seen beneficial in producing enjoyable test conditions, but the tasks used were too abstract for the participants. The Group Testing was only useful for validating the most severe problems found with Peer Tutoring. However the Feedback Game used with the Group Testing worked as well as the post task interview of the Peer Tutoring method for eliciting new ideas, although its results were similar to those found with Peer Tutoring. As a benefit the Feedback Game needs less resources compared to the Peer Tutoring method. Additional research about the effectiveness of the Peer Tutoring method with different audiences and tasks are needed. The usefulness of the Group Testing method should be re-evaluated in more thoroughly recorded conditions. The Feedback Game needs more research in order to prompt more discussion between peers during the game.

Keywords: usability, user-centered design, usability testing, children

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan koulutusohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Anna Kantosalo		
Työn nimi:	Käytettävyystestaus lasten kanssa Tapaus Runokone		
Päiväys:	14. marraskuuta 2014	**Sivumäärä:**	115
Pääaine:	Käyttöliittymät ja käytettävyys	**Koodi:**	T-121
Valvojat:	Professori Marko Nieminen		
Ohjaaja:	Tekniikan lisensiaatti Sirpa Riihiaho Diplomi-insinööri Mari Tyllinen		

Tässä diplomityössä sovellettiin vertaistutorointimetodia ja ryhmätestimetodia lapsille suunnatun luovan runonkirjoitustyökalun prototyypin käytettävyyden arviointiin. Käytettävyystestaukseen osallistui erään etelä-suomalaisen koulun kolmannen luokan oppilaita. Vertaistutorointimetodia on aiemmin sovellettu lasten kanssa tehtävään käytettävyystestaukseen, mutta käytetty ryhmätestimetodi on uusi. Ryhmätestiä varten suunniteltiin uusi kokeellinen palautteenkeräysmetodi, palautepeli. Testauksen tavoitteena oli tuottaa kattava lista käytettävyysongelmista, arvioida konseptin hauskuutta ja hyödyllisyyttä, sekä tuottaa deskriptiivistä palautetta ja kehitysideoita ohjelmakonseptin parantamiseksi. Työssä keskitytään metodien arviointiin näiden tavoitteiden saavutuksen suhteen, lisäksi arvioidaan niiden luotettavuutta, kattavuutta, analyysin laatua, ja soveltuvuutta lapsille.

Vertaistutorointimetodin todettiin toimivan hyvin käytännön arviointityössä, ja testin pariluonteesta havaittiin merkittävää etua miellyttävän testikokemuksen aikaansaamisessa, mutta testitehtävä osoittautui liian abstraktiksi osallistujille. Ryhmätestauksen tulokset olivat hyödyllisiä ainoastaan vakavimpien käytettävyyslöydösten validoinnissa. Käytetty palautepeli toimi kuitenkin vertaistutorointiin nähden yhtä tehokkaasti uusien ideoiden keräämiseen, vaikka sen kautta kerätyt ideat olivatkin samankaltaisia, kuin vertaistutoroinnissa kerätyt. Palautepeliin tarvitut resurssit, kuten analyysiaika, olivat kuitenkin pienemmät. Lisätutkimusta vertaistutorointimetodin tehokkuudesta erilaisten testitehtävien ja osallistujien kanssa tarvitaan. Ryhmätestin hyödyllisyyttä tulisi arvioida uudelleen tehokkaamman datankeräyksen kanssa. Entistä rikkaamman lastenvälisten keskustelun kehittämistä palautepelin avulla on tutkittava lisää.

Asiasanat:	käytettävyys, käyttäjäkeskeinen suunnittelu, käytettävyystestaus, lapset
Kieli:	Englanti

Acknowledgements

Many have provided me their insights, criticisms, as well as practical help in making this thesis possible. I wish to thank all of you for your support during the process!

I would like to thank especially both of my instructors, Sirpa Riihiaho and Mari Tyllinen, who have helped me throughout the process, providing criticism and practical advice first on the research plan outlining the experiment, and later throughout the analysis phase, finally culminating on this thesis itself. I also wish to thank Professor Marko Nieminen, who took the time to personally comment on the first drafts of the actual thesis.

I also wish to thank Karoliina Tiuraniemi and Mikko Hynninen, who were recruited with the kind help of Sirpa Riihaho, to act as observers and evaluators in the test sessions. And naturally I wish to thank all pupils and teachers, who participated in the various test sessions, as well as those teachers, and personnel who made the arrangement of the tests at a school possible.

This thesis has also been supported by the Helsinki Institue for Information Technology, HIIT and the University of Helsinki, where I have worked throughout the thesis writing process. The prototype used in the tests was developed by me and my colleague Jukka Toivanen for the Discovery Research Group lead by Professor Hannu Toivonen at the University of Helsinki. I am grateful for all my co-workers for pitching their ideas for the prototype, and commenting on the progress of the project. Laura Langohr deserves special thanks for her supportive feedback after reading the first draft through!

I also wish to thank my friends and family for supporting the thesis writing process and helping me to focus on actual tasks. And of course Jesse, for tirelessly answering my questions on English prepositions!

Espoo, November 14, 2014

Anna Kantosalo

Contents

Chapter 1

Introduction

Children are an interesting and challenging user group. Their physical and cognitive abilities are very different from adults [13], which puts the adult designer into a very difficult position. Therefore it is vital that the usability of children's products is effectively tested with appropriate methods with the target audience [9].

Children's software can be classified into three genres: enabling software, entertainment software and educational software [35]. The study of usability testing with children in the 1990's was more focused on the entertainment perspective, which can be seen in one of the most cited articles of the time, "Guidelines for Usability Testing with Children" by Hanna, Risden and Alexander. Lately we have seen an increase in discussing the evaluation of educational software for children. This thesis takes part in this discussion by examining the usability of one educational creative writing tool aimed at children.

The usability of educational software today is even more vital than before: Computers are now recognised by an increasing number of educators and laymen as a channel, which can reach modern pupils. The benefit of computers is their flexibility; they can offer a variety of experiences and learning opportunities [13]. Yet this flexibility must be channelled appropriately in order to produce meaningful and engaging resources for learning. Usability has been described even as the key issue of the success of e-learning applications by Lahti et al. [20]. It is considered as "vital" for edutainment applications by Egloff [6], and it is also an important factor for evaluating the capability of educational software to facilitate the acquisition of knowledge [24]. Therefore usability specialists should work hand in hand with educational specialists to produce meaningful learning experiences for children.

The benefits of testing with children are many: In addition to traditional benefits of usability testing, such as reducing cost and improving quality [13],

it will provide unique insight into the different perspectives children bring into computer products [9]. Testing with children can help professionals settle design debates and refine designs so that they become accessible for all ages [9]. Patel and Paulsen [29] even consider that not only can young children provide as much information and potential problems as adult expert reviewers do, but at times they can even contribute new information that could not be obtained through expert review.

Naturally, children's products should be tested by children [9]. This could be held as the baseline requirement for any user centered design effort in the design of children's products [25]. However, children's products can not be evaluated in the same way with software designed for adults [6]. Children often use computers in unstructured ways and outside the traditional office environment [13]. During testing they are not merely asked to use a software, but to adapt to the test environment, interact with the adult facilitators, follow processes and contribute to the evaluation by reporting their experiences [25]. This requires adjusting the test environment, tasks and procedures for the child user. Therefore, finding a selection of methods, best suited for testing a particular application with children is a real issue.

This thesis examines this issue through applying usability testing to a prototype of an educational poetry writing tool. The tool is based on computational creativity methods, and as such will have an artificial intelligence component, which will help the user in writing new poetry. The tool is designed to be used in school contexts by pupils in the age group of 9-12. The usability testing is a necessary part needed to develop the prototype into a program, which can be used to examine the co-creative work between humans and creative machines. As this is the first time this type of a co-writing tool is tested with users, it is also important to gather knowledge and ideas on improving the concept itself to establish new research goals in computational creativity.

Interestingly, the goals of the user testing are also related to current issues in usability testing with children. Research has requested more context aware testing [13] and including children also in different stages of development, especially in the early phases of conceptual design and evaluation [28]. To involve children more in the further development of such tools, this thesis suggests a new fast method for combining concept feedback discussion with context related testing at school. The new Feedback Game method is presented as a tool for guiding concept discussions within groups of children, but more rigorous usability testing, such as the Peer Tutoring method used in this thesis is recommended for collecting traditional usability error results.

1.1 Research Questions, Goal and Scope

This thesis examines the usability of a creative writing tool intended for children in school contexts. The work began by investigating the following research problem: *How should user-centered design methods be applied to testing an educational creative writing tool designed for children?*

During the background research phase, this research problem was divided into more approachable research questions: *How are usability tests with children conducted in practise? What aspects of usability need to be tested with children? What methods need to be chosen to collect information on the selected aspects?*

Later on into the study, topics for the actual practical usability evaluation were selected. These are *to find usability problems within the prototype*, *to evaluate the usefulness of the prototype* by identifying the most useful and utilized features of the prototype, *to evaluate the enjoyability of the prototype* through observed and reported indices of discomfort and fun, and *to elicit feedback* for the further development of the product concept.

It is important to note that the field of children's usability is diverse. Hourcade [13] reports five approaches, computing, education, psychology, art and design, and engineering, to studying children's usability. In this thesis I have focused on the usability engineering perspective and no exhaustive review is made on educational and psychological aspects. This thesis also focuses on testing approaches which involve users. Expert reviews of children's software are purposefully left out of this thesis.

It is possible to measure how well an application facilitates learning to some extent through pre- and post-test set-ups (see for example Sim et al. [35]), or with a test (see for example Costabile et al. [2]). But in general learning is very challenging to measure, as it's definitions change and it is hard to observe [7]. However the focus of this thesis is in measuring aspects of usability and therefore learning assessment methods are excluded from this study. Also no formal attempt is made to reviewing the tools capabilities to supporting creative work.

The field of e-learning itself can be divided into separate domains: adult education, higher education, and traditional school environment [20]. Because of the lack of resources focusing on the evaluation of educational software intended for traditional school environment, the literature review mostly focuses on children's products in general.

1.2 Research Methods and Data

A literature review was done to select suitable goals and methods for the user testing. Peer Tutoring was selected as the main method for usability testing based on the literature review. In addition to Peer Tutoring, I designed a Group Testing session, which included a group based feedback gathering method, called the Feedback Game. Six peer tutoring sessions were held over two days and two group testing sessions over one day. The tests were held within a two week period in the spring of 2014. All testing was conducted with one 3rd year class in a Finnish primary school located in southern Finland.

1.3 Context of This Research

I have acted as a user-centered design specialist throughout the Poetry Machine project at the Discovery Research Group in the University of Helsinki. My position has been funded by the Helsinki Institute for Information Technology, HIIT, and Finnish Centre of Excellence for Algorithmic Data Analysis Research, Algodan. University of Helsinki paid for the materials used in this work, except for the camera, which was loaned from Aalto University.

The idea for the Poetry Machine concept was invented together with doctoral researcher Jukka Toivanen and Professor Hannu Toivonen. I did the initial fieldwork for the concept, gathering information on the target user group and context, and continued with a literature research. Based on the information gathered I designed the user-interface for the concept and build the prototype based on the concept designs after briefly discussing them with the rest of the team and an outside specialist of educational software. Jukka helped me with connecting the prototype to his creative computing algorithms and 3rd party language processing resources.

My thesis instructors, Sirpa Riihiaho and Mari Tyllinen, guided me in designing the tests described in this thesis. The method selection and development process was iterative: Sirpa and Mari commented on my literature research and gave additional pointers, until I made a test plan. At this point it became clear that testing would require additional observers. Sirpa helped me to recruit two other students, Karoliina Tiuraniemi and Mikko Hynninen. Karoliina kindly offered to pilot test the background questionnaire with her son. Karoliina and Mikko acted as observers during the tests, and participated in the analysis. I wrote a poster [18] about the Feedback Game reported in this thesis with Sirpa for the NordiCHI 2014 conference. Other team members were unable to participate due to their tight schedules.

Chapter 2

The Poetry Machine

The Poetry Machine (PM) is a concept for a creative poetry writing tool developed at the Discovery Research Group at the Department of Computer Science of the University of Helsinki. The PM project was launched in 2013 as a result of an ideation session on practical uses for the computational linguistic creativity algorithms previously developed at the Discovery group. In the session it was decided that the adaptation work would focus on building a creative poetry writing tool intended for use in the Finnish primary school. Inspired by the STANDUP project ([33], [40]), which used User-Centered Design to put computational joke generation algorithms to work for children with complex communication needs, a User-Centered Design approach was adopted for the project.

The purpose of the PM is to give a way for children to explore the world of poetry through writing poetry of their own. The PM tries to help users by getting them over the empty paper stage, and giving them more material to work with by request. The concept does not try to formally educate its users on poetry, instead it is build to be more of a writing platform to allow for free expression. The web based tool is suitable to be used for fun at home, or teachers can build serious learning sessions around it.

Currently the PM tool exists as a prototype which can be run on a Linux computer, using a local python powered Django server, source files and 3rd party libraries. When the server is run, any major browser can be used to access the program. A copy of the source code can be requested from the author of this thesis. The current edition allows the user to select a pre-defined topic, proceed to editing a computer provided sample poem, ask for more machine input, and then move to a final stage, where the poem can be copied from the browser onto any program on the computer.

The PM concept is powered by the computational creativity algorithms intended for poetry generation developed at the University of Helsinki (see

[37] and [38]). From early on, one of the scientific goals for the PM project has been to utilize these algorithms. Therefore the possible concept designs for the PM are ultimately restricted by the underlying technology. Designing any solutions requires an understanding of the strengths and weaknesses of the algorithms. Therefore it was considered impossible to involve stakeholders as co-designers from early on and an approach to use them as informants was adopted instead. Once an initial design for the interface had been planned, I considered user-evaluation on a low-fidelity prototype. However I soon realised it would be impossible for a researcher to simulate the algorithms in real time, as it would require a lot of creativity. As Höysniemi et al. [15] write about Wizard of Oz testing methods: "The wizard's cognitive and motor skills restrict the interaction pace and the level of complexity of the system under analysis." The development process and these considerations are presented with more detail in [19]. After the initial designs were evaluated with an IT pedagogy expert, a functional prototype was designed for user evaluations. The rest of this chapter focuses on the details of the prototype tested in the usability tests described in this thesis.

2.1 The Poetry Machine Prototype

The tested prototype version includes the interface for composing poems. Login functions, and the intended on-line poem repository were left outside the scope of this test. The prototype consists of three screens, which are presented with screen-shots in images 2.1, 2.2, and 2.3. The prototype is fully interactive and semi-functional - all basic actions required to compose poems are implemented, but two features, the free input for poem topic, and dividing the poem into stanzas, were omitted to finish the prototype in time for testing. For the same reasons we had to replace some of the background functionalities with more simple ones to overcome some temporary problems with background libraries. However the appearances of the prototype were intended as final. Thus the tested prototype can be described as an interactive, functional, mixed-fidelity prototype, or a pilot system, using the terminology presented in [22].

The first screen, presented in image 2.1 is also called the starting screen. It has a dropdown menu for selecting a suitable topic for one's poem. It also features a large blue button styled to resemble an old feather pen, which is used to start the system. The upper corner has a shor text link, which in the final version would lead to a section giving related information on the PM concept and its use. This link was not functional in the prototype.

The second screen, presented in image 2.2 is also called the write mode.

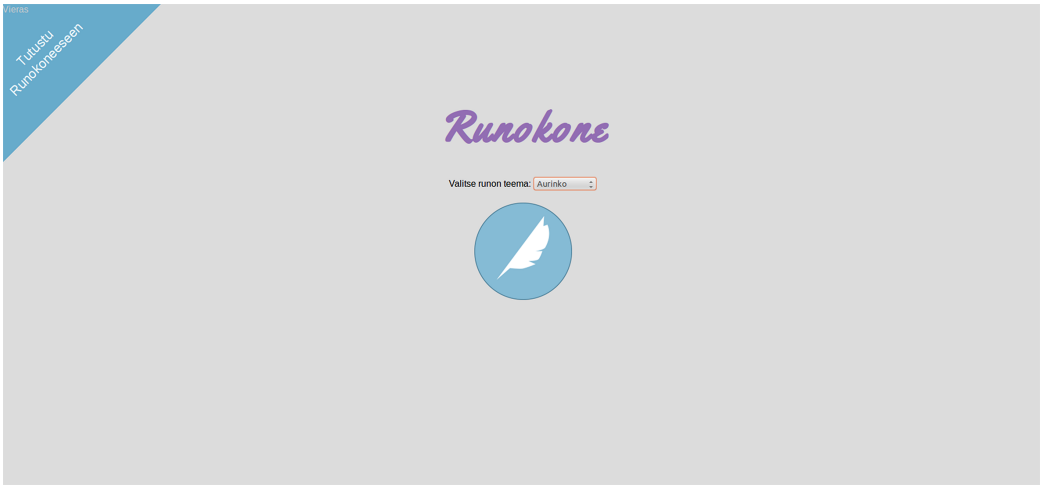

Figure 2.1: A screenshot of the starting screen of the prototype.

It is the main screen of the application, and the majority of the poem composition takes place in it. Once the user presses the feather pen-button on the first screen, the computer composes a short sample for on the user selected topic and presents it on this screen. From there on, the user can start modifying the poem. Within this window, the user can move words of the poem around by dragging them and dropping them on small indicator circles between words, or on top of the words themselves. The same can be done with rows. By double-clicking a word, the user can modify the word by typing on their keyboard. By double-clicking an indicator circle, the user can add a whole new word on it's place. The top row has four buttons - the robot-button, the rhyme-button, the trash-can-button and the finish-button. The user can remove words or rows by dragging them on the trash-can. If the user drags a word on the rhyme-button, the computer will find rhyming words for the target words. By dragging a word or a row on the robot-button, the computer will find related words or rows for the user to use. These new words and rows are presented in a speech-bubble box under the button row. Words or rows from the robot- or the rhyme-tool can also be dragged into the poem. Similarly, the elements dragged into the trash-can can be seen in the speech-bubble box and dragged back into the poem. If the user clicks on one of the robot, the rhyme-tool, or the trash-can, a short help text is displayed in the speech-bubble box instead. The user can move forward to the final screen by clicking the finish-button.

The final screen, presented in image 2.3 is also called the read mode. In this screen the poem is presented as text that can be copied to be included in

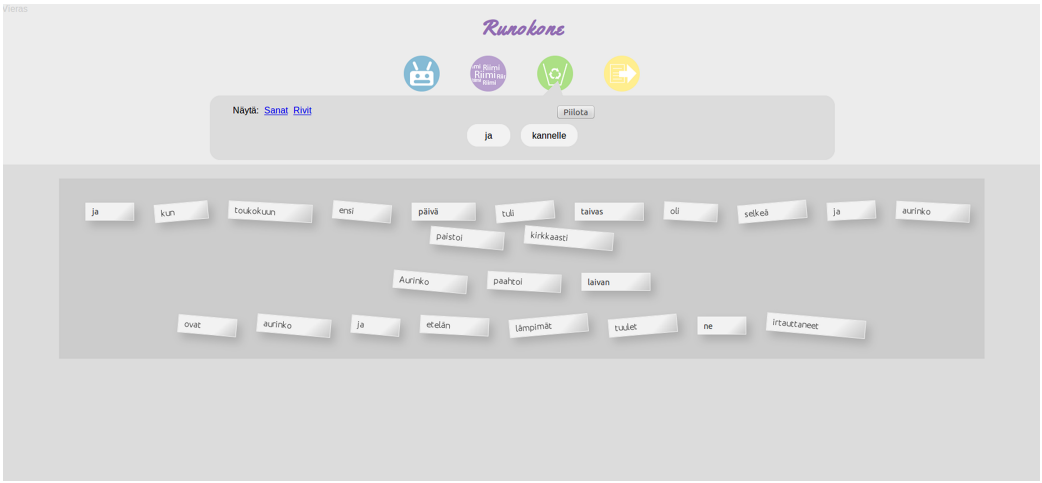

Figure 2.2: A screenshot of the prototype in the write mode, with the robot-tool open.

e-mails or other third party applications. It also allows the user to name the poem, enter a pen name, or change the theme of the poem. These changes are saved by pressing the button ”Päivitä”. The poem can also be published to be read on-line by clicking ”Julkaise runo”. The use can return to the write mode by clicking the feather-pen-button, which is similar to the one on the first screen. If the user wants to return to the starting screen, they can do it by clicking the link on the left most corner.

In this test version, the robot and the rhyme-tool will give similar rhyming answers for words, and the poems are composed by combining sentences with similar words.

2.2 Target User Group

The PM concept is targeted especially to the primary school level. It was originally developed based on user studies with one second and two ninth year classes. Basic reading and writing skills are needed for using the tool. The required level is usually reached by children during the second year of school. The national curriculum Finnish language has two poetry related courses, one on the third class and one on the ninth class. Currently the topic matter of the PM has been designed with the younger target group in mind and therefore the prototype is also designed to suit best the age range of 8-12-year-olds.

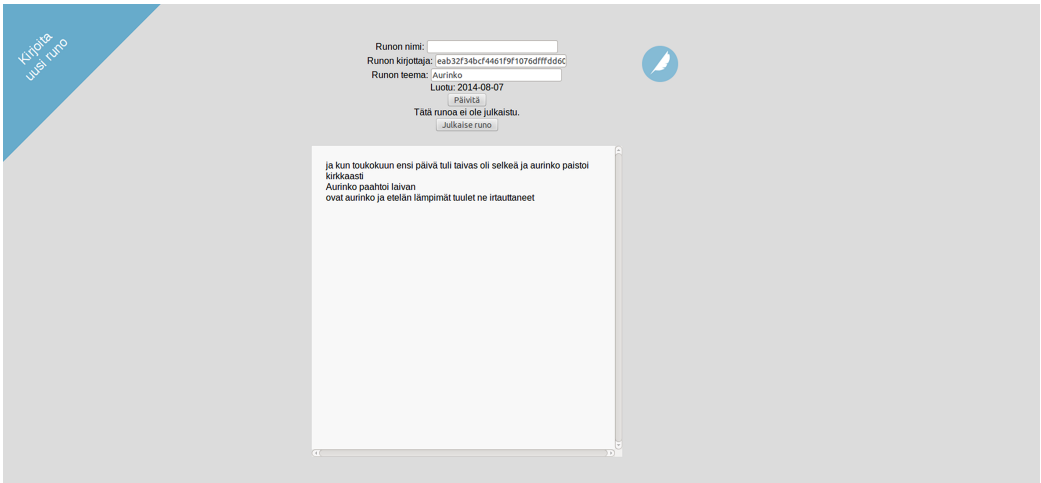

Figure 2.3: A screenshot of the last screen of the prototype.

2.3 A Typical Use Case

Leena goes to the third grade. She is a typical nine-year-old, who does averagely in school. Her favourite subject is biology, because she likes animals. Last year she wrote a presentation on dogs on the computer for her class. Leena is not particularly interested in writing stories, but it is ok when she has to. Her class has access to laptops during smaller group work times. Today she learns about poetry, one of the subjects of third grade Finnish education.

Her teacher Seppo begins the class by asking the children, what they know about poetry. Some kids tell a couple of examples, but Leena can not think of any. Then Seppo opens his computer and goes to the home address of the Poetry Machine. He tells, how today everyone is going to write a poem of their own. He starts by showing them an example: He selects a topic, and gets a small poem, which he reads aloud for the whole class. Then he asks, what the class thinks about it. He suggests they make a small change here and there and shows the children how it can be done. Finally he goes clicks the finish button and shows the whole poem to the class. Now it is time for them to try.

Leena goes to the Poetry Machine site and selects a topic. She selects the same topic as her teacher, but notices that the result she gets is entirely different. For a start, it is much shorter. She asks the teacher if she can get a longer poem somehow. Seppo shows the whole class, how the robot-tool works. Leena uses it to get more lines into her poem. She selects a few about

puppies. She decides to work the theme throughout the poem and writes the puppy into the existing sentences. But now she can not think of any good rhymes for the word "puppy". And she does not even know what rhymes really are. She selects the rhyme tool and drops the word "puppy" into it. She soon notices she is receiving more words with the same ending. Some of them are actually very funny and she chuckles to herself when she selects the word "tummy". After a while her poem is finished and she moves on to the last page. She copies her poem and sends it to her mum and dad by e-mail.

Once at home, she decides to take another look at the tool. This time she writes about cats.

2.4 Scope of the Prototype

The scope and fidelity degree of this prototype should allow the evaluation of the usability of the tool very well: According to Sim et al. previous studies have been inconclusive on the effects of prototype fidelity on the performance of adult user [34]. They note that sometimes adult users seem to prefer high-fidelity prototypes and may give more emotional responses, while they may overcompensate the aesthetic deficiencies of low-level prototypes. In their own study, testing the effects of prototype fidelity on a children's game by reverse engineering the game and testing two prototypes made from the game, they noticed there was little difference between the user experience of the prototypes and very few unique usability problems. In fact, low quality visuals seemed to be the greatest problem of their paper prototype, which had the most usability problems. Therefore, it seems likely that any usability problems found in the user testing of this prototype correspond very well to actual problems users would face when using the final version in a similar setting.

However, as we had to rely on more simple methods as a backup for generating the poems, children's reactions to the quality of the poems and the topics can only give some directions of how the computer is perceived as a creative writing partner. Some ideas and impressions may be collected to improve the quality of the material currently presented by the computer, but further evaluation with a fully functional prototype will need to be made to assess the viability of the computer as a co-writer.

Chapter 3

Background

This chapter reviews issues related to usability testing with children. The first section describes how usability is defined when working with children, visiting issues with traditional usability definitions and their applicability to testing with children. The concept of fun, is discussed as one of the usability factors often discussed with children. The second section visits practical issues applicable for most usability tests conducted with child users. Issues with selecting equipment, timing tests, defining suitable instructions and tasks, as well as selection of subjects, evaluator conduct and general ethical issues are discussed related to children's characteristics as a user group. Finally considerations on choosing methods, and the application of some specific methods for use with children are discussed in the final section.

3.1 How is Usability Defined when Working with Children

When assessing products with children it is important to look at how usability is perceived [25]. The dimensions regarded as most important will help setting goals for the usability evaluation and hence also affect the methods that are chosen for the evaluation. Children's software may have a different usability focus when compared to the traditional measures focused on productivity, speed and efficiency. Important dimensions often include engagement, or fun, although the traditional measure of usability error count is often still applied.

The ISO-9241-11 standard defines usability as the "[e]xtent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [16]. This can be considered as a very traditional definition and it has been seldom used

unmodified as the focus of usability testing with children, although Stinson et al. used the definition directly, and focusing their testing on performance indicators, such as fastness of navigation [36]. Diah et al. [3] report using another traditional definition: they use Nielsen's [1] definition, which considers learnability, efficiency, memorability, errors and satisfaction as suitable dimensions of usability to be measured.

These "traditional measures of usability" which focus on productivity, speed and efficiency are considered inappropriate for children's products by Hanna et al. [10]. Also Hourcade [13] criticises traditional measures unsuitable for children's use which often unstructured. However traditional ideas are rarely totally discarded, rather they are focused or augmented: For example Egloff [6] and Höysniemi et al. [14] report testing for learnability, and many authors collect usability errors (e.g. Lahti et al. [20]). When traditional measures are used, it is important to note their context is often related to office environments, which is rarely the case for children [13]. Therefore additional requirements need to be defined, or traditional constructs need to be re-defined to fit the new context better. For example, satisfaction should not be concerned only with freedom from discomfort, but with genuine fun. Additional criteria may include very field specific terms, such as assessment of game play, mechanics, playability, fun, feedback and immersion for games [41].

The idea of focusing on a subcategory of the traditional definitions is supported by Sim et al. [35], who note that the traditional measures can be used in the context of educational software, if the testing is focused on intuitiveness and reducing distractions for the objectives of the learner. Other focuses include engagement, which can be seen as a category of satisfaction. According to Hanna et al. [10] engagement includes constructs such as "familiarity", "control" and "challenge". The same constructs are also considered by Egloff [6] to be important when children decide, if a product is likeable or usable.

At times the focus of usability evaluation with children seems to be very specific: In addition to learnability, Egloff [6] measured also the clearness of used metaphors, ergonomics and gender preferences. And Höysniemi [14] et al. investigated how well their game avatars meet children's physical and social development needs. Höysniemi et al. [15] also conducted another study aimed at collecting user data to define suitable interaction patterns, instead of testing the usability of existing interaction possibilities.

The main additional usability dimension used for evaluating children's products is fun. It is a relevant element in keeping children's attention in

[1]Original reference: J. Nielsen (2003, August 25). Usability 101: Introduction to Usability. Alertbox, [Online] http://www.useit.com/alertbox/20030825.html

educational programs [24], and the most important goal for computer games [27]. Children seem to prefer fun products [24], and the willingness of children to use an interactive product seems to be greatly affected by how much fun they consider it to be [31, 35].

Fun is considered as a useful description [30], or even as a measure for user experience [34], and a useful differential when comparing products [31]. MacFarlene et al. [24] note that Carroll (2004)[2] has even suggested extending the concept of usability itself to include fun, but Sim et al. [35] do not consider Carroll's idea useful. But most importantly it has been shown that usability and fun are correlated in usability testing with children [24, 35].

Mac Farlane et al. [24] showed that fun in educational software correlates between observational data, user reports and ratings. The correlation of observed fun and observed usability is also verified by Sim et al. [35]. Observed fun is also negatively correlated with the number of usability errors in an interface [24, 35]. In addition, fun seems to be negatively correlated with negative attributes (ugly, boring, bad, childish, and confusing) and positively correlated with positive attributes such as 'great', 'fun', and 'exciting' [27]. However, other positive adjectives, such as 'beautiful', 'surprising' and 'easy', do not seem to have a similar correlation when used with 8-14-year-old children [27].

The description of fun is challenging: It can be seen as a manifestation of satisfaction in children [30], although MacFarlene et al. [24] consider it to be different, as it is not goal oriented. Read et al. [30] consider fun to have three dimensions, expectations, engagement, and endurability, which were identified by a previous study by Read and MacFarlane [3]. According to Obrist et al. [27] Hoonhout [4] considered seven factors, enjoyability, attention, challenge, curiosity, control, pride and presence, related to the fun experience. Sim et al. [35] however consider definitions of fun problematic, when they don't focus on pleasurable emotion. They argue that something can be engaging without necessarily being fun [35].

Despite problems with defining fun, the concept of fun still suits children's environment well [30]. Interestingly children as young as 7-8-year-old are already able to distinguish between the concept of usability and fun [24, 31, 35]. Although mac Farlane et al. [24] recommend fun as a way to keep children's attention in educational software, children may not see fun products suitable

[2]Original reference: Carroll, J. M. (2004). Beyond fun. Interactions, 11(5), 38–40.

[3]Original reference: Read, J. C., & MacFarlane, S. J. (2000). Measuring Fun. Computers and Fun 3, York, England.

[4]Original reference: J. Hoonhout. Development of a rating scale to determine the enjoyability of user interactions with consumer devices. Technical report, Philips Research, 2002

for the classroom [35]. Measuring fun is also important in order to avoid possible negative effects to learning [24].

3.2 Special Considerations Regarding Usability Testing with Children

Children as a user group pose some special requirements for usability test set-ups. In general, children are not able to concentrate on testing or solve problems as well as adults. In addition, children's motor coordination and perception limitations may restrict their performance as test users. The reliability of children as test participants, related to the recounting their experiences has also been questioned in literature. These features cause problems ranging from technical to ethical considerations. In general, testing with children requires readily adapting to children's own ideas during the testing, as children may spontaneously modify the test set-up itself, as witnessed by Costabile et al. [2], whose participants spontaneously took over a debriefing session.

The structure of this section has been inspired by the article "Guidelines for Usability Testing with Children" by Hanna, Risden and Alexander [9]. I have augmented the original guidelines with more details reported by practitioners in recent literature. I have also visited issues related to the cognitive, perceptual and motor development of children reported essential in Hourcade's [13] review targeted at usability professionals.

In the source literature, suggestions are usually made for specific age-groups, because children of a particular age usually share similar cognitive, perceptional and motor abilities. The factors discussed in this thesis apply mostly for children between the ages of seven and twelve. This group corresponds to the Finnish lower comprehensive school ("ala–aste"), and the intended user group of the Poetry Machine prototype fits into this age range. This categorisation roughly fits the elementary school group (6-10-year-olds) used by Hanna et al. [9] and Piaget's concrete operations stage (7-11-year-olds) discussed by Hourcade [13]. This division by school year highlights the idea of Hanna et al. [9], who consider that as children become accustomed to working with adults in the school environment, their ability to conduct themselves during usability testing is improved making them relatively easy to include in usability testing.

Children in the lower comprehensive school, have also other abilities making them more suitable for usability testing than their juniors: They are somewhat capable of describing what they see and do, although this ability

varies within the group [9]. Piaget's theory suggests children at this stage are more able to appreciate other perspectives in addition to their own, making them better team workers and possible design partners [13]. Markopoulos and Bekker [25] have found that some children in this group are already capable of thinking aloud. Patel and Paulsen [29] suggest that even users as young as nine are able to think aloud, and children may actually be more willing to do it than adults, as they are used to thinking aloud at school.

It is important to notice that categorising children by age is a generalisation, and individual variation occurs between test subjects. This individual variation may depend on inherent qualities, such as temperament of attention span [6] or be affected by task specific factors, such as information, social support and instructions [13]. For example, children, who enjoy themselves during testing will also be able to concentrate on testing for a longer time [25]. The individual variation is larger among younger children, meaning that "two five-year olds are more likely to show differences in the way they interact with software than two ten-year olds" [13]. Specifically within the group of 7-12-year-olds it is important to notice that the variation of practical skills varies also largely within the group, for example, children in the first year of school may not be able to read yet [20]. Specific domain expertise will also affect the performance of children, and experts may perform even as well as adults [13], or provide different kind of information during testing [4].

Some abilities of children are also related to the sex of the subject. For example the ability to distinguish object details has been found to be better in boys than girls of the same age group across all age groups [13]. According to Egloff, some studies suggest major differences in play or software preferences by gender, yet contradicting evidence exists including Egloff's own findings in studying 3-4-year old children [6]. Donker and Reitsma note that a study found that gender can affect the number of problems children cover during a usability test [4]. Due to the inconclusive evidence gender preferences were omitted from this literature overview.

In addition to individual differences, it is also possible for some of the group characteristics to change over time: Hourcade [13] notes that the basic IT skill level of children has risen throughout the years affecting the applicability of results between experiments. Some social aspects also affect how children use technology, as they learn from their seniors and adults [13]. Students in today's schools are also linguistically and culturally diverse and may have some special needs [20]. Therefore the guidelines presented in this section must also be adjusted according to the specific skills and motivations of the test participants.

3.2.1 Test Environment and Equipment

Usability testing may be a very unnatural experience for children, especially if they have to use unknown materials in unknown environments with strangers [14]. Höysniemi et al. conclude that this may easily result in the children feeling like they are being tested instead of the software [14]. Therefore the selection and set-up of the test environment and equipment are important. Giving children time to get used to the new environment will further minimize the amount of distractions [25].

Hanna et al. [9] recommend adjusting a traditional usability laboratory to make it more child friendly by adding some colourful images. Yet the excessive use of play items may direct the child's attention away from testing [9]. Nowadays many usability tests are conducted at schools (e.g. [15], [1], and [34]), at kindergarten (e.g. [3]), or even at shopping malls (e.g. [27]). Conducting tests at a familiar environment has its benefits, but Obrist et al. [28] note that according to Kaplan et al. [5] the school context may generate challenges of its own, including problems with gaining children's confidence and inspiring creativity. Yet these problems can be alleviated by using a different context, for example, by applying a paired test methodology [28].

Children's motor skills are restricted compared to adults; Although children's repetitive tapping and aiming skills are usually developed to a similar speed with adults, and by eleven they can copy simple shapes freehand, in practise seven to eleven-year olds still have problems with pointing [13]. Their visual abilities are also still developing [13]. In addition, children may be used to specific input devices. These factors have resulted in special recommendations for input devices for children's usability testing, especially in the early literature. Hanna et al. [9] recommend using input devices children are used to, especially with preschool children. High resolution monitors [13], or setting the mouse speed to slow [9, 13] are recommended to alleviate problems with unfamiliar pointers. However, later studies have found the mouse to be the best pointing device for children, regardless of the size of the mouse, additionally tampering with the speed of the pointing device may actually cause frustration [13]. Therefore these input device recommendations can be seen largely outdated. I would also argue that selecting specifically large monitors or tampering with the pointer settings may incorrectly represent or undermine usability errors children would encounter in everyday use with their typical equipment.

Using small recording devices and setting them up unobtrusively, for ex-

[5]Original reference Kaplan, N., Chisik, Y. and Levy, D. 2006. Reading in the wild: sociable literacy in practice. In IDC '06: Proceedings of the 5th conference on Interaction design and children. NY, ACM, 97-104.

ample to avoid children facing cameras directly, is recommended [9]. Yet Höysniemi et al. [14] report that contrary to some other studies, they have not found video cameras to be restricting or affecting the children's behaviour and Hanna et al. [9] also think that 6-10-year-old children are "generally not self-conscious about being observed as they play on the computer". In practise Egloff [6] suggests adjusting the location of cameras according to pilot tests.

The placement of cameras is related to the issue of the layout of the test space: Children should not face one-way mirrors [9] and setting of computers may need to be changed according to pilot tests [6]. In practise, compromises have to be made in arranging the layout: Höysniemi et al [15] found it beneficial to have the controlling facilitator of their Wizard of Oz prototype in the same room to make organizing at school and note taking easier. However, the clicking noise of the facilitators mouse made at times users pay unwanted attention to the facilitator.

Children are more likely to be relaxed in communicating with adults when a peer is in the test space. This is seen in action in many of the studies in which children are paired up, or act in groups. In practise, the same effect has been achieved by having children do tests alone but in the same room with other testers (for example [35], and [34]) or even just in the same room where peers are playing (for example [6]).

Reward practises used by different teams differ in the literature: Most importantly, researchers should make it clear to the participating children that their comments were helpful for showing what to fix in the tested software [9]. Material rewards mentioned by Hanna et al. [9] range from gift certificates and movie tickets to paying participants for their time. Some researchers do not use material rewards at all, but then again Patel and Paulsen even pay a finders fee for some parents who are willing to advertise the study to other parents [29].

3.2.2 Time Considerations

According to Hanna et al. scheduling enough time for usability test sessions with children is important [9]. Even though preschoolers may be able to focus on the test only for 30 minutes, it is important to reserve extra time for play and exploration [9]. Working with children is also tiring for the usability professionals themselves, and it is recommended to reserve extra time between participants [9].

The test itself should be kept short to allow for the short attention span of children [27]. Fun and engaging tasks may allow for longer test times [25, 27]. Shortness is also a virtue for surveys, as the drop in motivation

during a lengthy questionnaire will bias the results [31]. Even short breaks are a good idea in longer test sessions [9].

3.2.3 Test Instructions and Tasks

Testing with young children is often more focused on free exploration [9]. Freely determined tasks will also have the benefit of showing what children want to do with a product, instead of helping them to use it as intended [10]. However in practise, children as young as 3-6-year-old will need some structure to be applied into the test sessions in order to make the sessions more comparable to each other, control the type of information being gathered, and avoid boring the participants [6].

When designing test tasks, it is important to take into account that the working memory of a child is relevantly smaller than an adult's [13], restricting their problem-solving abilities. However, already children in preschool are capable of analogical reasoning and understanding causality, and children in the comprehensive school are in addition capable of using hierarchies in reasoning and problem solving [13]. Children in the lower comprehensive school are also able to reverse actions in their head when troubleshooting issues in software [13].

In general, tasks should be broken down to small segments where possible [9]. although even preschoolers are capable of carrying out tasks that involve a sequence of action, locations and objects [13]. However, short and varied tasks may help remedy short attention spans [6] [6].Mixing the order of test tasks is recommended, if possible, to avoid biasing the results due to the fatigue of participants [9]. If comparison across products is done as a part of the test, it is important to design the test tasks to be as similar as possible (see for example [35] or [6] for samples).

Tasks should be described in language understandable to children in order to avoid extra cognitive effort needed to understand complex instructions [39]. For example Höysniemi et al. [15] needed their test participants to play the tested game at a specific area of the test room, in their instructions they called that place 'The magic square'.

It is also important to remember suitable post task and post test activities. This debriefing is "supposed to capture subjective input from the child that cannot be captured using the task-based questions" [3]. Hanna et al. state that older children may enjoy giving improvement ideas [9]. They are also able to give reliable ratings of different software aspects with suitable

[6]Original reference: LARKIN, S. 2002. Usability, Jr. – How to run a successful usability test with children. STC Usability SIG Newsl. 8, 3 (Jan. 2002).

tools, such as scales relying on smileys [9].

3.2.4 Selection of Test Subjects

The selection an characteristics of suitable child participants for usability testing is addressed largely in the literature. Level of expertise, personality, relationships between researchers and among subjects is discussed, and strategies for recruitment shared. In general, abilities required from test participants depend on the method chosen, and Markopoulos and Bekker [25] have come to the conclusion that these requirements should rather be assessed directly for each participant than through age-group generalisations. Therefore it may also be viable to rule out children with serious disabilities, or children who have not yet reached the specific parameters required of a participant. For example Stinson et al. [36] excluded adolescents with cognitive impairments. Diah et al. note that participants should not only reflect the characteristics of target users, but also be likely to use the selected software [3].

In general, test subjects should not be familiar with the tester. Hanna et al. [9] recommend avoiding even the children of colleagues, as they may have too extensive knowledge of the products developed by their parents, or simply be reluctant to give negative feedback on the product. Patel and Paulsen [29] agree that recruiting children who have ties with the research organisation may bias the results and prefer to seek users elsewhere, however they have at times successfully recruited children of other employees.

In selecting participants for paired or grouped test settings, it is important to consider the relationship between users themselves. Views on whether or not subjects should know each other depend on the researchers and methods used. Patel and Paulsen consider asking children to bring their friends suitable for one-to-one interviews, but not for focus groups, although they note at the same time that bringing a friend may make the experience less intimidating and more fun for the children [29]. Höysniemi et al. specifically note that the participants of their pair testing experiment were paired up according to the children's own preferences [15]. Hanna et al. also reported using pairs of friends [8].

In practise, sample sizes of experiments are restricted by practical constraints [8]. I also noticed that many tests seem to use very unvaried users, in that users are often recruited within one class (e.g. [35], or [39]) or school (e.g. [34], [15], or [14]), and many other studies do not report the recruitment method used (e.g. [3], or [31]).

Pre-screening or selection of candidates is very often performed by teachers: Teachers formed the groups studied by Costabile et al. [2] to "guarantee

social and cognitive homogeneity". Teachers also selected participants for Höysniemi et al. [14]. And Egloff relied on nursery teachers to pick suitable candidates for her study [6]. Patel and Paulsen [29] then again call the parents to ask the necessary information.

Hanna et al. recommend that the test subjects should have at least some experience with a computer, however it is important to select children, who do not have too much computer expertise, or the results may be biased [9]. This idea is supported by Hourcade [13], who notes that children may perform as well as adults on their expert domains. It is recommended to document the background and expertise of participants in order to ensure that they have a similar level of expertise to the intended users of the software [13].

It seems that the level of expertise in a specific program can effect the type of comments made by children during usability testing. Donker and Reitsma [4] tested a game with a Talk Aloud-method with 6-7-year-old children and found that novices made more comments on the feedback and instructions of the program, and on wanting to stop practising, whereas experts needed more specific help. Children were not prompted during the test to talk aloud, but guided to do so before it, as a result children in the experiment made very few remarks – only 28 of the 70 participants made any remarks, resulting in a total of 54 remarks. More importantly, Donker and Reitsma concluded that testing with novices and experts will reveal different types of problems: problems observed only with novices are problems children can solve by themselves, whereas the behaviours of experts indicate persistence of problems and their remarks identify unanticipated problems and note their importance. However, in practise users with prior experience with the tested program are often excluded from testing. For example, Höysniemi et al. [14, 15] report specifically choosing participants, who had not played any previous versions of the game they tested in their studies.

The personality of a child participant may affect test results: According to Barendregt et al. [1] ideal child candidates for usability testing are curious and extrovert, yet at the same time not too friendly to be too polite. Barendregt et al. measured the personality traits of the participants with a Dutch personality test called 'Bilkvanger' which was completed for each child by their parents. The experiment found that the selection of curious participants correlated positively with the number of revealed problems and the selection of extrovert, curious, but not too polite participants had a positive effect in the number of problems identified verbally by the participants. While it is usually not possible to select ideal candidates by all the measurements, Barendregt et al. suggest practitioners choose curious children, if a large amount of revealed problems is desired, and extrovert, but not too polite children, if more verbal information on the problems is desired. The

researchers note that it might also be possible to prompt less extrovert children to talk more. Interestingly, Markopoulos and Bekker [25] reported that in their comparison of think-aloud, interviews and post task questionnaires, the number of problems found seemed to correlate with testing method and gender, but not with verbalisation or extroversion skills. They also found that girls reported more problems than boys [25]. In practise, personal characteristics have been used to select participants also by Hanna et al. [8], who relied on parents reports to exclude candidates characterized as shy.

3.2.5 Evaluator Conduct During the Test

Usability testing with children usually requires the attendance of adult evaluators. Here evaluators mean researchers, who can act in multiple roles. The roles can be specific to a method, such as a Wizard in the Wizard of Oz method, or more general, such as an observer. The evaluator, who's role is mainly to guide and interact with the child participants is usually referenced as a facilitator, or an interactor in the literature.

It is important to establish a relationship between the participating child and the evaluators before testing begins [6, 9, 28, 29]. A good relationship will reduce researcher impact [31] and guarantee success of the session [29]. A relationship can be established through small talk, and additional sense of trust and control by showing the child around the laboratory before testing [9]. It is also important to motivate the child and emphasize the importance of their role as tester, which also helps to set the child's expectations appropriately for what happens during the test session [9]. Careful planning is needed for design activities with children, to avoid frustrations of loss of interest [28]. Hanna et al. even recommend planning a test script for introducing the child to the test situation [9].

Hanna et al. note that younger children need an adult tester to remain in the room with them for reassurance and encouragement. As children may be agitated about being alone, or being alone with a tester, parents may be present during the test. Therefore it is also important to advice the parents not to participate, and stay out of sight. The parental influence can in part be reduced by keeping the adult tester present in the test situation. Siblings may be tempted to give instructions to the child participating in a test session, and therefore they should stay in another room.[9]

Children may need continuous encouragement and direction towards finishing up the tasks. Hanna et al. [9] note children should not be asked if they want to do something, but directed with active phrases, such as "Let's do this..." to do a task. Some children may also be encouraged by asking them to help the researcher to complete a task[9]. Generic positive feedback

is important if the child feels like they are failing to figure out the software[9]. Showing appreciation and gratitude to emphasize the child's input as a tester is recommended by Diah et al. [3] and praising children when they perform well may be helpful [14]. Some children will also need help in reading out instructions [9] or surveys [31].

However, instructions during the test should be given sparsely. Höysniemi et al. recommend giving instructions only when children seem frustrated or don't seem to understand what they are meant to do [14]. Donker and Reitsma recommend answering any questions with counter-questions [4]. In practise some number of prompting to keep children involved with a task is needed regardless of usability test method [39].

3.2.6 Ethics

Due to the vulnerable nature of children as a user group, ethical considerations are especially important with this user group. Institutional organs may need to approve test set-ups before testing and trust needs to be established between the research organisation and the parents of the test participants [29]. Investigating children in the context of school requires informing, motivating and integrating multiple stakeholders, including parents, teachers, principals, authorities and the children themselves [28]. Often permissions are required from all of these parties [28, 29]. Obtaining the needed permissions may take weeks [29]. It may also take a long time for parents to respond to any inquiries [29]. Possible confidentiality agreements should also be signed by the parents [9]. To make the child understand the agreement, [9] recommend telling that the design is "top-secret".

It is also important to remember that although consent is usually established through parents, it must be clear to participants that participating in a usability test is voluntary. MacFarlane et al. report that all children participating in their research were given the opportunity to leave the research activity before and during the work [24]. Waller et al. then again gave their participants a smiley stamp they could use to show their consent in participating in the testing [40].

Patel and Paulsen also recommend showing parents the test room to protect the participants and give some legal protection to the staff. They also recommend a videotape to be kept constantly on, to offer protection for everyone involved [29].

Ethics also affect the choice of methods, when planning a test set-up with children. For example Höysniemi et al. decided that a Wizard of Oz approach would be best to test possible movement strategies for a game, since this test set-up would be less unpleasant, frustrating, and discriminating than

conducting a usability test with a poorly designed early version of a machine vision system [15]. Additional procedures may be needed to make testing fair for all children. For example Sim et al. considered it important for ethical reasons, to give all children participating in the testing of low fidelity prototypes reverse engineered from an existing computer game a chance to play the actual game in the test situation as well [34].

3.3 Common Methods Used in Usability Testing with Children

Usability testing methods need to be tailored for the target user groups [25]. In this section I discuss criteria for selecting specific methods for usability testing with children and then visit some issues in applying some specific methods for use with children. I have limited the discussion of method details to those used in the experiments described in this thesis.

In general, usability test methods for children are valued by their ability to detect as many actual usability problems as possible [25], their running costs [25], as well as their ability to elicit authentic verbal comments from children [39]. Additional factors for comparing methods include required materials, participants and procedures, as well as what is required of the children participating in the tests [25].

Despite existing metrics, systematic evaluation of usability test methods for children based on literature is difficult. The ability to detect problems can not be measured across case studies as they test different products. Costs are usually not reported, and users used in case studies are usually incomparable, prohibiting analysis by the number of comments elicited. Additionally, language and details reported varies among studies, despite Markopoulos and Bekker [25] presenting a framework in 2003 for comparing usability test methods for children. Finally some comparative evaluations of usability test methods for children exist, but they are often weak due to small sample sizes.

Based on the literature reviewed for this thesis it is apparent that most usability testing methods used with children are derived from existing usability testing methods intended for adults. The details given on the modification of the methods for children varies a lot. For example Think-Aloud has been used with children by Markopoulos and Bekker [25] and Donker and Reitsma [4], but notes on its modification are limited to suggesting more frequent prompting, a larger amount of tests, and simultaneous observation to counter children's limitations with verbalising their thoughts [4].

Test settings involving multiple users are clearly popular for testing with

children. For example the rationale for choosing a Pair Testing includes discouraging shyness [15], balancing the ratio between adults and children in the test situation to help children communicate with adults in a more relaxed manner [15], naturally promoting discussion between the children [15] and promoting open and honest discussion [8]. A similar rationale has been used for Co-Discovery, which was used by Markopoulos and Bekker [25] to elicit comments more naturally. The problem of usability test set-ups involving multiple users is their vulnerability to peer bias. However Markopoulos and Bekker [25] considered this to be irrelevant for Co-Discovery and Hanna et al. [8] suggested starting paired interviews by individual ratings.

Based on the literature, I found that most usability tests with children use very basic methods, including observation, questionnaires, or interviews. These methods are used variedly as building blocks or support for more complex methods, or even as the main method for some studies. In general, it is advisable to select a method mix in order to gain various perspectives to the product under testing. For example, although behavioural signs are often trusted more than verbal reports in the literature, it is important to remember to support observational findings with suitable post task activities. For example Diah et al. [3] used a structured post task interview, resembling a questionnaire to find out about the subjective experiences of the users.

When selecting a suitable survey method, be it a questionnaire or an interview, the selection of methods depends on the type of information needed for the project [31]. A project should not rely entirely on one survey [31]. For example, different types of ranking tools reveal areas for development, but do not give concrete improvement ideas, which may need to be acquired through different means [31]. Read and MacFarlane [31] also advice against applying statistical tests to children's responses, as they are not necessarily stable over time [31].

With any method, be it complex or specific, special considerations have to be made to modify it for testing with children. The next subsections describe considerations related to modifying basic methods, which were used in this thesis to support other methods and as building blocks for designing the Group Testing setup with the Feedback Game. I then move on to describing the Peer Tutoring method, which was selected as the main test method for the study presented in this thesis.

3.3.1 Observation

Observation is one of the most used methodologies in usability testing with children. Many authors use observation either as a part of a test set up, or rely mostly on it during testing (e.g. [3, 6, 10]). The popularity of observation

can be partly explained by the inclination to think children's verbal comments unreliable. For example, Hanna et al. [9] consider observing behavioural signs to be more reliable, as children may be too eager to please adults with their comments. Egloff [6] reports similar notes from earlier research. In practise behavioural signs are trusted as indicators of engagement [9, 30, 35], enjoyment [35] and fun [24] and especially negative signs are considered important, as Hanna et al. [10] note "While children may not display much positive emotion in the laboratory, any sign of negative emotion deserves attention".

Diah et al.[3] describe the observation method as a data collection method in which users are observed while performing authentic tasks on a prototype or the evaluated application. Observation can be used in different situations and with a different number of users, including single users, pairs, or groups of users. Tasks may may also vary. To enable more robust measurements, observation tasks, help and questions may need to be carefully designed [6]. Observation results can be qualitative or quantitative. Quantitative observation results can be quantified in different ways. For example Sim et al. [35] counted a usability score for the software tested in their experiment by counting positive issues and subtracting the negative issues.

Observer tasks during a test vary a lot, some observers in testing with children are expected to remain silent, some are required to interact with the participant: For example Costabile et al. [2] used ten passive observers to observe groups of children in context when they tested a mobile learning tool. Egloff [6] and Diah et al. [3] both talked about testing with young children and decided to use only one observer, who would also act as a facilitator helping children and asking them questions during testing. At times the tasks of observers have been divided more clearly. For example Sim et al. [35] distributed the responsibilities of the two observers participating in their experiment by assigning the first observer to focus on usability issues and the second on the non-verbal signs of the participants.

Observation as a method is highly subject to observer effects resulting from the observer's subjective decisions. Different tactics for reducing observer effects in the literature include check lists, which can be comprised by eliciting a list of pre-defined problems through expert evaluation [4], or pilot tests [2], or simply by evaluating what is the ideal way to use the functions of the application [3]. However, sometimes it is too difficult to comprise a check list due to the wide range of possible options [24, 35]. Other methods include rotating observers between tests [24] and improving coding practises in a pilot test [2].

When observing behavioural signs, it is important to decide what gestures are noted and what kind of emotion they usually represent. Signs of

engagement usually include smiles or laughs [9, 24, 30, 35], or positive body language, such as leaning forward to try things out [9, 24, 35]. Commenting is usually considered a positive sign [24, 35], and Read et al. [30] consider positive signs to include concentration signs, such as fingers in mouth, or the tongue out. Signs of disengagement include frowns, sighs, yawns and turning away from the computer [9]. MacFarlane et al. [24] and Sim et al. [35] mention sighs and looking around the room. However concrete examples of gestures were not shown in any of the literature reviewed.

3.3.2 Interviewing

Interviewing can be used in a multitude of test situations for example to gather ideas or discuss problems after tasks. Interviewing children is challenging, as children's ability to answer questions greatly impacted by their developmental stage, such as language ability, as well as their temperament, including confidence, self-belief and the desire to please [31]. The interviewing process also requires children to express themselves verbally requiring more of them as test subjects [25]. Children are also tended to tuning their answers according to the person interviewing them [31]. However, Hanna et al. [8] found that 8-9-year-old children were already relatively consistent when interviewed.

Interviewing requires careful formalisation of questions. If children can not understand the question or interpret it incorrectly, they may answer superficially [31]. Special concern has to be paid to the question format: For example children tend to answer 'yes' to 'yes or no'-type-of questions, therefore free-recall type-of questions are preferable [31]. In practise it may be very difficult to ask direct questions from children. For example, Hanna et al. [8] noted that children seem to have problems with coming up with suggestions for improving game concepts and tend to offer suggestions from familiar games. Instead of asking directly, they suggested listening to comments and observing what children like and dislike.

3.3.3 Questionnaires

Questionnaires are data collection format that requires the user to fill in a form, which consists of open and closed questions, including multiple choice. Questionnaires are usually used with other techniques to collect for example background data, but Obrist et al. [27] relied mainly on questionnaires for their massive testing of a game at a local mall. Questionnaire items need a lot of consideration, and well tested tools are recommended, as they reduce

misunderstanding and adaptation to facilitator's opinions [31]. Younger children have been found to be tended towards answering questionnaires over positively compared to their elders [27, 31].

Parental help may be needed to answer some questionnaires, such as background questionnaires for collecting demographic data. This approach has been used for example by Diah et al. [3] and Höysniemi et al. [15] [14]. Hanna et al. [10] have applied this even to other questionnaires during tests involving participants under 5 years of age, asking parents to answer two questionnaires, one for their child and one for themselves. Many authors also administer questionnaires verbally to avoid the need for reading and writing. For example Diah et al. [3] used this approach.

The considerations applied for interview questions apply for questionnaires as well. In addition, Read and MacFarlane [31] advice against requiring lengthy writing tasks from children during questionnaires. Some research projects have successfully applied well known questionnaires for use with children, including SUS (System Usability Scale) used by Lahti et al. [20] with slightly older children, and standardised questionnaires for measuring learning motivation, strategies and behaviour used by Costabile et al. [2]. But details to the modification of these questionnaires are few. Costabile et al. [2] mention mostly making the questionnaire visually more appealing and adding two open questions.

Fowler [7] reported that some previous studies using Likert-scales have found young children putting their answer to the anchor points, despite observations supporting a much more varied range of experiences. An alternative question format for questionnaires are Visual Analogue Scales (VAS), which allow children to answer on a scale represented with images related to a feeling. According to Read and MacFarlane, the VAS has been found useful, although it should be used with older children as younger children have a tendency towards being positively biased [31]. It is also important that the VAS options are completely labelled, as this improves reliability of the answers [31]. It is also important that the language used is clear, avoiding any vague or ambiguous words [31].

Perhaps the most well researched questionnaires intended for children belong to the Fun Toolkit. It is intended for measuring fun with 5-10-year-old children [30]. It has been researched extensively through many statistical studies including those of Read et al. [30], MacFarlane et al. [24], Read and MacFarlane [31], and Sim et al. [35]. The toolkit itself consists of four tools: The Funometer, the Smileyometer, the Fun-Sorter and the Again-Again Table [30], [31], which can be used to evaluate different constructs related to fun. As the tools measure similar constructs it is not advisable to use all or them, but to choose the tool by the age-group and the number of

objects that are being measured. Most tools can be used to evaluate different types of products and they can be administered verbally, leaving the children only the task of marking down the desired rating, or order of instances.

The Funometer, originally developed by Risden, Hanna, and Kanerva (1997) consists of a vertical analogue scale with a smiling face on top, and a frowning one on the bottom [30][7]. It is used by the child by drawing a vertical line along the scale, imitating a thermometer reading [30]. Hanna et al. [10] found children to respond more reliably on this pictorial representation and the concepts of more and less than to Likert-type scales. It is considered to be more useful with older children [30], and not very useful for rating concepts [8].

The Smileyometer is a discrete version of the Funometer. It is a VAS, corresponding to a five step Likert-scale with options ranging from 'Awful' to 'Brilliant', with all steps presented by smileys [30]. The pictorial presentations were developed with children [30]. Later it has been considered that the Funometer and the Smileyometer produce very similar results and therefore it is recommend to use only one of them [30], [31]. Young children (7-8-year-olds) tend to report their experiences with the Smileyometer over-positively [24, 31]. Regardless Fowler considers Smileyometer a useful tool for children aged seven and above [7].

The Fun-Sorter is a table which can be used to rank different objects, such as activities, or options by specific criteria, such as their relative funniness, or ease of use [30]. The use of the Fun-Sorter requires clearly identifiable objects in order to be used. In practise the Fun-Sorter can be completed by using pictures to represent different objects and giving children a chance to physically rank each object on the table, making it easy to edit for the children [30]. This makes the activity similar to other card sorting tasks, such as that used by Hanna et al. [10] or the Talking Mats®[26] used as a discussion aid with people with communication problems. It seems that children are clearly able to distinguish different ranking criteria when using the Fun-Sorter [24, 30, 31]. However there are some indications that some criteria, like 'good for learning' may still be difficult for 7-8-year-olds [35]. Additionally Read et al. [30] found that some children in their experiment tended to order objects similarly for every criterion or adjust their charts if they felt an object was doing too poorly. To avoid the tendency young children have to rank objects 'fairly' Read and MacFarlane [31] recommend having the Smileyometer and the Fun Sorter on different papers if both are

[7]Original reference: Risden, K., Hanna, E., & Kanerva, A. (1997). Dimensions of intrinsic motivation in children's favorite computer activities. Society for Research in Child Development, Washington, DC.

used. If movable cards are used to represent objects, researchers will have to copy the results on paper [30]. Also, having too many rankable objects may make the Fun-Sorter too cumbersome to fill [30].

The Again-Again Table is a simple form in which children can mark how happily they would like to do an activity again [30]. It is laid out as a table with activities listed on rows, and options 'Yes', 'Maybe' and 'No' on columns [30]. In a trial by Read et al. [30] the results of an Again-Again Table filled immediately after an activity seemed to point to same activities the children identified as most fun in a post test session conducted after two weeks. The Again-Again Table seems to measure similar aspects with the Fun-Sorter [30], but Read and MacFarlane [31] consider it is better to use the Again-Again Table to measure fun, and use the Fun-Sorter for measuring other constructs, such as 'ease of use'. Again some results indicate that the Simleyometer and the Again-Again Table measure similar aspects of fun, and Read and MacFarlane [31] consider the Again-Again Table to be more useful for measuring fun with young children.

3.3.4 Peer Tutoring

Peer Tutoring is a usability test approach, in which users tutor each other in the use of the evaluated software [14]. The use of the method for testing with children is based on the idea that the application under evaluation can be looked at as part of a child's play, making the teaching situation analogous to explaining the rules of a game [14]. The benefits of the method include the promotion of peer-to-peer communication in the test situation, providing information on the teachability and learnability of the tested application [14]. It offers testing a real social context in which users are equal in authority and knowledge and offers the children a chance to take an active role in the usability testing diminishing problems caused by the child-adult relationship unbalanced by authority and knowledge differences [14]. Peer Tutoring is unobtrusive, as it is intended to be conducted in the natural environment of children and engage children so that they do not even notice that testing is happening [14]. Höysniemi et al. [14] also report that previous studies have shown the peer tutoring approach to be effective in fostering creativity, experimentation and problem-solving skills. The method is thoroughly described in the article by Höysniemi et al. [14], and additionally it has been used in two comparative analyses where it was found to be useful for eliciting comments [5, 39].

Höysniemi et al. [14] defined the role of a peer as children of similar age and status belonging to the same classroom. In their study, tutors had a possibility to try out the game for a couple of times before teaching the

other child. They note that it has been reported that peers may be more able to help each other than adults would be, but at times the tutee may feel less competent than the tutor. Therefore the tutor and the tutee should not differ too much in knowledge. Edwards and Benedyk also tried Cross-Age Tutoring [5], but in comparison, it elicited fewer comments from participants.

The key of Peer Tutoring is in peer-to-peer communication, which is a good alternative for thinking aloud [14]. This communication can be used to evaluate how well children have learned skills needed for using the application, how they perceive the interaction, and what kind of language should be used instructions intended for children [14]. Höysniemi et al. [14] note that children only teach thinks they like, understand, and feel important. The teaching in the sessions usually follows similar patterns, which show how children behave and categorise information [14].

Höysniemi et al. [14] tried the peer tutoring approach with two different set-ups, two-on-one tutoring and one-on-one-tutoring. In both of the set-ups researchers first introduced themselves to the children. In the first option, two of three children first participated in tutor training, while the tutee waited in the classroom. The two tutors were showed how the game is played, after which they took turns in trying it out. After the rehearsal, the tutee entered the room and the tutors were asked to teach him the game. The tutors gave praise and further instructions during the game. In the one-on-one approach a group of four children goes through the test, paired tests chained after each other. It was started by teaching the first child of the four to be a tutor. He then continued to teach the second child, until that child had a chance to play the game alone. After this the second child could tutor the third and so forth. Hösniemi et al. preferred the each-one-teach-one approach as it combined both paired and individual testing at the same test. Edwards and Benedyk [5] used the each-one-teach-one approach in pairs.

During the peer tutoring session Höysniemi et al. [14] at times had to help the younger children to give instructions to their peers. They used a simple question asking protocol, in which they asked the tutor a question, such as "where should you be when playing the game", to get the tutor to give more detailed instructions. This approach ensures that the researchers hear authentic language and honest answers from the children [14]. As questions can be directed to the tutor during the test, the facilitator does not have to bother the tutee, who can concentrate on the task [14]. The only problems with the question asking recorded by Höysniemi et al. [14] is that the children tend to direct their answers at the researcher instead of the tutee.

Höysniemi et al. [14] note that for peer tutoring to work it is important that same sex participants should have a friendship relationship in everyday life, as otherwise the participants may act hostile towards each other. Dif-

ferent sex tutor and tutees tended to behave more respectful towards each other [14]. Van Kersten et al. [39] reported that in their test tutors sometimes took over the task completely. Höysniemi et al. [14] suggest battling this problem by setting up the test so that it physically discourages these types of intervention.

Höysniemi et al. [14] conclude that the problem of the Peer Tutoring approach is that it demands a lot of work, both in organizing the test sessions as well as analysing the video material. In practise the waiting time for the tests held by Höysniemi et al. [14] was also too long. Höysniemi et al. [14] also think that the method is restricted in giving information on children's mental models and features they do not understand or like, and recommend using additional methods to investigate these factors. Previous experience in working with children is beneficial and Peer Tutoring could be paired with some participatory design methods [14].

Chapter 4

Applying Usability Testing Methods for the Poetry Machine

This chapter first describes the rationale behind the selection of methods for evaluating the usability of the Poetry Machine concept. This is followed by a description of the criteria for selecting participants for the testing and finally this chapter presents two methods, Peer Tutoring and Group Testing, which were selected to be used in the evaluation. Additional survey methods, actual participants and the analysis methods used with each method are described in the corresponding sections.

4.1 Selection of Methods

In choosing the evaluation method it is important to establish the purpose of the evaluation and what type of data is to be captured [34]. To establish this, three dimensions, usability problems, usefulness, and enjoyability, were selected as evaluation criteria for the prototype. Further requirements, such as the context of testing and the type of collected data also affect method selection.

The selected dimensions depict two traditional aspects of the ISO-9241-11 definition of usability: effectiveness and satisfaction. I discarded efficiency, as I do not consider it a major goal for a creative writing application. Also no relative metrics for efficiency can be determined as the evaluation is not comparative. Enjoyability was selected here as a term to describe fun in addition to satisfaction.

Evaluation Goal	Related Questions		
Usability Problems: Is the product usable for children?	Are children able to use the product?	Do the users find all features?	Are the users aware of all features?
			Do users understand the features in the same way the designers do?
			Do children understand the actions needed for using a feature?
		Are features easy to use?	Are users making mistakes with the feature?
			Is the sequence needed to activate a feature easy enough to complete with the physical and mental skills of the users?
		Do the children exhibit signs of frustration due to usability problems?	
		Do children verbally indicate problems with a feature?	
	Is the interface graphically pleasing to the children?	Do users indicate disliking the interface colours, fonts or shapes?	
		Are the chosen icons identifiable to the users?	
Usefulness: Is the concept useful for children practising creative writing?	What features of the program are the most useful for children?	Do children use all of the features or stick with a few?	
		Which features are used most often?	
		Which features do the children name when talking about the concept to a peer/ to a researcher?	
		When asked about the features, what are the children's motivations for using/not using a feature?	
	Does the concept make creative writing easier for children?	Do the adults (parents or teachers) notice a change in the capability of the children when working with the program versus working on a creative writing task on their own?	
Enjoyability: Is the concept fun for children	Do the children exhibit negative signs, such as signs of boredom or frustration?		
	Do children exhibit positive signs, such as smiling, or willingness to continue the activity for a longer period of time?		
	What activities do the children name when asked about the most fun/boring items in the program?		

Table 4.1: Goals for evaluation

To select suitable methods, a study question for each goal of the evaluation was formed. The questions were further divided into sub-questions. These questions are presented in Table 4.1. Noticeably the usefulness and enjoyability parameters are represented by questions focusing on the concept instead of the prototype. This is important, since early testing of the concept with a paper based prototype was not possible. Therefore this test session is the first contact children will have with our poetry writing system and it is also important to gather user impressions and ideas related to the concept itself.

Many of the related sub-questions illustrate that most of the questions can be answered with observation based methodologies: Usability problems can be gathered by signs of frustration and the enjoyability of the concept can be evaluated with similar indicators. Usefulness of the concept can be also determined partly by observation, but it gathering opinions from secondary sources, such as teachers, is also important. Yet many of the questions can only be answered either by asking children directly, or preferably by listening to the discussions among peers.

The main context of use for the PM tool is school. But it will also be possible to use it at home. Therefore we need to test it in two types of situations: Unassisted use, and in assisted use with a teacher. Although Lahti et al. [20] note that the usability of e-learning applications should be considered for two separate end-user groups: teachers and students, this thesis focuses on the usability evaluation from the children's perspective and methods for evaluating the usability from adult perspective are not considered. Therefore the teacher should be concerned with the testing only from the perspective of the pupil and as a possible informant to assess the usefulness of the concept.

Assisted use can be covered well by observing a lesson held by a teacher using the PM prototype and interviewing the children after the session. To accommodate the tight schedules of school life and the attention span of the pupils, we will need to design this session to be short, and the related interviews need to be conducted simultaneously. To fill these needs, we designed a Group Testing set-up, with a Feedback Game as a post task.

Evaluation of unassisted use is more challenging to organise, since it is preferable that children will participate in it with a friend to make the testing as comfortable for them as possible to avoid bias from discomfort. However I also wanted to see how users perform alone. Therefore the Peer Tutoring approach, using a paired each-one-teach-one system discussed in section 3.3.4 was selected. Peer Tutoring has been previously successfully implemented in school conditions and it has the opportunity to offer a natural context for the use of the tool, while reducing problems resulting from the unbalanced adult-child relationship. According to Höysniemi et al. [14] it can additionally give

information on the learnability of the tool, which is important for unassisted use of the tool in home conditions.

Additionally suitable interviewing and questionnaire methods were considered for the Peer Tutoring approach, such as the Fun Toolkit. The Fun Toolkit was not used, as it was considered more important to gain insight into why children had problems during the test, instead of how much they enjoyed specific features. It was also difficult to distribute the use into distinct features in order to evaluate them on the Fun Sorter. Therefore a short post-test interview with more open questions was designed instead. However the FunToolkit was used as an inspiration when designing the Feedback Game for the Group Testing session.

4.2 Recruiting Participants

As discussed before in section 3.2.4, gaining access to suitable test users is difficult when children are in question. We decided to use a similar approach to many other studies, and first recruited a teacher, who chose suitable test subjects from his class. I approached the same school in Espoo, in which I had observed the second year class on the previous year to gain insight into how young children work with computers in practice. Since the school has a history of participation in different kinds of educational research activities, the children are fairly good candidates for such research: They are used to adult visitors and therefore some of the problems related to meeting strangers could be alleviated. Also the pre-established relationship between the author of this thesis with the children may be beneficial for gaining their trust.

The observed class uses a co-teaching approach in which all pupils of the same age group are taught together by multiple teachers. We contacted one of the teachers directly by e-mail and phone, to set up a meeting to discuss the possibility of conducting usability evaluation, and presented what kind of methods we intended to use. The teacher was very enthusiastic about participating in the research and agreed that both the Peer Tutoring and Group Testing approaches would be suitable for his class. The tests could also be arranged at the school, with the benefit of having an environment designed to be child friendly, without being too distracting. As a benefit of the co-teaching approach, the class has multiple rooms in use and it would be very natural for the children to take part in the usability testing during school hours in one of the classrooms intended for small group teaching. The teacher would also be able to act as the tutoring teacher for the Group Testing set-up, while other teachers would continue the normal school day with the rest of the class.

After the initial details were clarified, we discussed the amount of pupils needed for testing. We agreed upon six pairs for the Peer Tutoring and two small groups for the Group Testing. The teacher would select suitable pupils from his own class by the following parameters:

1. Users must be 8-11-years-old
2. Users must go to the Finnish comprehensive school
3. Users in pair testing must be friends with each other and preferably the same sex
4. Users participating in the group test must be from the same class
5. Users must have elementary knowledge of computer use (from home, or preferably from school)
6. Users must have elementary reading and writing skills (they must be able to write complete sentences and read short paragraphs of text)
7. Users must be able to follow short instructions

We also noted that it would be good to get a balanced distribution of boys and girls. Most of the parameters were quite clear, since we were testing within a single classroom of 3rd year pupils, but we discussed the required skill set more specifically. The teacher noted that his class did have a number of pupils with special needs and we discussed the possibility to include them into testing. We agreed that if a pupil did not suffer specifically problems related to the understanding of language the pupil would be eligible for the research. Also the requirement of the Peer Tutoring participants preestablished friendship and the pupil's ability to follow short instructions were discussed in more detail, in order to ensure that all participants would be co-operative with other children as well as with the researchers.

The teacher selected the pupils according to the plan. As the school already had a functioning approach to securing the research permissions from the parents, the related forms were given to the teacher, who distributed them to the pupils taking part in the research. Each pupil brought the permission form and a leaflet with information on the study home. The teacher collected the written permissions from the children and the forms were checked by the researcher at the beginning of the tests.

4.3 Peer Tutoring

The Peer Tutoring method used in this evaluation is based on the each-one-teach-one style Peer Tutoring method as it is presented by Höysniemi et al. in [14]. In the each-one-teach-one Peer Tutoring method first one child is tutored by an adult, after which the child continues to tutor another child. The tutoring can be continued by having the second child tutor another and so forth. The approach used by Höysniemi et al. is described in detail in section 3.3.4. The main differences to the approach by Höysniemi et al. is that in our method we use pairs of children, instead of groups, and the first child is not directly tutored by the adults, but is encouraged to try out the tool with as little adult guidance as possible. We decided to use pairs instead of groups, because we wanted to interview the children after the test, and it was deemed better to do it in pairs than alone, or in groups. We also wanted to gather more information on how children used the tool alone and therefore optioned for a larger amount of pairs, versus a smaller amount of groups. The each-one-teach-one approach has been used for pairs by Edwards and Benedyk [5].

The Peer Tutoring set-up can be divided into five phases: In the first phase, tutor introduction, the researchers introduce themselves to the first pupil, who fills out a background questionnaire. In the second phase, tutor training, the tutor familiarises him or herself with the prototype and writes a short poem with it. In the third phase, tutee introduction, the researchers introduce themselves to the second pupil, who fills out the background questionnaire. In the fourth phase, peer tutoring, the tutor guides the tutee in writing a poem with the prototype. In the final, fifth phase, the pair interview, the tutor and tutee are interviewed as a pair about the prototype, and some of the problems the pair faced during the testing.

4.3.1 Participants

Twelve pupils participated in the Peer Tutoring. Six of the pupils were male and six female. Tests were conducted with six same sex pairs. The pupils were selected according to the requirements stated in section 4.2 by their teacher. The teacher also formed the pairs Before starting the actual testing with either the tutor or the tutee present, each of the participants filled in a background questionnaire. The questionnaire can be seen in appendix A. It was designed to be easy for the children to fill: Instead of tick-boxes users may circle suitable alternatives, and they are encouraged to write more, if needed. A lot of room is given for the free text fields. The questionnaire

was pilot tested with a nine-year-old. During the test, the researchers read the questions aloud, and asked additional questions to ensure that the pupils answered thoroughly. Table B.1 in the Appendix B lists the results of the questionnaire. Additional written inputs and verbal comments are shown in the comments field for each pupil.

From the table, we can see that all female participants were nine-year-old, while the majority of the male participants were ten-year-old. Based on the open question "Do you like writing", we can see, that majority of the participants have a positive attitude towards writing. Noticeably most participants had difficulties in telling what kind of poems they like, many did not indicate anything, or picked up an adjective at random, when asked about it by the facilitator. This can be seen as an indicator of the participant's uncertainty of what a poem is. Also majority of the pupils had no previous experience with writing poems. All in all the participants were not very familiar with poems, nor with poem writing, while they remained mainly positive towards writing in general.

The table also shows participants' IT use preferences. As we can see, the majority of male pupils uses a computer quite often, while most females indicate using it less than once a week. All pupils use a computer at home, but unexpectedly, some of the male participants do not seem to count the use of the computer at school. Most of the participants also wanted to indicate specifically that they use the computer alone, while half of the pupils also admit to using the computer with friends. The teacher is considered as someone to use the computer with by three pupils, and three pupils use the computer with their parents, and again three pupils use the computer with siblings. The computer is mostly used for playing games, while only seven indicate using the Internet, and five doing school work with it. However, since many participants seemed to specify carefully, what they did on the Internet, it is possible that children have understood this question differently to adults. The tablet is used by eight users, while six have used a smart phone. Three pupils specifically talked about having a normal phone, without a touch screen. One pupil elaborately told about gaming consoles he owned. In a summary, all participants seem to have basic skills with the computer and can thus be seen as suitable participants for the study.

4.3.2 Materials

Each Peer Tutoring session required the attendance of at least two researchers: One for acting as the facilitator of the test, and one acting as the main observer for the test. The facilitator's responsibilities were similar to the responsibilities of the "interactor" described by Höysniemi et al. [14], but since

the word facilitator is more often used in literature, we have selected that term to be used in this thesis. The facilitator was responsible for fetching the participants from the classroom and introducing the participants to the test and the researchers. The facilitator also interviewed each pupil with the background questionnaire and conducted the post task interview after the test. During the test the facilitator was responsible for helping the pupils out with problem situations, and prompting them to tell their thoughts. The author of this thesis acted as the facilitator for all Peer Tutoring tests. The original purpose of the observers during the test was to fill in an observation form. The observation form was prepared to include all the features of the system, and the observer's responsibility was to evaluate during the test if these features were easy to use for the children. However the observation form turned out to be too cumbersome to fill out during the test, and it directed the observer's attention too much away from the comments and gestures of the users. Therefore the observation form was abandoned after the first test and the observers took free form notes instead to support the observations made later from a videotape. Two different observers participated in testing, observer 1 and observer 2. Observer 1 attended the first thee Peer Tutoring sessions, whereas observer 2 attended all six Peer Tutoring sessions.

The testing was conducted on laptops brought in by the researchers. Two Lenovo x301 laptops were reserved for testing, but only one was used in each session. The prototype was running on a Ubuntu 13.10, 64 bit operating system on an Intel Core 2 Duo processor with 3.8 GiB of memory and 121.8 GB of disk space. The laptops had a 13,3" display, and the resolution set to 1440x900 (16:10). However, since the additional display, showing a duplicated view of the screen for a camera was in letterbox format, the system had to be run in a letterbox setting for testing. To ensure all pupils used the same input mechanism, the touch-pads of the computers were disabled for testing, and a simple mouse with left, right, and middle scroll buttons was used instead. The mouses used were similar to the ones used by the pupils at school.

All of the sessions were videotaped with one camera. The test computer was set on top of two school desks, and an additional flat screen monitor was connected to the computer, so that the camera was able to capture both the users' expressions as well as what happened on the screen. This test set-up is shown in Figure 4.1, where the placement of the participants during the tests has been marked down. The pupils were positioned in user position 1 when they filled in the questionnaire, and in user position 2 during testing. The tutor of the test changed back to position 1 when tutoring the tutee. The facilitator moved around in three positions: During the background questionnaires, the facilitator sat in position 1, during the tutor training in

position 2, and finally during the peer tutoring in position 3. The layout in real life is depicted in Image 4.2.

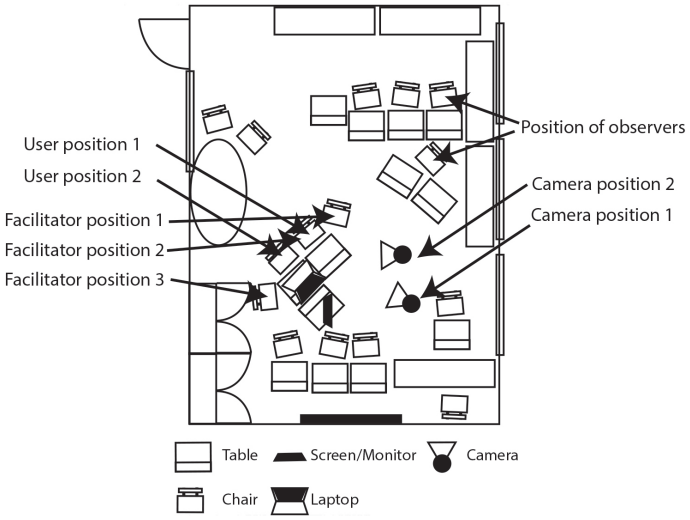

Figure 4.1: The set-up of the classroom during the Peer Tutoring Sessions.

As the prototype was run locally on each computer, the pupils did not have a way to save their poems for later. Therefore a selection of postcards was reserved for the children to write their poems in. The children received the postcards as a reward for their work, in addition to a sheet of stickers and candy. A sample of the reward materials can be seen in Image 4.3.

In addition to the background questionnaire, which was printed out for each participant, a post test interview was prepared and one sheet was printed for each session. The interview sheet can be found in appendix C. At the beginning of the interview the pupils were asked to assess how many stars they would like to give the tool. The purpose of the question is to ease the children into the interview and assess, if their overall experience was positive or negative. In next questions, the pupils were asked to name a specific feature they liked and one they did not like. More specific questions about the named features were asked to better assess them. After this, the facilitator asked more about specific problem situations or unused features according to what happened during the test. Finally the children were asked, if they would like to recommend the tool for a friend. Children were asked to consider another child, as it was considered this might reduce the bias children have towards pleasing adults. Additionally children were given a chance to say some free comments and express their desire to participate in the evaluation again. These questions were intended to evaluate the test set-

Figure 4.2: The set-up of the classroom during the Peer Tutoring Sessions.

up itself. The question for additional comments is very typical and a similar question has been asked for example by Donker and Reitsma [4].

4.3.3 Procedure

The Peer Tutoring tests were held on two days, three sessions each day. During the first day, the we arrived two hours before the first scheduled test to set up the equipment in the allocated classroom. This in itself was very fast, but the extra time allowed us to chat with the teachers for a while. Likewise we had ample time to participate in the morning assembly of the class to introduce ourselves and the project. We each gave our name and told the children how pleased we were to be there to test with them. I also reminded the children that I had already met most of them last spring, when I had visited their class. We emphasized the idea that we were there to recruit their help in order to enhance the tool for all children of their age. The teachers reminded the pupils that this was a special assignment and although everyone could not participate they could represent the class some other time on other special assignments.

Once we were ready to start the tests, the facilitator went to the class next door, and contacted the teacher to find the first tutor. All children in the class were used to reading a book, when they had finished a task, and the tutor was asked to bring this book with her to the test. Once the tutor entered the class, the facilitator first presented herself and the observers present. After this, the facilitator explained the testing procedure to the pupil in extent, asking if she needed to use the restroom and if she was ready

Figure 4.3: A selection of postcards, sticker sheets and candy used for rewarding the test participants.

to start. The introduction process was simplified after the first test, since it seemed to rather increase the anxiety of the pupil, instead of reducing it. Later we used phrases like "This is not an exam" and "Not even your teacher is going to see the video we take." to reduce the anxiety of the pupils. After the introduction the tutor was given the background questionnaire. The facilitator read the questions aloud, as well as the answers provided by the pupil, asking some additional questions, such as "How about last week? Did you use [the computer] a lot?". This was done the help the observers stay on track about the background of the pupil.

Once the background sheet was filled, the tutor training phase started. As the PM concept is supposed to promote creative writing, it was deemed impossible to give the user a very specific task to complete during the test. Instead, the users were simply asked to write a poem. A specific list of features was comprised before the testing began, and the facilitator and the observers were able to follow the list during the test to establish test coverage. If the user got stuck for too long a time, or indicated discomfort due to not knowing what to do next, the facilitator started giving hints on the features. Hints were also used, when a pupil announced the poem was ready at a very early phase, when very few features had been used. For example, some pupils left extra words on their screen after finishing the poem and the facilitator hinted at removing them, by using phrases like "Do you think you will still need these words?". This would usually get the pupils to find the unused feature quite fast on their own. All in all the tutor phase of the test came to resemble an informal walkthrough, which according to Riihiaho [32], is a

suitable method for testing when specific tasks are inappropriate.

The tutor training phase ended with the tutor selecting a card and writing the poem down on it for safekeeping. Due to the school schedule, the tutor training phase of the first session ended when the break had just begun, and as pupil 1 was very anxious at the time, she was encouraged to take a break herself, and bring her tutee back with her after the break. The anxiety of pupil 1 was entirely gone after the break. Likewise, the tutor training of session 5 ended during lunch break and it was deemed more humane to let pupil 9 to go have lunch and continue after.

The tutee introduction phase usually began with the facilitator telling the tutor that it was time to get the tutee, after which the tutor would teach the tutee how the program worked. The tutor was instructed to read the book he or she had brought from the class, while waiting for the tutee to enter and fill the background questionnaire. This worked well, except with pupil 7, who had forgotten his book and had problems staying on one place while waiting. The tutee was given a shorter introduction about the test, she/he was greeted by all the researchers and told that her/his friend had already tried the program out, and would now teach its use. The background questionnaire was filled in in a similar manner as with the tutor. Once it was finished, the facilitator asked the tutor and the tutee to change places. The tutee was instructed to start writing a poem, and the tutor encouraged to start giving instructions.

The peer tutoring phase usually went smoother than the tutor training. Tutors gave the tutee instructions, especially at the beginning. If the tutee seemed to be stuck, asking advice, the tutor was encouraged to answer. If the tutee clearly asked the facilitator instead of the tutor, the question was redirected to the tutor, saying for example "I am sure Jussi will be able to answer this.". If the tutee was stuck, but was not asking any advice, the tutor was prompted to participate by asking him a question, like "Jussi, what did you do, when you didn't know what to write?". This would usually result in the tutor telling about a specific feature, or telling about their writing process in general. If the help was not enough, the tutor was asked more specific questions, such as "Did the robot help you with this? Would you tell how it is used?". Once the tutee was finished, he or she also received a card to write the poem in.

Finally, the pair interview was conducted after the tutee had also finished writing down the poem in the card. The facilitator started by telling that she had a couple more questions about the tool. The facilitator would ask the questions, and give both children a chance to answer them. If one of the children seemed to be talking less, the facilitator would encourage them by asking them specifically, like "Well, what did you think about that Jussi?".

As van Kesteren et al. [39] suggested that watching important situations retrospectively from a videotape during post task interview may improve the information content of children's answers, we decided to use the prototype to illustrate some of the most important issues. The facilitator urged pupils to show the difficult situations, by asking questions such as "What was so difficult about the robot? Could you show me again?". This seemed to help to clarify the issues and remind the children about the context of the problems. Finally, after the interview, the children were thanked for their great help. As a thank you, they could choose a sticker sheet and a piece of candy. These rewards, especially the stickers were met with great pleasure.

All in all, six tests were held. On average the tests lasted for about 40 minutes each, with an additional 10 minutes it took to write down the poems. Specific test times are shown in Table 4.2. Noticeably, the most time was usually spent in tutor training, while the tutees were much faster to complete their poems. Likewise the interviews were usually very short, however, test 4 and test 5 both had an interview little over ten minutes long.

Session	Tutor introduction	Tutor training	Tutee introduction	Peer tutoring	Interview	Total
Session 1	3:55	20:06	02:57	9:35	4:25,	40:53
Session 2	4:21	18:57	2:48	11:22	7:42	45:10
Session 3	3:30	13:44	3:29	10:31	6:19	37:33
Session 4	2:51	8:37	4:23	5:28	4:34	25:53
Session 5	4:12	18:43	2:41	15:34	10:30	51:40
Session 6	3:21	8:42	3:22	12:57	10:30	38:52
Average	3:42	14:48	3:17	10:55	7.19	40:00

Table 4.2: Peer tutoring test times

4.3.4 Analysis Methods

Both qualitative and quantitative analysis was performed on the video tapes gathered from the Peer Tutoring tests. Qualitative analysis was aimed at gathering usability problems and notes about the test procedure. Quantitative analysis was based on observing the frequency of use, number of comments made during testing, and gestures of the test participants. The qualitative analysis is based on notes done by two researchers, whereas the quantitative analysis relies on one.

The qualitative analysis is based on notes made of the video tapes. Each tape was watched at least twice, once by one of the observers and once by the author of this thesis. Each made their own notes independently of the other. Notes were gathered in a spreadsheet. The analysts marked each note down on their own row, with the corresponding test phase and screen related to the

note. Additionally, the analysts could add free comments such as a suggestion on how to remedy a problem. The analysts were also asked to pre-classify each note into five groups: "Test technicalities", "Conceptual", "Usability", "Bug", or "Background". All notes were given a code, marking the test session and author of the comment, note T1_A1, would mean the first note of the first test made by the author of the thesis. Analysts were encouraged to make as many notes they liked, with as much detail as possible, preferably dividing complex ideas into multiple notes. Before starting the analysis, the author of the thesis made an example of the ten first minutes of the second test to show in practice how the notes should be made.

All in all, 1180 notes were made about the Peer Tutoring sessions. The author made 776 notes, while observer 1 made 148 and observer 2 256. On average 197 comments were made of each session. It was impossible to analyse so many notes with an affinity diagram and a digital qualitative analysis tool called Atlas.ti 7 was used instead. In Atlas.ti each note, called a quote, can be marked with multiple codes. Codes can then be used to build groups, in a similar manner to affinity diagrams. I started by putting in the notes made by the analysts as a pdf file. Each note was selected and turned into a quote. After this, each quote was renamed and encoded. I started by using the pre-defined classification groups as codes, and added other information, such as the writer of each note, test phase, and screen. Then I moved onto building relationships between notes – similar markings done from the same test by different analysts were connected. Gradually, more specific codes such as "Usability of robot" were added. Finally, affinity diagrams were build with the diagram tool of the Atlas.ti program, resulting in 82 unique usability problems. These problems can be seen in the appendix in Table E.

All problems were evaluated on a severity scale from one to three. Each grade of severity was given a description in order to correctly classify the problems. The descriptions of the severity grades used are as follows:

- Severity grade 3: A usability problem of this degree prevents the writing of poems, or the use of a feature. A conceptual misunderstanding of this degree may have serious consequences.

- Severity grade 2: A usability problem of this degree preventes the use of a feature, or seriously affects the efficiency of a feature, but the user is able to find a way to work around it. A conceptual misunderstanding of this degree seriously affects use.

- Severity grade 1: A usability problem of this degree affects efficiency, of bothers the user, but the user is able to work tolerably well despite

of it. A conceptual misunderstanding of this degree does not affect use considerably.

The given classification can be used to classify both classical usability errors as well as conceptual misunderstandings. This helps the classification of problems, which seemed to be the result of a conceptual misunderstanding or differences between the mental models of the user and the creator. Using this scale, 37 of the problems were identified to fit the description of severity grade 3, 25 grade 2, and 20 grade 1. The ability of a child to find a way to work around a problem was taken into account in the classification. Therefore two problems, which may seem as severe may be classified differently. For example, problem number 7: "Concept: The purpose of the robot is unclear" was classified in severity grade 3 as it prevents the use of the robot feature entirely, but problem number 39: "Creating a new word: The pupil does not know how to add a word" was classified in severity grade 2, as many users were able to divert the problem by replacing existing words.

In addition to assessing the severity of a porblem, a frequency was calculated for each problem. Problems were identified separately for the tutor training, peer tutoring, and interview phases. As such the maximum number of occurrences during testing is 18, while the maximum for each phase is six. The separation between phases was done, as it is interesting to see if the same problems were encountered by both tutors and tutees.

Since we did not use pre-defined tasks, a form was designed for observing the use of different features. For each test session, a form counting the number of uses was filled, for the tutor training and peer tutoring individually. I also marked down, if the specific occurrence had any problems. This analysis was done to support the qualitative findings and to evaluate how well the test covered each feature.

Since testing with children is very much dependent of the non-verbal signs of enjoyment or discomfort, I did a similar quantitative analysis for the gestures of the children. Again, a form was designed to be used once per tutor training and once per peer tutoring, to count the indices of different gestures during the testing. Finally the videos were also analysed for the comment types made during testing. Since it was deemed a futile task to transcribe the videos in a way that would make sense, re-occurring comments were quickly classified into suitable classes based on the first two tests, and a form was made to mark down the number of each comment type during each test. Only the interview phase was transcribed. The forms used for the quantitative analysis can be seen in appendix F.

4.4 Group Testing

The Group Testing method used in this evaluation was developed as an observational method, intended for evaluating the suitability of the PM concept in authentic classroom use. To gather feedback as a part of the session, we designed the Feedback game, a tool for leading group discussionw with children. The Feedback Game method has been previously reported in [18]. The Feedback game was inspired by the Talking Mats®[26], the Fun Toolkit [30] and the Focus Group methods. We also wanted to hear the teacher's thoughts on the usefulness of the concept and his views on how well the children were able to work with the system during the testing. Therefore the Group Testing was finished with a teacher interview.

The Group Testing had also five phases: During the first phase, introduction, the participating children filled in the same questionnaire as the peer tutoring participants, but working on their own. In the second phase, instruction by the teacher, the teacher introduced the PM prototype to the children, showing how to compose a simple poem. During the third phase, the poem writing, each child wrote a poem, which again, was written down on a postcard. The fourth phase, the Feedback Game, collected the children's impressions on the concept, while the fifth phase, the teacher interview, focused on the teacher's thoughts about the system.

4.4.1 Participants

Ten pupils participated in the Group Testing. Five of the pupils were male and five female. The testing was conducted in two sessions, with five pupils in each. The first session had three male pupils and two female pupils, the second session three female, and two male. The groups were formed by the teacher participating in the test in the teacher role. Table B.2 in the Appendix B shows the results of the background questionnaire.

From the table we can see that the group test participants also have mostly positive attitudes towards writing, with the exception of one participant, who stated he did not like writing. All participants, except for one came up with an adjective to describe what kind of poems they liked. Especially pupil number 16 mentioned liking poems that rhyme, indicating she knew more about poetry, than the question "Have you written any poems yourself?" would indicate. Three pupils had written poems before.

The Group Testing participants, who were not instructed during the background questionnaire, indicated considerably less specific things about their computer use preferences. Four pupils report using the computer 1-3 times a

week in the first session, while pupils in the second session use the computer a little less frequently. All participants reported using the computer at least at home, and half of the pupils considered use at school worth mentioning. Two children stated using the computer somewhere else, but only one of them elaborated it by stating she uses the computer at friends. Most of the pupils (9 out of 10) use the computer together with friends. Additionally four have mentioned they use it alone. Again pupil 14 has not specified, with whom he uses the computer in addition to friends, parents and siblings, although he has marked also the option "With someone else". Parents were additionally selected by two pupils, and siblings by one. Playing games with the computer is again popular, seven pupils mentioning it in their form. Six pupils use the computer for school while five use the Internet. All pupils indicated a history with tablet use, and seven had smartphone experience. A normal cellphone was used by three. Again, all participant seem fluent with the computer, and thus are suitable testers.

4.4.2 Materials

Each Group Testing session was attended by three researchers: the author of this thesis and two additional observers. The author acted again as a facilitator, introducing the test and the background questionnaire to the children at the beginning of the session. Additionally she served to help the teacher if he faced a problem he could not work out on his own. The first observer used the camera during the test, moving the camera around in the classroom, looking for the most interesting interaction. The second observer wrote down free form notes. The first observer changed positions with the author and took the role of the facilitator in the middle of the second test, as the author was losing her voice.

Six laptops were reserved for the testing and the author's personal laptop served as a spare. The two Lenovo laptops used in the Peer Tutoring were used also in the Group Testing. Additionally four Dell Latitude E4200 laptops were brought in for the test. The Dell laptops were running the same Ubuntu 13.10 64-bit operating system. The memory capacity of the Dell computers was slightly better, 4.8 GiB in comparison to the 3.8 GiB of the Lenovo laptops. The processor was the same, Intel Core 2 Duo processor, and the Dell computers also had 120.7 GB Disk space. The screen of the Dell laptops was slightly smaller, allowing only for a resolution of 1280x800. The same prototype was installed on all computers, however one bug was fixed on the basis of the Peer Tutoring tests.

The test was set up in the same room, as the Peer Tutoring tests. This time, the room was arranged to suit traditional teaching, with the exception

that the computers were arranged into groups of three and two (The two Lenovo computers formed the pair, while three of the Dell computers were used in the group). The computers were grouped in order to prompt talking between students. This was also similar to the arrangements normally used by the class, as observed the year before, when pupils were usually working in pairs when working with a computer. One of the Dell computers was given for the teacher to plug it in to the smartboard. This test set-up is shown in figure 4.4. A photograph taken of the test set-up before the first test can be seen in 4.5.

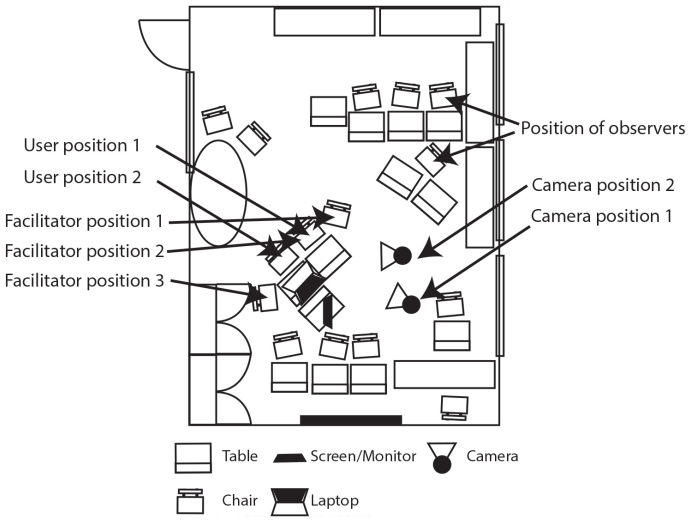

Figure 4.4: The set-up of the classroom during the Group Testing Sessions.

Additional supplies were needed for the Feedback Game. It is played around a physical gaming board. The board consists of eight questions and five corresponding smiley tiles each. The smiley tiles form a likert scale ranging from completely disagree - a very sad face - to completely agree - a laughing face. The scale resembles the Smileyometer, a part of the Fun Toolkit developed by Read et al. [30], while the table layout of the board has been inspired by that of the Talking Mats®[26] and the Again- Again Table [30]. The questions were based on the questions used in the pair interviews held after the Peer Tutoring. Some questions were omitted, and some added based on the challenges in the Peer Tutoring tests. All questions were reformatted into likert-scale suitable statements for the board. The gaming board used is shown in Image 4.6.

The questions on the board are:

- "Was it fun to use the poetry tool?"

Figure 4.5: The set-up of the classroom during the Group Testing Sessions.

- "Was it easy to use the poetry tool?'
- "Was the purpose of the Poetry Machine clear?'
- "Were the poem beginnings suggested by the Poetry Machine good?"
- "Were the words suggested by the robot good?"
- "Would you be able to write a poem without the Poetry Machine? "
- "Would you recommend the Poetry Machine to a friend?"
- "Would you like to participate in this test again?"

In addition to the gaming board, each player of the game has their own set of tokens. Physical tokens were selected as the experiences of Read et al. with the Fun Sorter [30], Murphy and Cameron with the Talking Mats Talking Mats®[26], and Hanna et al. with card sorting tasks [10] support the use of physically manipulatable items as a survey aid. The game tokens were wooden cube shaped crafting beads. There were eight colours available and each colour had ten to twelve pieces (only eight were needed for the game). Tokens were held in small resealable plastic bags. The game tokens can be seen in the same Image (4.6) with the gaming board. Similar postcard and sticker reward system was also used in the Group Testing.

Additionally an interview was prepared for the teacher. The interview sheet can be seen in appendix D. In the questions, the teacher is asked to review, how similar the test situation itself was to an average group lesson,

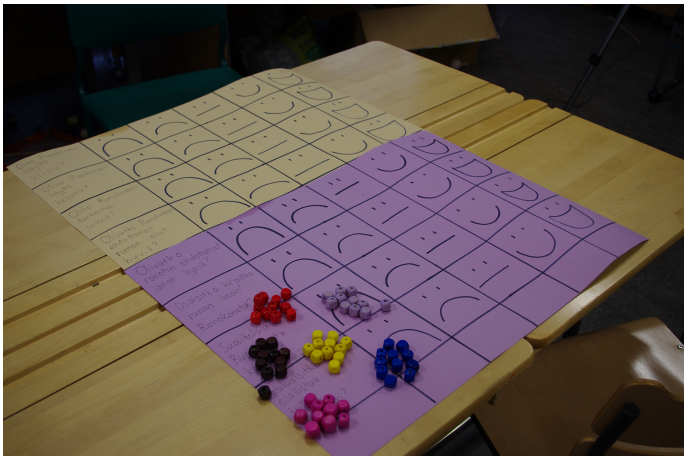

Figure 4.6: The feedback game game board and game tokens.

what kind of ideas he has for supporting the students' work better and what kind of feedback he would like to give about the test itself. He was also asked what kind of informal ideas he had heard from the participants of the previous tests and if he thought the children behaved differently to their usual self. As such, the teacher interview was designed to gather insight into how the children behaved during the test lesson opposed to normal lessons, and act as a validator for the test session itself.

4.4.3 Procedure

Two Group Testing sessions were held on the same day. We arrived early, to set up the environment and go over the last minute details with the teacher. We started the first session right after the morning assembly. When children entered the class they were encouraged to choose any one of the computers set up for testing. The children chose to sit with like gender, and each found a place without problems. The first test took 52 minutes while the second took an hour and two minutes. Times of specific parts of the test can be seen in Table 4.3.

Test	Introduction	Instruction	Poem writing	Card writing	Feedback Game	Total
Group 1	4:58	7:32	14:15	10:25	14:52	52:02
Group 2	4:04	8:11	19:21	14:43	15:52	1:02:11

Table 4.3: Group testing activity times

The facilitator first re-introduced the researchers to the participants and

then distributed the background questionnaire. This time, the questions were not read out aloud, nor pupil specific help given during the filling. The facilitator reminded the pupils they could write down a more suitable answer, if they felt that the possibilities given were not appropriate. She also answered a couple of questions about the correct format of writing down one's name and such practicalities. The children were asked to put down their pens, once ready, and once everyone had finished, the facilitator collected the forms and encouraged the teacher to start.

In the first session, the teacher started the instruction right away. He was a little bit nervous about not having used the prototype himself overly much at first, but regarded the assignment positively. In the second session, the teacher first asked the children about poetry, and only then started the actual lesson. In both sessions, the teacher went through the creation of one poem, after which he ushered the children to start writing one of their own.

After the initial instructions, the teacher left his computer to circle among the pupils, giving help when needed and praising the children for their poems. The children went about the task, occasionally stopping to watch what a friend was doing, or to ask the teacher help with something. If multiple children were having the same problem, the teacher returned to his computer to show the solution to everyone. Some pupils were very eager to comment on their writing, especially pupil 14 in the first group and pupil 18 in the second. All in all, the first group performed very well, without any serious errors. The second group had some more problems, especially with the prototype freezing at times. However all managed to write a poem. After the children had finished their poems, they were given a selection of post cards to choose a card to write their poem in.

Once the writing was over, the pupils were asked to assemble around the game board, which had been set up on top of tables pushed together. The Feedback Game session started with the participating pupils having their choice of colour from the game tokens. Once finished, the facilitator – called during the game a game master– explained the rules of the game: The game master asks the next question on the board and all players may answer it by placing their tokens freely on the most suitable smiley face. After this a round of further questions will follow on the subject and each player may answer it on their own turn. After each turn, the used tokens would be left on the board while the group proceeded to the next question.

During the game, the game master addressed the children through their chosen colours. The game master would for example ask for the arguments of "the blue player". To avoid biasing the arguments of the participants, each new game round was shifted to begin from the second player of the previous round. For each round of questioning one question was prepared

for all participants based on the board question. For example we used the following question pair: "Was it fun to use the poetry machine?" - "What was so fun/boring about it?" The game master was allowed to add further questions for pupils, who expressed particularly interesting arguments, or to clarify the arguments of the pupils. A feedback game in progress can be seen in Image 4.7.

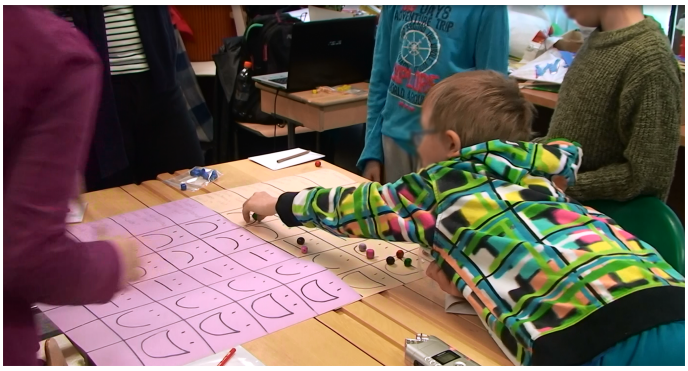

Figure 4.7: The feedback game in progress.

The teacher was interviewed twice with the same questions. Therefore the second interview was less formal than the first one. The first interview was administered by the author and the second by the acting facilitator, with the author asking one additional question, as the teacher showed interest in the Feedback Game as a educational method.

4.4.4 Analysis Methods

The Group Testing was less rigorously analysed, due to the fact that the video record was of a more general nature. Again, the author made notes of the video, in a similar manner to that of the Peer Tutoring videos. This time, the videos were only watched by one person, as others were unable to participate in this action. The head researcher made 201 notes about the first session and 290 about the second session. In addition the Feedback Game was transcribed.

This time, the number of notes made was more manageable. An affinity diagram was built of the notes, which were printed out. In order to see, if any of the problems found in the Peer Tutoring were seen in the Group Testing, a form was used to count indices of problems per pupil, based on the problems faced in the Peer Tutoring. The form can be seen in Appendix F.

Chapter 5

Assessment of the Usability Testing Methods

This chapter discusses the methods used in this study, their reliability, coverage, and more specifically possible error sources in their application. In this section I first review how well the selected methods performed against goals set for the usability testing. Example data is given to illustrate what can be learned with these methods. Next I move on to evaluate the methods against more general criteria, usually considered for usability evaluation methods. These criteria include the reliability of the methods, their test coverage and the quality of used analysis methods. I then consider the suitability of these methods to usability testing with children. Finally I discuss more specific issues related to using these methods in evaluating the Poetry Machine.

5.1 Method Performance in Collecting Problems, Eliciting Feedback, and Testing Usefulness and Enjoyability

The goals for the usability evaluation were to collect usability problems and evaluating the usefulness and enjoyability of the concept. Additionally I wanted to collect user impressions and ideas. I managed to find interesting use-patterns to illustrate the use in practise and some participants came up with new ideas.

The testing uncovered 82 unique usability problems in the prototype. The problems are described in detail in Table E.3 of Appendix E. The number of identified problems is high, but many of them seem to be related to each

other and can be corrected by using similar means. For example, issues with moving words are described by eight problems (problems 1,3, 4, 13, 19, 28, 54, and 56). The Peer Tutoring also produced a frequency rating for each problem: For example, problem 1 was identified with all tutors in training and four of the tutees implicating its importance. In addition to traditional usability problems, some problems were identified as conceptual, implicating the idea that users were not understanding the concept behind a feature, such as publishing a poem (problem 8). These conceptual findings were uncovered by dialogue between the users or between the users and the facilitator. Finally, two problems (problems 81 and 82) came directly from the Peer Tutoring participants, who expressed a desire for shortcuts for moving all or extra words. The Group Testing was useful for understanding the severity of some problems, such as problems with buttons and their identification better.

Enjoyability was measured by evaluating the observed and experienced fun during testing. Observation forms were used to analyse the number of positive and negative gestures during Peer Tutoring. Figure 5.1 shows the total number of negative an positive gestures during Peer Tutoring per pupil. In general it seems that pupils make less gestures when they are tutoring another, than when they are working on their own. Only four pupils made more positive than negative gestures. However, it is important to note how behavioural signs were classified: Since Hanna et al. [9] consider frowning a negative sign, frowning has here been considered negative. Yet in many cases, frowning seemed to be rather an indicator of concentration, which has been interpreted as a positive sign by Read et al. [30]. Hanna et al. also indicated later that children make in general less positive gestures during testing than negative [10]. Therefore it seems overly hasty to conclude that children were displeased with the program. Also, most of the comments made by children during testing were negative as seen in Tables F.3 and F.4 of Appendix F. However many of the comments were related to not understanding what the prototype was for, which can be interpreted more as a problem with the test task. Less comments were made during Group Testing. The teacher confirmed that the number of help requests was actually smaller than in his average lessons.

The feedback from the pair interviews and the Feedback Game is in contrast to the observed fun. All peer tutoring participants gave great scores for the prototype when asked to rate it during the interview. Pupils 1, 2, 3, 4, and 8 gave the tool five stars out of five. The rest gave four stars out of five. Notably, pupil number 10 originally gave three stars, but changed his rating to four after trying out the robot tool during the interview. The rating was clearly linked to fun in children's minds, as the main argumentation for the

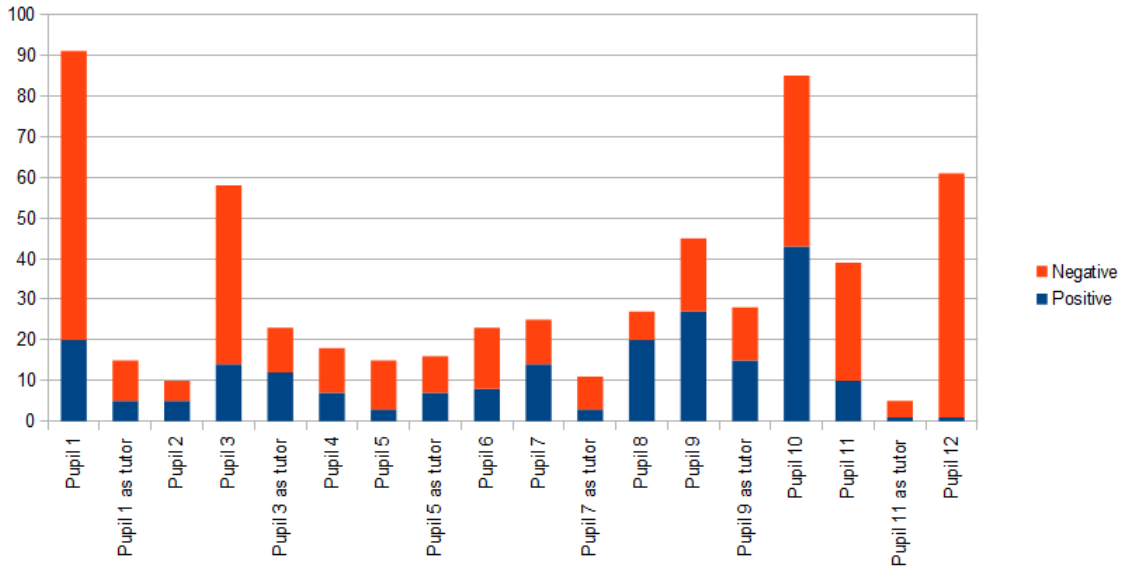

Figure 5.1: Total number of gestures made by pupils during peer testing. Gestures made by tutors are listed separately for tutor training and peer tutoring phases.

scores by pupils 3, 4, 5, 6 and 8 was that they had fun. In addition students 1, and 4 mentioned having fun as the main reason for recommending the tool to friends. Also group testing participants were of the opinion that using the Poetry Machine was fun: They all put their tokens enthusiastically to the fully laughing face when asked about it during the feedback game. Pupils 13, 15, 18 and 21 also told they would have liked to participate in the test again, as writing poems was so much fun. The evidence for fun was further backed by anecdotes from the teacher, who told one of the group testing participants had said him that she had been waiting to participate in the test.

The PM concept seemed to be useful for writing poems: Many users also mentioned either during the peer tutoring interview, or the feedback game that they enjoyed writing poems with the tool. Some users also considered the tool helpful to writing poems. Pupils 1, 2, 3, and 4 thought writing was the best thing about the tool. Pupils 1, 2 and 7 would have recommended the tool especially because writing with it was fun and pupils 2, 5, 6,12, 13, 18 and 21 would have liked to participate in the test again especially because writing poems was so much fun. Pupils 2, 8, 11, 13, 14 and 16 also considered that writing poems with the tool was easier than writing otherwise. Especially pupil 16 mentioned that the existing words helped

him. The teacher noted also that it seemed pupils were more faster to get to work, and they also were more engaged with the program. For example, he mentioned that pupil 14 usually had difficulties with coming up with ideas, but now he was working very autonomously throughout the session. Later on, he also wrote in an e-mail that one of the pupils participating in the test had been so enthusiastic about it that she had been inspired to take poem writing as a hobby.

No formal evaluation of the educational value of the tool was made, and children were not asked to evaluate the learning capability. However, pupils 4, 9, 13, 14, 19, 20 and 22 stated they would recommend the tool, because they saw it good for learning. Especially pupils 9, 10, 19 and 22 mentioned younger siblings or friends as a suitable target group. Pupils 5, 18, 19 also autonomously told they had themselves learned to write poems with the tool. The teacher saw possible use scenarios for the tool in his future lessons.

In addition, observing the use of different functions was useful for analysing different use patterns. For example, instead of adding words, users seemed to replace existing words, and some users used another row as a storage unit for words they did not want to use right away. The robot was also used less than initially expected, but pupil 9 was clearly using a strategy, in which he always tried the robot first to get more ideas before adding new words.

The interview and the Feedback Game were also fairly good for gathering improvement ideas for the concept. Ideas can be seen in Tables E.1 and E.2. Users asked for more relevant and familiar words, better rhyme suggestions, and shortcuts for removing words and adding punctuation. Finally pupil 10 suggested a change to how the word options were displayed in the prototype. He suggested displaying alternatives under each word, when a word was clicked. Notably, ideas gathered in the Feedback Game seem to deal more with the material, while ideas gathered during Peer Tutoring deal with the interface.

5.2 Reliability of the Chosen Usability Test Methods

It is very difficult, if not impossible to evaluate, if all usability errors existing in the program were found through the tests conducted, and if there are false positives in the set of found errors, as there is no standard usability problem set to compare the used methods to. Hartson et al. [11] suggest that when evaluating two methods against each other, a standard problem set could be produced through known usability problems, laboratory testing, asymptotic

laboratory testing or as a union of the compared methods.

In our case, the system has not been tested previously and therefore the only option would be to combine usability problems found in Peer Tutoring with problems found in Group Testing. However, we were not able to analyse the group testing as rigorously as the Peer Tutoring, since we did not have a method for recording each participant's computer carefully enough. We also could not capture pupil expressions during the test. Yet it seems that the group testing participants faced only a fraction of the problems faced by Peer Tutoring participants. This would suggest that Peer Tutoring is better for recording usability problems, or it records a large number of false positives. However the Group Testing participants were likely to face less errors because of the formal instruction they received for the program. As intended the two methods test for different scenarios: While Peer Tutoring tested uninstructed use and peer instructed use, the Group Testing tested formally instructed use. Therefore, it is not recommendable to compare errors found by tutors in training, tutees, or group test participants with each other. Still by looking at a specific subgroup of the errors, specific type of use can be improved.

5.3 Test Coverage

Three factors are relevant for determining the test coverage: number of participants, representativeness of the participants, and task coverage.

The number of test participants needed to cover a prominent number of usability errors in an application is a well studied issue in usability testing. The original statistical model was developed by Lewis, based on binomial probability and the probability of discovering one problem [21]. According to Lindgaard and Chattratichart [23] the two most well used formulas today are Virzi's model, based on the Monte Carlo procedure and Nielsen and Landauer's formula based on the Poisson distribution. Both formulas assert that 5 users will be enough to reveal 80% (Virzi) or 85% (Nielsen and Landauer) of usability problems in any given application [23]. Egloff [6] has used the formula of Nielsen and Landauer for children's usability testing and determined her pool of 15 users was more than ample to provide a general feel for the ease of use and user preference. Our pool of 12 users or 6 pairs is also over the threshold of five. However these formulae may not be applicable for paired testing scenarios. It is uncertain, if paired testers are more or less efficient in finding problems and different studies have found contradictory evidence [17]. Additionally, Lewis [21] notes that since the probability of discovering a problem changes by context, these results are not applicable to all

situations. For example Lindegaard and Chattaratichart [23] did not find a correlation between the number of testers and problems in their experiment.

The variability of users seems to be relevant to covering usability problems in user testing [23]. It is an obvious downfall that the test was conducted only with subjects from one class. Additional tests with a fourth class were designed and booked, but regretfully we were unable to complete those tasks, as the whole research team fell ill, and no more free times were available in the schedule of the children. However, many studies in the field of children's usability evaluation have been conducted with very narrowly selected samples. For exampe Sim et al. [34, 35], Costabile et al. [2] and Höysniemi et al. [14], [15] were content with recruiting children in one school. In addition, multiple studies, including [30], [31] and [24] do not state where the users were recruited.

Likewise, the selection of test subjects for testing must be reconsidered, especially for tutors. It seemed that Peer Tutoring worked better, if less shy partners were in the tutor role, so it might be beneficial in the future to ask the teacher, or other party responsible for forming pairs, to note which of the children is less shy. Otherwise there were no notable problems with the pairs participating in Peer Tutoring tests. However the second group participating in the group testing had two pupils, who once commented on each other negatively during the feedback game.

In the study by Lindegaard and Chattaratichart [23], tasks defined for usability testing were found to be more nominal for finding a high number of usability errors, than the number of test participants. As the Peer Tutoring task was very general – to write a poem – it considerably resembles an informal walkthrough. Therefore we are not able to determine the coverages of the tasks by task description, instead we have to look at the actual usage of features during the tests: As Table F.1 shows, some features, such as theme selection were not used by all pupils. Also, two features, moving rows and changing theme were not used by any of the pupils. Therefore the test does not represent problems with all features in a unified way. However, many features are very similar to each other. There is reason to believe, that moving rows would have similar problems to removing rows, which displayed some problems with accurately grabbing rows. Likewise it is likely to have similar problems to moving words, in that the targeting works similarly with rows. Changing the theme then again works similarly to changing the name or author. Therefore it can be said that no feature was completely left without notice.

5.4 Quality of Analysis

The integrity of usability evaluation does not depend only on the number of users, used materials, or methods, but the evaluators used to analyse the data. Hertzum and Jacobsen [12] studied what they call the 'evaluator effect' in their article. Their conclusion was that independent from the method used to obtain the results, all usability evaluation methods are subject to evaluator effect, which means each evaluator will come to different conclusions based on the data. Although strict goal analysis, evaluation procedures and problem criteria can help to alleviate the evaluator effect, the only solution for reducing it significantly is to apply multiple evaluators instead of one [12].

To reduce the evaluator effect, two independent observers took notes of each video record of the Peer Tutoring sessions. However the collection of errors and comprising a list of them was an individual exercise done by the author of this thesis. Therefore it is quite likely that a number of false positives or misses of actual problems have happened during the analysis. However, using two observers for taking notes of the Peer Tutoring sessions was still beneficial: As seen in Table E.4 of the Appendix E, there are 18 problems which were only recorded by one observer. 12 of these notes were made by the author, 4 by observer 2 and 2 by observer 3. Therefore it is clear that six errors would have gone unnoticed had the author performed all of the observing alone. However, it is impossible to judge the accuracy of findings between observers, since there was no possibility to discuss the findings further among team members.

The reliability of the quantified results is more questionable, since only the author recorded observations with the observation forms. Although the use of forms should reduce observer bias, it is highly likely that these records suffer from what is called a 'coder-drift', the tendency of a single observer to categorise information instances differently across different examples [1]. To evaluate the amount of coder-drift in this study, the analysis of the first video was repeated. For the feature usage, the differences between markings for one column was at its worst 7 markings (the modifying words column). However the biggest percentual difference was in the column for robot markings, which was found to have 4 markings less in the second observation round. Differences in gestures and comments were even larger, with similar gestures or comments coded differently. Therefore additional codes for gestures and comments were merged with similar notes to reduce the error rate. For example, all instances of a user covering his or her mouth were noted as instances of the user covering his or her face, to reduce possible miscategorisation of the face-covering instances.

It is remarkable, how difficult it was to achieve consistent coding for gestures and comments: I took screenshots of the pupils' gestures to get references to use for coding other gestures, but this did not help to avert the issue. During the first run, I counted instances of knitting one's face, but on the second run, I coded the same gestures as knitting one's brows. Similar problems were faced with comments. Comments were coded more specifically towards the end. Also some comments were not collected at all for the first test. However, already the second test was much more closely coded.

The validity of the gesture and comment observation is therefore questionable. But it is important to note that as the Peer Tutoring method is formative in nature, meaning, it is done during the development of a tool to find usability problems to improve the design of the interface, the purpose of this quantitative data is not to be statistically significant but to guide the design [11]. Therefore we consider that a higher error tolerance for this activity is allowed and general results, such as the number of positive vs. negative gestures can be seen valid enough to estimate the general feel of the situation. However, the usefulness of the results is also in question: The results are in strong disagreement with the interview results and later results from the Group Testing. And children seemed to have discussed the testing positively with each other outside the official sessions, as participants grew more eager to participate throughout the testing.

5.5 Suitability of the Selected Methods for Usability Testing with Children

All children were willing to participate again, either in the Peer Tutoring or the Group Testing. Most children wanted to participate in the Peer Tutoring test again, because of fun had during the test and with the creative writing task. Children wanted to participate in Group Testing again because of the fun had during the test with computer, working with friends, and the card received as rewards. In the Feedback Game, the enthusiasm for participating again was clearly shown by how all participants decided to put the rest of their tokens on the board to show how much they liked the test.

Therefore, we may conclude the methods were not unpleasant for the children. Special admissions were made to alleviate any discomfort children were having during the testing by giving two of the test participants in the Peer Tutoring a chance to take a break during testing. Since both participants came willingly back with their partner, they did not consider the activity

overly unpleasant. The Peer Tutoring, pair interviews and the Feedback Game were all good for eliciting comments from the children. In addition, the Peer Tutoring was good at covering problems children had with the interface. Therefore we may assume that all methods used were suitable for testing with children.

5.6 Lessons Learned from Peer Tutoring

This section discusses practical issues in conducting the Peer Tutoring sessions, as well as possible error sources and strategies for averting them.

5.6.1 Issues with the Test Set-up and Materials

The background questionnaire clearly would have needed more testing than just one pilot tester. Many pupils added choices to the sheet, indicating that the alternatives would have needed more work. Even though it distracted some children, it was important to interview the children while they filled the questionnaire, as this revealed which questions worked and which not. For example, many pupils did not seem to understand the question about what kind of poems they liked and seemed to pick up an adjective used by the facilitator to elaborate the question. This behaviour indicated clearly that most pupils did not know much about poetry. Therefore this question was unsuited for the Group Testing, as it was not possible to interview each pupil individually. In addition I noticed we should have asked the native languages of the pupils, since one mentioned being bilingual only when she told she liked poems in her other native language. Likewise it seemed many pupils did not think computer use at school to be relevant, and it would have been beneficial to ask about it separately or from the teacher. The interviewing also seemed to help the facilitator to make some small talk to break the ice.

In the test itself, it became apparent that most pupils liked to browse the system a little at first, while others were too shy to do it. It would have been good to include free browsing in the test like done by Hanna et al. [8]. It might also have been a good idea to use some sort of a warm up task to show the children that we were indeed not trying to evaluate them, but the system. The task of writing a poem seemed to be a very sensitive one. Researcher 2 wrote to her notes that it seemed that pupils were trying to write as good poems as possible, and suggested this might be the reason why pupils 2 and 4 seemed to be writing an existing poem from memory. Additionally pupil 10 expressed that it seemed unfair pupil 9 had seen his

poem, while he had not seen the poem written by pupil 9. The task was also too general. As noted by Markopoulos and Bekker [25], children may have difficulties understanding such abstract tasks. However, with enough prompting and using more explicit language to describe the interface, when children showed they had no idea what to do, the process started to work.

In the physical set-up of the test, it would also have been better to place the camera a little differently. Now, the pupil working on the computer was partly hidden behind the screen hindering the observation of gestures. Likewise some comments, spoken with a shy voice may have been better understood with the video reference. The camera was also clearly bothering some of the users, especially when an old cassette started to make scratching sounds during one test.

The test sessions were ocasionally disturbed by other activities in the classroom. Sometimes teachers had to drop in to get something, which considerably bothered the users. The worst disturbance to testing was, when the headmaster popped in during the fifth session. He was very enthusiastic to see how we were doing, but the participants of the test suddenly grew silent and it was hard to get them back into the task after the headmaster had left.

The prototype had a considerable amount of bugs. The bugs have been described in Table E.5 of Appendix E. Notably, the first bug, problems with updating, was faced in all Peer Tutoring tests. More rigorous testing under prolonged conditions would have been needed before the user testing to correct all of the bugs. We would also have benefited from setting up an automatic system for recognizing the use of features in the prototype. This would have saved time from observing the amount of feature use during the test.

The rewards given to children for participating in the tests were met with enthusiasm. Children were seemingly happy about the stickers, as well as the cards. The cards were a bit difficult for the children to handle, as some of them were unfamiliar with envelopes, having lived their lives in the era of e-mail. In addition to the practical issue with the envelopes, the cards may also have caused a bit of bias, since they were given before the feedback sessions of the Peer Tutoring and Group Testing: Later on, in the Feedback Game, one pupil told he liked the test most because of the fancy cards.

5.6.2 Conduct of Test Personnel

As can be seen in Table F.3 of Appendix F, the facilitator had to make a lot of comments during the tutor training phase of the Peer Tutoring tests. To alleviate the problems with the abstract task, the facilitator had to explain

the tool more than originally expected, by saying that the system is giving the users examples to work on. Different support strategies were used with different pupils, supporting the idea of van Kersten et al. [39] that children need different prompting techniques dependent of their character. At times the facilitator seemed to have given a little too much information too early, when pupils might have just needed to think for a while. However, if the facilitator had not directed the test forward, the time taken with one pair would have become overly cumbersome for the children. Related to this, the facilitator needed to help some pupils to get an idea for a poem, to get them started on writing. Despite using the question asking protocol as described in section 4.3.3, some pupils seemed to have grown overly dependent on the facilitator during the test. For example, pupil 1 asked the facilitator, if she needs to use punctuation marks. Luckily, the peer tutoring phase worked better, with the facilitator needing to give much less directions as shown in Table F.4.

We noticed that children were bothered whenever the facilitator wrote notes. Therefore we recommend giving the note taking responsibility entirely over to observers. However some notes were needed to keep track of problems that were selected for further discussion in the interview. The facilitator also wrote down the names of the children after having embarrassing problems remembering the names of the pupils of one test. It might also be good practise to give participants name tags to help making personal contact with them.

Also, we have to note that three adults were present during the first three sessions of Peer Tutoring, which may have put over-much pressure to the first tutors during tutor training. Therefore we recommend only one observer in addition to the facilitator.

5.6.3 Conduct of Test Participants

Throughout the peer tutoring phase, tutors used five techniques for guiding their tutees. The first technique was directly pointing at objects on the screen and describing what to do next. The second technique was to try and describe what can be done with the interface, or what they had done themselves in a certain situation. The third technique was showing physically what to do next. However, since the tutee was placed on the right side of the tutor, blocking access to the mouse, the tutors only used this technique when explaining how to move the cursor with the arrow buttons. The fourth way of instructing, was to answer specific questions asked by the tutee. Tutors themselves usually asked their tutees only, if they were ready. The fifth way was to generally ask the tutee to try things out. This was only used by pupil

1, who clearly was using the technique similarly to the facilitator.

Although most tutors were quite happy to guide their friends, some of the pupils had problems with the tutor and tutee roles assigned to them: First of all, some tutees tried to ask the facilitator first, when faced with a problem. Secondly, some tutors waited for the facilitator to give them permission to answer the tutee. Thirdly, at times tutors were bored, and unable to understand what kind of help their friends needed. For example pupil 7 ignored some questions from pupil 8, until the facilitator prompted him to answer. Then he often explained something that was irrelevant to the task the tutee was performing. Finally, pupil 3 was seemingly shy, and first seemed to dislike the idea of tutoring her friend. All of the tutees seemed to react positively to the test. Most of them seemed more confident at the beginning of the test and started on the task faster than their tutors.

5.6.4 Thinking Aloud

The pupils participating in the Peer Tutoring test were able to think aloud to different degrees. Some were mostly silent, asked a lot of questions, or complained about things, but a couple of students were more descriptive. For example, pupil 7, when clicking words on the screen initially tried to write on a word, but then grabbed the word instead, at the same time describing what he was finding out: "It's not budging... Whoa... you can move it!" ("Ei toi ota tosta... Oo... tota voi siirrellä!").

During the tutor training phase, most pupils were able to express some of their problems, once asked by the facilitator. For example Pupil 9 had problems coming up with an ending for his poem. The facilitator asked "What are you thinking about now?" ("Mitä sä nyt ajattelet?") Pupil 9 initially answered "I don't know." ("Emmä tiiä."). But when the facilitator asked "Is it hard to think of an ending?" ("Onks vaikee keksii loppuu?") the pupil confirmed this by saying "I don't really know what to do with the ending then..." ("Ei tiiä mitä tekee tohon loppuun sitten...").

During the peer tutoring phase, the facilitator did not want to disturb the tutee, but prompted the tutors instead. After having some tutors autonomously start describing how they had written their own poems, we started to use it as a prompting technique. This gave us a unique view to see how children perceived the program. However, situations in which tutors gave wrong instructions, or surprisingly instructed the tutee correctly in a task they had not tried out for themselves were also valuable in pointing out mechanisms behind problems with certain features and their discoverability.

Facilitator: "So nothing was boring? That's pretty surprising. Well, was there something you didn't think worked that well or...?"
Pupil 3: "I donno.... No..."
Facilitator: "You can say it, if there is something, just say it bravely."
Pupil 3: "Well maybe that you had to first remove the words..."

Ohjaaja: "Ei ollu mitään mikä oli tylsää? Aika yllättävää. Mitäs... Oliks siin jotain mikä ei teidän mielestä toiminu kauheen hyvin tai...?"
Oppilas 3: "Emmätä... ei"
Ohjaaja: "Sano ihan rohkeesti, jos on joku."
Oppilas 3: "Tai no ehkä se, kun ensin piti poistaa ne sanat..."

Figure 5.2: An excerpt from the discussion between pupil 3 and the facilitator during a pair interview

5.6.5 Pair Interview Bias

Observer 2 noted that the pairs participating in the first three Peer Tutoring test gave exactly the same amount of stars to the program. She also notes that pupils 2 and 4 seemed to imitate their partners in argumenting for the star rating. Also, pupils 7 and 10 were clearly influenced by their friends when they considered to whom they would recommend the tool. The bias may also be stronger with younger children, as the only two pupils, pupils 7 and 9, who rated the program differently to their partners, were already 10-years-old.

At times users were also clearly affected by the facilitator's choice of topics: It seemed that during session 4 pupils 7 and 8 did not use the robot tool at all. The facilitator asked during the interview, if they had had difficulties with it. Neither pupil could say why they did not use it, or what they would have used it for, had they known about it. Yet, after trying it out during the interview, pupil 8 said later on it was one of the best features. Therefore it cannot be said if the robot was only mentioned because the pupils wanted to please the facilitator, or just because they actually liked the robot. Later, a similar case happened during session 5. However pupil 10 was so carried away by the robot it was clear his enthusiasm was not faked; He wanted to carry on playing with the robot until the facilitator had to remove the mouse from him, in order to have him concentrate on the task.

The facilitator had to balance the discussions dominated by either the tutee or the tutor: In session 2, the tutor, pupil 3, was too shy to give

her opinions, but making eye-contact with her and encouraging her made her state one of the most interesting things about the system as illustrated by the discussion in Figure 5.2. In session 5, pupil 10 was so incredibly talkative that it was almost impossible to give pupil 9 enough space to give his opinions. All in all, it helped to ask pupils personally for their opinion, if they seemed to be left out of the discussion.

5.7 Lessons Learned from Group Testing

This section discusses practical issues in conducting the Group Testing sessions, as well as possible error sources and strategies for averting them.

5.7.1 Test Set-up, Materials, and Conduct of Test Personnel and Participants

The feedback form had similar issues to those uncovered during Peer Tutoring. In addition one pupil forgot to write down any additional options, even though pupils were reminded of it during the task.

Pupils were very enthusiastic throughout the Group Testing sessions. They also stayed very concentrated during the Feedback Game. Having the discussion in game form seemed to help pupils concentrate and understand that everyone had their own turn. While the facilitators sometimes forgot to ask questions from all pupils, the pupils were quick to remind the them about it. The rule about leaving tokens in their place was a little difficult for some pupils, who were keen on moving the buttons already placed on the board. Also, one pupil looked very bored during the end of the game, playing with his tokens. To encourage all participants, the facilitator commented enthusiastically all pupils for their answers during the game.

We could have used additional cameras or screen capture for analysing the results of the test. During the Feedback Game, the camera could have been positioned in a way that showed the board more clearly. However, we managed to stay quite unobtrusive throughout the test, giving the teacher and students a chance to work undisturbed. We had a couple of similar distractions as with the Peer Tutoring, but they were less obtrusive as the users were more used to teachers moving around when teaching was going on.

Some of the questions of the Feedback Game were misunderstood by some pupils. For example question 6 was misunderstood by pupil 13, who answered it as if he was asked how he would write a new poem with the machine, instead of without it. Also, some pupils had difficulties in giving structured answers

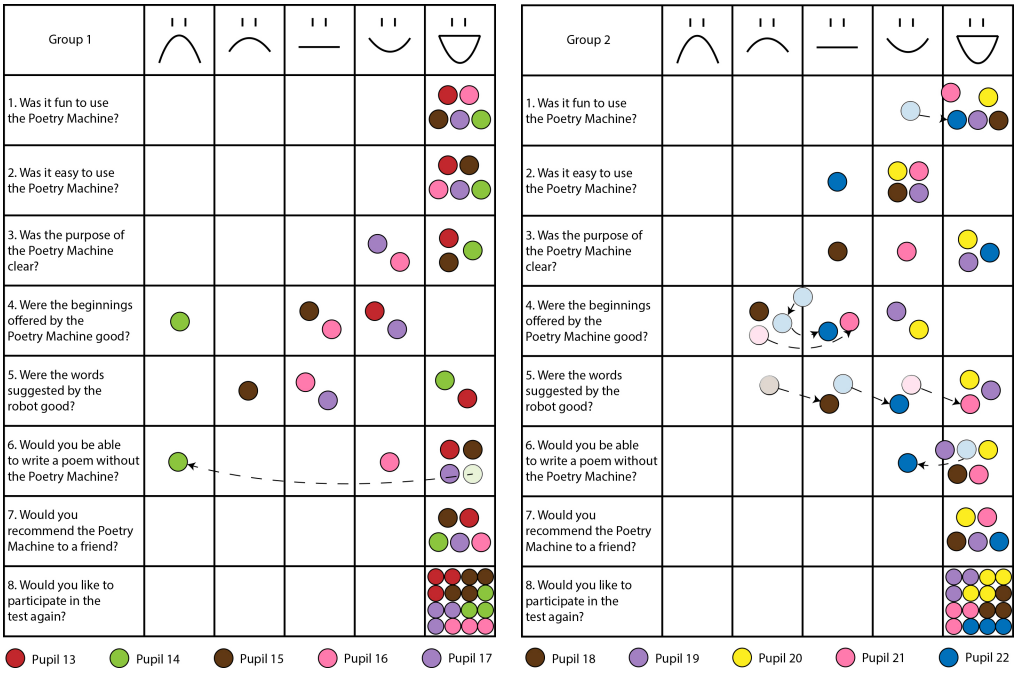

Figure 5.3: Results of the feedback game

to some questions. For example question 3 on the clearness of the concept was difficult to answer for three pupils. However, having pupils explain their answers out aloud helped us to see, what kind of difficulties they had had with answering and what kind of concepts were difficult for them. Even though pupils had difficulties in rationalising some of their opinions, it is important to note that all pupils were confident in giving an opinion. An additional option for the game board, indicating the pupil did not try the feature in question at all, would have been good, as one pupil indicated he had used the straight face option for the robot only because he had not used it.

5.7.2 Bias in the Feedback Game

As a group based activity the feedback game is vulnerable to group bias. However, we found pupils to be able to give quite independent answers. The moves made by pupils on the board during the feedback game are showed in figure 5.3. Changes made by players during the turn are indicated with arrows. Notably, none of the pupils copied all of their answers from others. The changes made by pupils during testing also make hidden biases visible: Pupil 21 originally moved the token of pupil 22 in to the frowning face on

the fourth question, when pupil 22 was thinking between the frowning face and the straight face. As a result, pupil 22 clearly moved his token to the straight face and was followed by pupil 21. Also, originally, pupil 18, 22 and 21 answered the fifth question originally more negatively than in the end. Interestingly, pupil 22 and pupil 14 also changed their answers to question six to more negative, which shows they were also able to disagree with others.

Chapter 6

Discussion

This chapter discusses the benefits and restrictions of the chosen methodologies in the light of the information they provided as well as the problems faced with their application. We first review the benefits of each method and follow that by the restrictions of each method.

6.1 Benefits of the Chosen Methods

Peer tutoring offered us a way to observe the use of the prototype very closely. This helped us to elicit a lot of usability problems with very detailed descriptions. When tutors instructed, or were prompted to instruct their tutees, by telling them what they had done in similar situations, we were given a unique chance to learn, what kind of mental models children had while writing poems with the prototype. Additionally, the peer tutoring situation gave us a chance to see how children were able to perform alone versus with a more experienced friend. The more confident attitude of the tutees participating in the test clearly showed the benefit of having children test with a friend instead of alone.

Children were capable of giving quite good answers to retrospective questions during the paired interview following the Peer Tutoring sessions. In my opinion, giving the children a chance to try problem situations hands on helped them to discuss the problems. Asking them for improvements or suggesting improvements based on their problem descriptions helped them to pinpoint the source of problems during the discussion. All in all children were able to make good improvement suggestions after the Peer Tutoring when they were reminded of problems with hands on examples.

The Group Testing, while not very accurately recorded, acted as a good method for evaluating the persistence and severity of the problems recorded

during Peer Tutoring in a new context. Clearly children faced less problems in its more realistic scenario, where they had formal instruction and the support of a teacher. Observing the use of the prototype in a more realistic educational setting also helped to evaluate the usefulness of the system.

The Feedback Game proved to be a very fast method for eliciting user feedback. It was also a good method for controlling the discussion of a group of children while simultaneously showing possible group biases visually on the board. Biases in the argumentation were also effectively reduced by having each participant start a turn at least once. Pupils seemed to be able to distinguish between different features and qualities of the program, such as the suggestions made by the robot versus the words given at the beginning, or fun versus ease of use. Notably, almost as many improvement suggestions were collected with this method, as in the paired interviews conducted after the Peer Tutoring. The suggestions are also very similar. Therefore we may note that the feedback game is roughly as accurate, and verbose a method as the interviews, but much faster to conduct. The results of the Feedback Game are also very fast to analyse, as each question can be analysed individually and pupils either tend to note they agree with a former player or build up on each other's answers producing a more manageable amount of data.

6.2 Restrictions of the Chosen Methods

Setting up Peer Tutoring sessions, holding a number of them and analysing all their data is cumbersome. Test materials have to be carefully planned and the test has to be set up outside the laboratory, amounting to a lot of equipment that has to be moved around and stored in between sessions. Each session usually takes at least an hour, but additional time has to be reserved to accommodate the schedule of the young users. Sessions are also demanding for the facilitator, as especially the tutors demand a lot of attention and encouragement to make the test successful. Children are also easy to take note if the facilitator is not completely engaged with the test. In addition analysing the sessions takes a lot of time. Analysing one record took at least three hours of an evaluators time. Additional time had to be reserved for counting user actions, gestures and comments. Finally the analysis of the raw notes also requires a lot of time and preferably multiple evaluators.

Pupils chosen for a session play a large role in the success of each Peer Tutoring session. If possible, the more talkative and active member of each pair should be used as the tutor, in order to make the time during tutor training easier for the pupil, and to elicit as much comments during the peer tutoring phase as possible. Vague tasks should be avoided with Peer Tutor-

ing, as they require more prompting to help the tutors overcome problems during the tutor training phase. This may make the tutors more dependent on the facilitator than is preferred for the peer tutoring phase. The pair interview is also subject to bias, as some pupils may be more careful to agree with their partners.

Group Testing sessions are difficult to record for detailed analysis. It was impossible to analyse the sessions for exact usability problems, with so few resources allocated for recording them. Therefore Group Testing was only useful as a filter for analysing the problems already found during peer tutoring in more detail.

The Feedback Game can only have a few questions. Because the amount of participants is quite large, compromises have to be made with the number of questions in order to keep the game short to keep all participants interested and engaged in the game. The Feedback Game also requires quite specific questions – the more specific the question the more specific the answers of pupils seem to be.

Chapter 7

Conclusions

In this thesis we have used Peer Tutoring in pairs, and a group based observation test with an experimental feedback eliciting tool called the Feedback Game to investigate the usability of a creative poetry writing tool with 9-10-year-olds. Usability testing with children is an interesting subfield of user-centered design. Its interestingness and challenges originate in the user group itself: Children are in many ways less capable than adults, including restrictions in memory, concentration, and problem solving, as well as in the physical execution of tasks with accuracy and swiftness. However, children are not a heterogeneous user-group, but rather there exists a number of different age-groups who tend to have different capabilities. Individuals then again will fit loosely on these age categories depending on their individual skill and expertise. These conditions have to be kept in mind when designing a usability test for children.

7.1 Main Findings

At the beginning of this thesis I determined three study questions related to usability testing with children: *How are usability tests with children conducted in practise? What aspects of usability need to be tested with children? What methods need to be chosen to collect information on the selected aspects?* In this section, I review most important lessons learned from literature and how they were applied in practise in this thesis.

7.1.1 How Are Usability Tests with Children Conducted in Practise?

When setting up usability tests with children, special care has to be taken to account for the special characteristics of the user group. The test environment should be familiar and child friendly. Following the approach used in many case studies, in this study, the tests were set up at the school of the participants. Literature is divided in what kind of recording equipment is suitable for children. In this study, I found that even one small film camera may notably disturb children, especially at the beginning of the test. Therefore I recommend investigating the possibility of web cameras for further studies. Material rewards are not necessary according to the literature, but I found that children were very happy with the small sticker, candy and card-rewards offered.

The number of adults present during the test situation is delicate: It is recommended to have at least one adult, a facilitator, supporting children throughout the test and literature recommends a balanced number of children and adult. I found that in addition to the facilitator, one additional observer is recommendable, as the facilitator needs to focus on supporting the children entirely as children may be disturbed by notes taken by him or her. However, two additional observers seems to be overmuch in situations where only one child is present. Detailed guidance for the facilitators actions can not be given, as different children seemed to require different methods for encouragement and guidance during the testing.

Literature recommends that tests are kept short and additional time is reserved for the researchers to catch their breath between sessions. I agree with this, since I found that the duration of sessions varied a lot and testing with children demands the undivided attention of the crew. Test tasks should be short and clear. I found that our task, writing a poem, was too general and vague and children needed a lot of help to get started when working on it alone. This countered the intention of the task, which was to promote creative thinking.

In practise, children are often recruited from small pools, with help of adults, such as parents and teachers. I also used this approach to recruit our participants and called on one teacher, who selected suitable participants from his class. Literature suggests that children with specific personality traits may be more suitable for testing. I found that in Peer Tutoring it might be more useful to have the least shy child in each pair acting as the tutor.

Ethical considerations are also important when testing with delicate participants, such as children. To this end I approached the parents of the

children explaining our research and the suitability of the methodology for children. We also interrupted two of our Peer Tutoring sessions for ethical reasons. The first interruption was made to give one tutor a chance to relax after having difficulties with the first part of the test, and the second to give another tutor a chance to have lunch at the appropriate time.

7.1.2 What Aspects of Usability Need to be Tested with Children?

Some authors, such as Hanna et al. [10] and Hourcade [13] consider traditional measures of usability, especially measures of productivity, speed and efficiency unsuitable for testing with children. Yet based on case studies, some traditional measures, such as the amount of usability errors remain important in testing with children. Instead of completely abandoning traditional measures, specialist often seem to focus on some specific measures and add additional ones. The most important additional usability measure in testing with children is fun, which has been found to correlate with usability.

In this thesis I tested for traditional usability in the sense of testing for usability errors. Additionally I was interested in the usefulness and enjoyability of the tool, as well as in collecting feedback and improvement ideas for the concept itself.

7.1.3 What Methods Need to be Chosen to Collect Information on the Selected Aspects?

Comparison of usability test methods against each other is difficult, as the results of testing depend on the context, product and users. Additionally reporting practises in the literature are varied. In general, usability test methods for children are valued by the number of correctly detected usability errors [25], cost efficiency [25] and ability to elicit authentic verbal comments from children [39]. Usability methods used when testing with children are modified from methods used with adults. Especially generally applicable methods, such as observation, interviews and questionnaires are frequently used, but with age-appropriate modifications. We selected Peer Tutoring from existing literature and combined it with a post test interview. Additionally, we developed a Group Testing method, for which we developed a feedback discussion tool aimed at small groups, called the Feedback Game. Group Testing is based on observing simulated small group education, and the Feedback Game is based on lessons learned from the Fun Toolkit.

The Peer Tutoring method was selected to capture usability problems. It

was selected as it allows for testing of both, uninstructed and peer instructed use, which is important for understanding the severity of problems in different contexts. It also gives a unique chance to understand how pupils perceive and use new concepts, when tutors instruct their tutees by telling what they did and why. Peer Tutoring was successful for collecting a number of usability problems as well as for describing them accurately, which made the large number of resources spent in holding and analysing the tests worthwhile. The Group Testing method was not very useful, but it helped to confirm the most serious errors.

Interviews and the Feedback Game were used to evaluate the usefulness of the tool. Questions such as "Would you recommend the tool?" and "Why?" worked well, and elicited comments such as "I would recommend it to my sister, because she would learn to write better". Chilldren talked about learning and ease of writing during both interviews and the Feedback Game. Interviewing the teacher for additional remarks were useful for evaluating the usefulness of the tool in a larger context.

I consider the enjoyability of the tool most difficult to evaluate. I tried observing behavioural signs throughout the Peer Tutoring sessions, as recommended in the literature. However the observation results were in clear disagreement with the feedback from the children. Self reporting abilities of children have been put to question by many, including Hanna et al. [9], and my findings may have been biased, as no well established questionnaires or tools were used to evaluate fun. However I consider the self reports more reliable in this case, as the observational data could be also interpreted as signs of concentration instead of frustration, and children seemed to have communicated with each other outside the test sessions positively about the application.

Additionally Peer Tutoring was good for analysing use patterns of children, and the interviews and the Feedback Game provided us with useful development ideas for the concept. The Feedback Game was more efficient for collecting ideas and experiences of the users.

7.2 Implications for Design

User evaluations are a key part of any user-centered design project and proper preparation is the key to their success. When selecting evaluation methods for a delicate user group, such as children and a specific context, such as school, designers will need to understand the situation from a larger perspective. To gain a balanced understanding of the use of a product in this larger framework, a mixture of test methods is recommendable. Sometimes new

methods will have to be designed to gain insight into specific issues. Mixing methods is also vital to prevent disasters resulting from mismatch between specific users and methods. With children especially, methods also require balancing and modification during testing to accommodate unexpected user behaviour, such as users having problems starting on a task, or users willing to continue playing with the prototype even after testing, as seen in this study.

Gaining insight into the use of a creative application demands a high sense of trust between the users and the designers. This trust may be easily disturbed by using too obtrusive methods, or implying judgment with your attitude. In preparing for the unexpected nature of this user-group, experience working with children helps, but it can be well simulated with reading case studies. In addition, a positive attitude, willingness to listen, and preparedness to admit your own mistakes help with gaining the trust of this user-group. Recognising their ability to give feedback is the first part of involving them more into the user-centered design process.

7.3 Future research

Although experts in general seem to consider usability testing with children to require modification of methods to suit the users' needs, accurate descriptions of modifications made to methods are strikingly few. In the literature reviewed for this thesis, Peer Tutoring was the best described method. Because Peer Tutoring demands a lot of resources, better descriptions of methods used with other contexts and tasks, as well as in more budgeted conditions are needed. Faster and cheaper methods are also required to better enable the use of iterative design practises. In its current form, Peer Tutoring takes too much time from schoolwork to be used for a large number of iterations. In addition, faster methods may enable the recruitment of participants from more varied sources, if test sessions can be arranged faster.

More research is also needed in how test tasks affect the results of usability testing with children. Lindgaard and Chattaratichart [23] proposed moving the focus of test coverage from the number of participants to tasks in the context of usability testing with adults. However, varied rules for designing test tasks for children are reported in the literature, ranging from free play to very short, specific tasks. It would be beneficial to investigate formally, what kind of tasks work with which age-groups and with what methods.

Peer Tutoring itself needs more research on what kind of participants are most suitable for it and in which roles. In our study it seemed it would have been beneficial to select less shy participants in the tutor role. However no

formal measurements of the personality traits of the participants were made. Barendregt et al. [1] studied the effect of personality traits with individual child participants. A formal evaluation of the effect of personality in Peer Tutoring is required.

Additional strategies for recording Group Testing sessions are required to investigate its usefulness as an evaluation method. For example, automatic logging of moves could help analysing the material in more detail. For now, the Feedback Game used in the Group Testing sessions seems more promising as a future research area. Further research is needed to develop unbiasing ways to support more discussion among peers. Additional investigation is needed to find what kind of questions work best for the game board. Statistically relevant samples to investigate bias within the game are also required to ensure the validity of the method.

Bibliography

[1] Barendregt, W., Bekker, M. M., Bouwhuis, D. G., and Baauw, E. Predicting effectiveness of children participants in user testing based on personality characteristics. *Behaviour & Information Technology 26*, 2 (2007), 133–147.

[2] Costabile, M. F., De Angeli, A., Lanzilotti, R., Ardito, C., Buono, P., and Pederson, T. Explore! possibilities and challenges of mobile learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008), CHI '08, ACM, pp. 145–154.

[3] Diah, N. M., Ismail, M., Ahmad, S., and Dahari, M. K. M. Usability testing for educational computer game using observation method. In *International Conference on Information Retrieval & Knowledge Management, (CAMP), 2010* (2010), IEEE, pp. 157–161.

[4] Donker, A., and Reitsma, P. Usability testing with young children. In *Proceedings of the 2004 Conference on Interaction Design and Children: Building a Community* (2004), IDC '04, ACM, pp. 43–48.

[5] Edwards, H., and Benedyk, R. A comparison of usability evaluation methods for child participants in a school setting. In *Proceedings of the 6th International Conference on Interaction Design and Children* (2007), IDC '07, ACM, pp. 9–16.

[6] Egloff, T. H. Edutainment: A case study of interactive cd-rom playsets. *Computers in Entertainment (CIE) 2*, 1 (2004), 1–22.

[7] Fowler, A. Measuring learning and fun in video games for young children: a proposed method. In *Proceedings of the 12th International Conference on Interaction Design and Children* (2013), IDC '13, ACM, pp. 639–642.

[8] Hanna, L., Neapolitan, D., and Risden, K. Evaluating computer game concepts with children. In *Proceedings of the 2004 Conference on*

Interaction Design and Children: Building a Community (2004), IDC '04, ACM, pp. 49–56.

[9] Hanna, L., Risden, K., and Alexander, K. Guidelines for usability testing with children. *Interactions 4*, 5 (1997), 9–14.

[10] Hanna, L., Risden, K., Czerwinski, M., and Alexander, K. J. The role of usability research in designing children's computer products. *The design of children's technology* (1999), 3–26.

[11] Hartson, H. R., Andre, T. S., and Williges, R. C. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction 15*, 1 (2003), 373–410.

[12] Hertzum, M., and Jacobsen, N. E. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction 13*, 4 (2001), 421–443.

[13] Hourcade, J. P. Interaction design and children. *Foundations and Trends in Human-Computer Interaction 1*, 4 (2008), 277–392.

[14] Höysniemi, J., Hämäläinen, P., and Turkki, L. Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers 15*, 2 (2003), 203–225.

[15] Höysniemi, J., Hämäläinen, P., and Turkki, L. Wizard of oz prototyping of computer vision based action games for children. In *Proceedings of the 2004 Conference on Interaction Design and Children: Building a Community* (2004), IDC '04, ACM, pp. 27–34.

[16] ISO. 9241–11:1998. Ergonomic requirements for office work with visual display terminals (vdts). part 11: Guidance on usability (iso 9241-11:1998). *The international organization for standardization* (1998).

[17] Kallio, T., and Kekäläinen, A. Improving the effectiveness of mobile application design: User-pairs testing by non-professionals. In *Mobile Human-Computer Interaction - MobileHCI 2004*, S. Brewster and M. Dunlop, Eds., vol. 3160 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, pp. 315–319.

[18] Kantosalo, A., and Riihiaho, S. Let's play the feedback game. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (2014), NordiCHI '14, ACM, pp. 943–946.

[19] Kantosalo, A., Toivanen, J. M., Xiao, P., and Toivonen, H. From isolation to involvement: Adapting machine creativity software to support human-computer co-creation. In *Proceedings of the Fifth International Conference on Computational Creativity* (2014), pp. 1–8.

[20] Lahti, J., Siira, E., and Törmänen, V. Development and evaluation of media-enhanced learning application. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia* (2012), MUM '12, ACM, pp. 5:1–5:10.

[21] Lewis, J. R. Sample sizes for usability tests: Mostly math, not magic. *interactions 13*, 6 (Nov. 2006), 29–33.

[22] Lim, Y.-K., Stolterman, E., and Tenenberg, J. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction 15*, 2 (2008), 7:1–7:27.

[23] Lindgaard, G., and Chattratichart, J. Usability testing: what have we overlooked? In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), CHI '07, ACM, pp. 1415–1424.

[24] MacFarlane, S., Sim, G., and Horton, M. Assessing usability and fun in educational software. In *Proceedings of the 2005 conference on Interaction design and children* (2005), IDC '05, ACM, pp. 103–109.

[25] Markopoulos, P., and Bekker, M. On the assessment of usability testing methods for children. *Interacting with Computers 15*, 2 (2003), 227–243.

[26] Murphy, J., and Cameron, L. The effectiveness of talking mats ®with people with intellectual disability. *British Journal of Learning Disabilities 36*, 4 (2008), 232–241.

[27] Obrist, M., Igelsböck, J., Beck, E., Moser, C., Riegler, S., and Tscheligi, M. Now you need to laugh!: investigating fun in games with children. In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology* (2009), ACE '09, ACM, pp. 81–88.

[28] Obrist, M., Moser, C., Fuchsberger, V., Tscheligi, M., Markopoulos, P., and Hofstätter, J. Opportunities and challenges when designing and developing with kids @ school. In *Proceedings*

of the 10th International Conference on Interaction Design and Children (2011), IDC '11, ACM, pp. 264–267.

[29] Patel, M., and Paulsen, C. A. Strategies for recruiting children for usability tests. In *Meeting of the Usability Professionals Association: FL (June 2002).* (2002).

[30] Read, J., MacFarlane, S., and Casey, C. Endurability, engagement and expectations: Measuring children's fun. In *Interaction design and children* (2002), vol. 2, Shaker Publishing Eindhoven, pp. 1–23.

[31] Read, J. C., and MacFarlane, S. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 Conference on Interaction Design and Children* (2006), IDC '06, ACM, pp. 81–88.

[32] Riihiaho, S. User testing when test tasks are not appropriate. In *European Conference on Cognitive Ergonomics: Designing Beyond the Product — Understanding Activity and User Experience in Ubiquitous Environments* (2009), ECCE '09, VTT Technical Research Centre of Finland, pp. 21:1–21:9.

[33] Ritchie, G., Manurung, R., Pain, H., Waller, A., Black, R., and O'Mara, D. A practical application of computational humour. In *Proceedings of the 4th International Joint Conference on Computational Creativity* (2007), pp. 91–98.

[34] Sim, G., Cassidy, B., and Read, J. C. Understanding the fidelity effect when evaluating games with children. In *Proceedings of the 12th International Conference on Interaction Design and Children* (2013), IDC '13, ACM, pp. 193–200.

[35] Sim, G., MacFarlane, S., and Read, J. All work and no play: Measuring fun, usability, and learning in software for children. *Computers & Education 46*, 3 (2006), 235–248.

[36] Stinson, J., McGrath, P., Hodnett, E., Feldman, B., Duffy, C., Huber, A., Tucker, L., Hetherington, R., Tse, S., Spiegel, L., Campillo, S., Gill, N., and White, M. Usability testing of an online self-management program for adolescents with juvenile idiopathic arthritis. *Journal of medical Internet research 12*, 3 (2010).

[37] Toivanen, J., Toivonen, H., Valitutti, A., and Gross, O. Corpus-based generation of content and form in poetry. In *Proceedings of the Third International Conference on Computational Creativity* (2012), pp. 175–179.

[38] Toivanen, J. M., Järvisalo, M., and Toivonen, H. Harnessing constraint programming for poetry composition. *Proceedings of the Fourth International Conference on Computational Creativity* (2013), 160–167.

[39] van Kesteren, I. E. H., Bekker, M. M., Vermeeren, A. P. O. S., and Lloyd, P. A. Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In *Proceedings of the 2003 Conference on Interaction Design and Children* (2003), IDC '03, ACM, pp. 41–49.

[40] Waller, A., Black, R., O'Mara, D. A., Pain, H., Ritchie, G., and Manurung, R. Evaluating the standup pun generating software with children with cerebral palsy. *ACM Transactions on Accessible Computing 1*, 3 (2009), 16:1–16:27.

[41] Yue, W. S., and Zin, N. A. M. Usability evaluation for history educational games. In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human* (2009), ICIS '09, ACM, pp. 1019–1025.

Appendix A

The Background Questionnaire

Mikä on nimesi?

Minkä ikäinen olet?

Monennella luokalla olet?

Pidätkö kirjoittamisesta?

Millaisista runoista pidät?

Oletko aiemmin kirjoittanut itse runoja?

Kuinka usein käytät tietokonetta? (ympyröi sopivin vastaus)

Kerran päivässä

1-3 Kertaa viikossa

Kerran viikossa

Harvemmin

En koskaan

Käännä ->

Figure A.1: Background Questionnaire Front

Missä paikoissa käytät tietokonetta? (ympyröi kaikki sopivat vastaukset)

Kotona

Koulussa

Kerhossa

Jossain muualla

Kenen kanssa käytät tietokonetta? (ympyröi kaikki sopivat vastaukset)

Opettajan

Kaverien

Äidin tai isän

Siskon tai veljen

Mitä yleensä teet tietokoneella? (ympyröi kaikki sopivat vastaukset)

Teen koulutöitä

Kirjoitan sähköposteja

Surffaan netissä

Pelaan pelejä

Jotain muuta

Mitä muita laitteita käytät? (ympyröi kaikki sopivat vastaukset)

Tablettitietokonetta

Älypuhelinta

Tavallista kännykkää

Figure A.2: Background Questionnaire Back

Appendix B

Results of the Background Questionnaire

Table B.1 presents the background of Peer Tutoring participants

Table B.2 presents the background of Group Testing participants

Session	Participant		Sex	Age	Grade	"Do you like writing?"	"What kind of poems do you like?"	"Have you written poems before?"	"How often do you use the computer?"						"Where do you use the computer?"				"With whom do you use the computer?"					"What do you usually do with a computer?"						"What other equipment do you use?"		
									Once a day	1-3 Times a week	Once a week	Less	Never		At home	At school	At a club	Other	With a teacher	With friends	With mom or dad	With sister or brother	With someone else	School work	Write e-mail	Use the internet	Play games	Something else		A tablet	A smartphone	A normal cellphone
Session 1	Pupil 1	Written	F	9	3	I write stories	short ones	I have not		x				Varies a lot	x	x			Alone									x		x		
		Comment																														
	Pupil 2	Written	F	9	3	Yes	Animal poems	A couple	x						x	x		x	x	x		x		x		x			x		
		Comment						A couple of poems							At the library				Alone					rehearses for exams, reads blogs								
Session 2	Pupil 3	Written	F	9	3	I don't know	I don't know	No				x			x	x		x				x		x			x		x	x	
		Comment													At friends																	
	Pupil 4	Written	F	9	3	Yes	Nice poems	Yes				x			x	x				x						x	x				
		Comment					Does not know how to elaborate	Likes writing stories better											Alone					Listens to music online								
Session 3	Pupil 5	Written	F	9	3	I like to write poems	Story-poems	Not much				x			x	x		x			x			x			x		x		
		Comment						3 or a couple											Alone					Does schoolwork and rehearses for exam								
	Pupil 6	Written	F	9	3	Yes	Funny poems	no				x			x	x										x	x			x	x
		Comment				Ok to write at school	More familiar with poems written on other native language												Alone					Plays and searches for information online								
Session 4	Pupil 7	Written	M	10	3	Yes	Old ones	Yes		x					x	x			x	x					x		x	x		x	
		Comment				Writes mostly at school	Selects adjective based on facilitators suggestions																	Writes to a friend, uses Skype								
	Pupil 8	Written	M	9	3	Sometimes	-	-							x				x	x	x	x		x		x	x			x	x
		Comment				Writes for fun																		Plays and does schoolwork online							Has several game consoles	
Session 5	Pupil 9	Written	M	10	3	Yes	-	No							x															x		
		Comment				Writes mostly for school													Alone					Plays and rehearses for exam online								
	Pupil 10	Written	M	10	3	Yes	Exciting ones	I have not		x					x	x						x		x		x	x				x
		Comment					Can't think of an example												Alone					Uses the Internet for school								
Session 6	Pupil 11	Written	M	10	3	Yes	-	No							x					x									x	x	
		Comment				Writes stories																		Watches Youtube								
	Pupil 12	Written	M	9	3	Yes	Funny ones	No							x			x			x	x					x	x		x		
		Comment				Writes stories																		Writes with the computer, searches information online								

Table B.1: Peer Tutoring participants' answers to the background questionnaire

Session	Participant	Sex	Age	Grade	"Do you like writing?"	"What kind of poems do you like?"	"Have you written poems before?"	"How often do you use the computer?"				"Where do you use the computer?"				"With whom do you use the computer?"					"What do you usually do with a computer?"						"What other equipment do you use?"		
								Once a day	1-3 Times a week	Once a week	Less	Never	At home	At school	At a club	Other	With a teacher	With friends	With mom or dad	With sister or brother	With someone else	School work	Write e-mail	Use the internet	Play games	Something else	A tablet	A smartphone	A normal cellphone
Session 1	Pupil 13	M	9	3	I like writing	I don't know	I have		x				x					x					x					x		
	Pupil 14	M	10	3	No	Funny ones	No		x				x			x		x	x	x				x	x	x	x	x	x
	Pupil 15	M	10	3	Yes	Funny ones	Yes		x				x	x				x	x					x		x	x	x	
	Pupil 16	F	9	3	I do	Ones with rhymes	No						x				Alone												
Session 1	Pupil 17	F	10	3	Yes	-	-				x		x	x		x		x					x		x		x	x	
													At friends				Alone												
Session 2	Pupil 18	F	9	3	Yes	Animal poems	No				x		x					x						x	x		x	x	x
																	Alone												
	Pupil 19	F	9	3	Yes	Exciting ones	I have	x					x	x				x									x	x	
	Pupil 20	F	9	3	Yes	Funny ones	No		x				x	x				x				x		x	x		x		
	Pupil 21	M	10	3	Somewhat	Exciting ones	No				x		x						x			x					x	x	
	Pupil 22	M	9	3	Yes	Funny ones	No			x			x	x				x				x	x		x		x	x	
																	Alone												

Table B.2: Group Testing participants' answers to the background questionnaire

Appendix C

Post Task Interview Questions

	Original	Translation
1.	Montako tähteä antaisitte työkalulle?	How many stars would you give for the tool?
1a.	Yksi tähti on tosi huono työkalu, viisi tosi hyvä	One star means a very poor tool, five a very good one
2.	Mikä työkalussa oli kivaa teidän mielestä?	In your opinion, what was fun about the tool?
2a.	Miksi? Millä tavalla kivaa?	Why? In what way?
3.	Mikä työkalussa oli tylsää teidän mielestä?	In your opinion, what was boring about the tool?
3a.	Miksi? Millä tavalla tylsää?	Why? Boring in what way?
4.	Ongelmatilanteet (1-3)	Problem situations(1-3)
4a.	Käyttämättömät featuret	Unused features
4ai.	"Näytti siltä, että te ette käyttäneet tätä nappia ollenkaan - muistatteko mitä se tekee?"	"It seems like you did not use this button at all - do you remember what it does?"
4aii.	"Oliko joku erityinen syy, miksi te ette käyttäneet sitä?"	"Was there a specific reason to why you did not use it?"
4aiii.	"Jos te neuvoisitte nyt vielä jotain kaveria, kertoisitteko te tästä napista?"	"If you would instruct another friend, would you tell them about this button?"
4b.	Ongelmatilanteet	Problem situations
4bi.	"Näytti siltä, että te etsitte tätä nappia aika pitkään. Oliko siihen joku syy?"	"It seemed like you searched for this button for a long time. Was there a specific reason for it?"
4bii.	"Miten te muuttaisitte sitä, että teidän kaverit löytäisi sen helposti?"	"How would you change it, so that your friends would find it easily?"
5.	Suosittelisitteko tätä ohjelmaa omille kavereille?	Would you recommend this program for your friends?
5a.	Kelle?	For whom?
5b.	Miksi?	Why?
6.	Haluatteko vielä kommentoida jotain tähän ohjelmaan liittyen?	Would you like to comment the program freely?
7.	Haluaisitteko osallistua uudelleen tällaiseen testiin, jos siihen olisi mahdollisuus?	Would you like to participate in another test like this, if you had the chance?
7a.	Miksi/ Miksi ei?	Why/why not?

Appendix D

Teacher Interview Questions

	Original	Translation
1.	Miten oppitunti mielestäsi sujui?	In your opinion, how did the class go?
2.	Millä tavalla Runokone–tunti erosi tavallisesta oppitunnista?	In what ways did the Poetry Machine–lesson differ from a regular lesson?
2a.	Huomasitko eroa yksittäisissä oppilaissa?	Did you notice any difference in specific pupils?
3.	Mitkä olivat testatun ohjelman hyvät puolet?	What are the good sides of the program tested today?
3a.	Miksi?	Why?
4.	Mitkä olivat ohjelman heikot puolet?	What are the bad sides of the program?
4a.	Miten korjaisit ongelman?	How would you fix the problem?
5.	Miten kehittäisit työkalua, jotta se tukisi oppilaiden työskentelyä?	How would you develop the tool, so that it would support the students' work better?
6.	Miten kehittäisit työkalua, jotta se tukisi opettajien työskentelyä?	How would you develop the tool, so that it would support teachers' work?
7.	Haluaisitko käyttää työkalua jatkossa osana oppituntia?	Would you like to use the tool as a part of your lessons in the future?
7a.	Miksi/ Miksi ei?	Why/Why not?
7b.	Millaisen oppitunnin osana käyttäisit työkalua?	What kind of a lesson would you use it as a part of?
8.	Ovatko oppilaat, jotka osallistuivat aiempiin testeihin puhuneet Runokoneesta?	Have the pupils, who participated in previous tests talked about the Poetry Machine?
8a.	Millaista palautetta olet saanut?	What kind of feedback did you hear?
9.	Millaista palautetta antaisit koeasetelmasta?	What kind of feedback would you give of the test situation itself?
10.	Haluaisiko luokkasi osallistua vastaavaan projektiin uudelleen?	Would you like to participate in a project like this again?
11.	Jäikö sinulle vielä mieleen jotain, mitä haluaisit sanoa työkalusta?	Was there still something, you would like to say about the tool?

Appendix E

Usability Evaluation Results

Table E.1 presents ideas gathered during Peer Tutoring

Table E.2 presents ideas gathered during Group Testing

Table E.3 presents usability problems gathered with Peer Tutoring

Table E.4 presents the distribution of problems during Peer Tutoring

Table E.5 presents bugs and their distribution during Peer Tutoring

	Idea	Session
1	Words should automatically be divided into their own boxes	1
2	Users should be able to remove all words, or all extra words automatically	2
3	Words given by the machine should be more familiar to the users	3
4	Words given by the robot should be more related to the words given to the robot	5
5	New options could be displayed under the word to be replaced	5
6	A quick way to add punctuation is needed	6

Table E.1: Ideas collected in the Peer Tutoring

	Idea	Session
1	The beginnings should have more familiar words	2
2	The words given by the robot should be more related to the topic	1 and 2
3	The words given by the robot should have better rhymes	1
4	The beginnings given by the computer should have more rhymes	1

Table E.2: Ideas collected in the Group Testing Feedback Game

Table E.3: Usability Problems

					Frequency				
No.	Screen	Problem	Description	Improvement suggestion	TT	PT	Int.	All	Severity
1	Screen 2	Moving words: Problems with targeting	A majority of the pupils has problems targeting the white circles when moving words. Especially moving words into the beginning or the end of a row is difficult, as they mostly try to target the words themselves.	The targeting areas should be enlarged, preferably so that the whole area between words can be used.	6	4	1	11	3
2	Screen 2	Modifying words: Moving the cursor	Pupils are unable to move the cursor with arrow buttons. The cursor spawns at the beginning and it can not be moved with the mouse	The cursor should be made to work "normally"; it should appear at the end of the box and be movable with the mouse	6	4	0	10	3
3	Screen 2	Moving words: Moving a word on top of another	Many of the pupils assumed that the targeting of the dropped word within the word it is dropped on matters. If they wanted to drop the word before a word, they targeted the beginning of the word	There should be two areas for drop targeting inside a word	3	3	0	6	3
4	Screen 2	Moving words: Undefined problems	Some problems were described so vaguely it was impossible to determine their reason		3	3	0	6	3
5	Screen 2	Robot: Pupils do not understand that they should drop words on the robot	Pupils needed extra instructions for dragging words to the robot, or they gave wrong instructions to the tutee.	A tooltip instruction is needed. The existing instructions should be more clearly visible from the background	3	1	1	5	3
6	Screen 2	Buttons: The buttons in the menu are not seen as buttons	Many pupils ignored the top menu entirely, until they were told about it.	Make the menu more button-like. Use stronger colors.	4	0	0	4	3
7	Screen 2	Concept: The purpose of the robot is unclear	Some pupils are unclear about what the robot does or react negatively to it.	Improve instructions.	2	0	2	4	3
8	Screen 2	Publishing: What does the publishing do?	Pupils who published their poems did not know what publishing menas.	Improve instructions. Add a mouseover tooltip.	2	2	0	4	3
9	Screen 2	Buttons: Targeting the robot is difficult	Some pupils can not target the button. Instead they try to drag elements to very specfic places, such as under the button, even when the speech-bubble-box is not opened	Increase buttons so that the area around them can also be used for targeting.	1	3	0	4	3
10	Screen 2	Buttons: Targeting the trashcan is difficult	Pupils click past or drop past the trashcan.	Use the area around the trashcan for targeting as well	3	0	0	3	3
11	Screen 3	Saving poem details: The pupil forgets to save the changes	Students do not notice the button used to save modifications to the poem details	Change the name into "Save"	2	1	0	3	3
12	Screen 2	Concept: Some of the pupils think, all words and rows need to be used	The starting point of the Poem Machine is not clear	Improve instructions.	2	1	0	3	3
13	Screen 2	Moving words: Grabbing words is difficult	Pupils are unable to grab a word as they point besides it.	Enlarge the padding of the word so that grabbing is easier. Change cursor when grabbing is possible.	1	1	1	3	3
14	Screen 1	Buttons: The start button does not look like a button	Some pupils asked how the tool is started or where to press after selecting the theme	Make the start-button look more like a button.	1	2	0	3	3
15	Screen 2	Modifying words: Leaving the modify-state	Some pupils had problems closing the modify-state of words	If possible, close the modify-state automatically	2	0	0	2	3
16	Screen 2	Buttons: The trashcan button is not recognized as a trashcan	Some pupils can not find the trashcan and try all kinds of ideas, including right-click instead to remove words	Make the trashcan in the button look more like a trashcan	2	0	0	2	3
17	Screen 2	Robot: Problem with moving the words from the robot	Pupils are unable to grab words of the robot	Increase the size of the words	2	0	0	2	3

18	Screen 3	Buttons: The feather pen is not understood as a way to return to the writing window	Many pupils clearly wanted to go back to the writing window. They did not notice the feather-pen-button. Some used the back-button of the browser or the facilitator helped them to return.	Design a button for returning to the writing mode	2	0	0	2	3
19	Screen 2	Moving words: Opening the modify-state by accident	Pupils tried to grab words, but opened the modify-state by accident instead	Enlarge the padding of the word so that grabbing is easier. Change cursor when grabbing is possible.	1	2	0	3	3
20	Screen 2	Moving rows: Grabbing rows is difficult	Pupil has problems grabbing the row. It will not stick.	Enlarge the padding of the row so that grabbing is easier. Change cursor when grabbing is possible	1	1	0	2	3
21	Screen 1	Concept: The pupils assume that the program starts automatically	Some pupils wait for the program to start on its own	Make the start-button look more like a button.	1	1	0	2	3
22	Screen 3	Buttons: The write a new poem-button does not look likea button	Pupils do not realise they can return to the beginning by pressing "Write a new poem"	Design a button for returning to the beginning	1	1	0	2	3
23	Screen 3	Buttons: Returning to the writing window with the feather pen button is associated with the first screen.	Two of the pupils tried to go back to the first page by pressing the feather-pen-button	Design a button for returning to the writing mode	1	1	0	2	3
24	Screen 2	Robot: The words of the robot do not look movable	Pupils do not realise that the words given by the robot can be dragged directly into the poem	Make the words in the robot more similar to the ones in the poem	1	1	0	2	3
25	General	Appearance: The bug states are not clearly shown	Pupils often continue to work without noticing a bug state	If a bug happens, notify the user and reload automatically	1	1	0	2	3
26	Screen 3	Saving poem details: The save button is not clear	The text on the save button says "Update". It is bad.	Change the name into "Save"	1	0	0	1	3
27	Screen 2	Appearance: Activation of the modify-state is not shown clearly	Pupils do not notice the cursor	Make the font larger, change color of the background	1	0	0	1	3
28	Screen 2	Moving words: Grabbing a row by accident	Pupil tries to grab a word, but grabs a row instead by accident.	Enlarge the padding of the word so that grabbing is easier. Change cursor when grabbing is possible.	0	2	0	2	3
29	Screen 2	Material: Sentences given by the robot are too long	Some of the sentences given by the robot were too long	Shorten the sentences used by the robot	1	0	0	1	3
30	Screen 2	Robot: The robot does not react to words with punctuation matks	The robot is not working when users drag words that end with a punctuation mark into it	Remove punctuation from the input automatically	1	0	0	1	3
31	Screen 2	Buttons: The trashcan button is not recognized as a button	All of the pupils are not understanding that the trashcan button is a button	Make the trashcan in the button look more like a button	1	0	0	1	3
32	Screen 2	Robot: Words moved from the robot replace the word they are dropped on	A word dragged from the robot replaces the word it is dropped on, instead of appearing after it	Change the behavior of the words dragged from the robot into the same with the rest. Problem can then be fixed by removing the new word.	1	0	0	1	3
33	Screen 2	Robot: It is impossible to drag words to the robot if they are in the modify-state	Pupil tries to move a word with the modify-state still open	If possible, close the modify-state automatically	1	0	0	1	3
34	Screen 2	Moving words: Undoing moves	One pupil would have tried to put a word he dragged from the robot back	Change the behavior of the words dragged from the robot into the same with the rest. Problem can then be fixed by removing the new word.	0	1	0	1	3

35	Screen 2	Moving words: It is impossible to grab a word, when the modify-state is active	Pupil tries to grab a word, but it does not move as the modify-state is still open	If possible, close the modify-state automatically	0	1	0	1	3
36	Screen 2	Robot: Sentences dragged from the robot change into parts of an existing sentence	Sentences dragged from the robot turn into parts of existing sentences instead. Pupils are unable to tell which part is the new sentence and which belongs to the old poem	Sentences dragged from the robot should behave like sentences in the interface.	0	1	0	1	3
37	Screen 2	Rhymes: Problem with grabbing words from the rhyme-tool	Pupils are unable to grab words from the rhyme-tool	Remove the rhyme tool, the robot should be enough alone	0	1	0	1	3
38	Screen 2	Creating a new word: Writing words into the same box	Pupils write multiple words to the same word box	Divide words automatically, if more words are added into the same box.	3	3	2	8	2
39	Screen 2	Creating a new word: The pupil does not know how to add a word	Pupils were unable to add new words. They used different strategies or asked for help. It was not always clear even after they accidentally had added a word	Improve instructions. Add a mouseover tooltip.	4	2	1	7	2
40	Screen 2	Appearance: Words are moving between rows when their length changes	The placement of the words on rows changes automatically when the words are edited. This is puzzling to the pupils	It is impossible to prevent this problem entirely, but it can be alleviated, if it is possible to animate the changes. Targeting should be made possible into all parts of the sentence.	3	3	0	6	2
41	Screen 2	Removing: Pupils do not think of removing extra words	Pupils do not think they need to remove extra words.	Show more clearly, which words have been modified. Possibly animate words that have not been used.	3	1	0	4	2
42	Screen 2	Concept: The difference between the robot and the rhyme-tool is unclear	In the prototype, the robot and the rhyme-tool have the same functionality. Pupils mix them up, the researchers mix them up, and finally only the robot is used.	Remove the rhyme tool, the robot should be enough alone. Add more specific options into the robot	1	2	0	3	2
43	Screen 3	Saving poem details: It is not clear when the saving has finished	Pupil presses the "Update"-button multiple times	When the saving is completed, show a message to the user	1	2	0	3	2
44	Screen 2	Moving rows: Pupils are unaware that you can grab a whole row	Pupils do not think they can also move the rows	Improve instructions. Change the cursor to show grabbing is possible.	2	0	0	2	2
45	Screen 2	Material: The quality of the starting options is poor	Some pupils critisised the quality of the starting options. Pupil 1 did not think she could make complete sentences out of them, she started removing words with the same content	Change the poem generator into a better one	2	0	0	2	2
46	Screen 2	Material: Forgetting about the robot	Pupil had problems coming up with new words, but forgot to use the robot, or did not remember how to when asked	Possibly animate the robot and play the animation, if long pauses in the workflow are noticed.	2	0	0	2	2
47	Screen 2	Appearance: Spelling mistakes are not seen	Spelling mistakes are not easy to notice in the writing window	Enlarge the font size. Change background color more visible.	2	0	0	2	2
48	Screen 3	Concept: The concept of the read mode is unclear	Pupils do not understand what is the purpose of the read mode, apart from one, who clearly uses it to review her poem	Improve instructions	2	0	0	2	2
49	Screen 2	Appearance: Pupils do not notice extra elements	It is hard to notice extra elements in the writing mode	Show more clearly, which words have been modified. Possibly animate elements that have not been used.	1	1	0	2	2
50	Screen 2	Material: Quality of the words given by the robot is poor	Pupils 5 and 6 think the words given by the robot are unclear. Sometimes the robot also gave English words.	Improve word quality by using more modern sources. Make the search more contextually dependent.	0	1	1	2	2
51	Screen 2	Robot: It is not possible to drag words from the robot on the robot	One pupil tried to use words given by the robot to generate more with the robot	Make it possible to drag words from the robot into the robot.	1	0	0	1	2

52	Screen 2	Buttons: A pupil presses the finish-button by accident	One pupil was investigating the top menu and accidentally pressed the finish-button	Make a more clear distinction on which buttons work by draggin and which not.	1	0	0	1	2
53	Screen 1	Concept: One pupil is too shy to press the starting button	One pupil does not dare to press the start button, instead just keeps the mouse on top of it.	Make the start-button look more like a button. Add a tooltip.	1	0	0	1	2
54	Screen 2	Moving words: Dragging words is difficult	One pupil had clearly problems with the dragging. He was able to do the drag-and-drop move, but just did not think of it intuitively.	Improve instructions	1	0	0	1	2
55	Screen 2	Buttons: One pupil does not understand that most menu buttons work by draggin and dropping	Pupil does not understand that elements should be dragged onto the buttons	Make a more clear distinction on which buttons work by draggin and which not.	1	0	0	1	2
56	Screen 2	Moving words: The concept of moving words	Pupil asks if it is possible to move words, but does not understand how the white circles should be used for targeting	Improve instructions	1	0	0	1	2
57	Screen 2	Rhyme: Words from the rhyme-tool can not be dragged and dropped on the tool itself	Pupil tried to use words given by the rhyme-tool to generate more words with the rhyme-tool	Remove the rhyme tool, the robot should be enough alone.	0	1	0	1	2
58	Screen 3	Appearance: The poem in the read mode looks modifyable	Pupil 10 tries to remove words from the read-mode poem	Remove the border, which looks too much like a text field	0	1	0	1	2
59	Screen 2	Buttons: One pupil drags elements on the finish-button	Pupil drags elements on top of the finish-button. Assuming it works similarly to the rest of the buttons	Make a more clear distinction on which buttons work by draggin and which not.	0	1	0	1	2
60	Screen 2	Robot: It is not clear to the pupils that it is possible to drag rows to the robot	Pupils do not realise that it is possible to drag the whole row to the robot	Make it more clear that rows can be moved too, by changin cursors when entering the row and giving better instructions	0	0	1	1	2
61	Screen 2	Robot: Words can not be written directly into the robot	Pupils have to add extra words into their poem to drag them onto the robot	Add a small input field into the robot	0	0	1	1	2
62	Screen 2	Material: More options are needed in the robot	Pupil 10 would have liked more options to drag directly from the robot	Add a button for getting more options into the robot	0	0	1	1	2
63	Screen 2	Mental model: Pupils do not seem to care to divide the poem into rows	It is not clear to the pupils why the poem is structured into rows	Make the distinction between rows more clear, try to fit elements of one row on to the same row	3	3	1	7	1
64	Screen 2	Removing: Pupils are removing whole rows word by word	Pupils tried to remove rows one word at a time. This also leaves empty rows into the poem	Improve the instructions of the trashcan. Show with a cursor change that grabbing is possible	3	2	1	6	1
65	Screen 2	Material: Too many starting options	Some pupils got many starting options. They had to scroll, which in itself was not difficult, but some forgot to.	Make number of possible beginnings more limited, for example between 3-6	2	2	0	4	1
66	Screen 2	Material: Too few starting options	Some pupils got very few starting options. Pupil 6 also commented on that	Make number of possible beginnings more limited, for example between 3-6	1	3	0	4	1
67	Screen 2	Concept: The robot was not needed	The robot was not useful for some pupils, as they did not use it, and were even unable to think of how they would use it	Improve instructions	0	1	3	4	1
68	Screen 2	Removing: Pupils try to remove words by emptying the container	Many pupils tried to remove words by emptying the word containers	If a word box is emptied, remove the word automatically. Animate trashcan when this happens	2	1	0	3	1
69	Screen 2	Creating rows: New rows created by accident	Pupils added or moved new words on new rows	Improve instructions	1	1	0	2	1
70	Screen 1	Selecting theme: Pupils are unsure, if the selection affects anything	One pupil asked if the selection of the theme has any affect, another wondered if it is needed at all.	Improve instructions, add a mousover tooltip.	1	1	0	2	1

71	Screen 2	Robot: It is unclear, if the robot needs to be closed	It was unclear to some pupils, if the robot speech-bubble-box needed to be closed	Change the close-button into an x, reducing the mental load.	0	2	0	2	1
72	Screen 1	Selecting theme: Pupil does not notice the theme selection	One pupil did not seem to notice the theme selection	Increase the size of the theme selection, make the area of active elements more visible	1	0	0	1	1
73	Screen 1	Selecting theme: One pupil asks if he can write the theme down	One pupil asked if the theme needs to be written down, or selected	Increase the size of the theme selection.	1	0	0	1	1
74	Screen 2	Modifying words: Pupil does not know how to use delete	Pupil was unable to erease the word until the cursor had been moved to the end	Make the cursor work so that it is possible to paint text	1	0	0	1	1
75	Screen 2	Appearance: Pupil does not like that the word boxes are tipped	Pupil 1 did not like the word containers being tilted	The appearance of the boxes does not seem functional to the pupil. Make it more clear	1	0	0	1	1
76	Screen 2	Trashcan: Pupil tries to move a targeting circle into the trashcan	Pupil tried to drop a targeting circle into the trashcan	Show with a cursor change that the circle can not be moved	1	0	0	1	1
77	General	One pupil would like to use the keyboard for everything	Pupil tries to use the keyboard at first	Improve instructions.	1	0	0	1	1
78	Screen 2	Consistency: The ”close” button in the robot seems too much like a word	The closing button of the robot, rhyme-tool and the trashcan looks too much like a word	Change the close-button into an x, reducing the mental load.	0	1	0	1	1
79	Screen 2	Creating words: Creating a word by accident	The pupil stops adding a new word, but adds an empty one by accident	If the new word is left empty, do not create one	0	1	0	1	1
80	Screen 2	Concept: Purpose of the rhyme-tool is left unclear	Pupils do not understand what the purpose of the rhyme-tool is	Remove rhyme-tool, the robot should be enough alone.	0	0	1	1	1
81	Screen 2	Removing: A shortcut for removing all words is needed	Two pupils would have liked a button for removing all words in order to begin from an empty poem	Add a possibility to remove all words into the trashcan	0	0	1	1	1
82	Screen 2	Removing: A shortcut for removing all extra words is needed	Pupil wanted the extra words to be deleted automatically	Add a possibility to remove unused words into the trashcan	0	0	1	1	1

Problem	Recorded in																		Recorded by											
	T1			T2			T3			T4			T5			T6			T1		T2		T3		T4		T5		T6	
	P1	P2	I1	P3	P4	I2	P5	P6	I3	P7	P8	I4	P9	P10	I5	P11	P12	I6	R1	R2	R1	R2	R1	R2	R1	R3	R1	R3	R1	R3
1	1			1			1	1	1	1	1		1	1		1	1		2	1	1		3	4	5		3		7	1
2	1	1		1			1			1	1		1	1		1	1		4	1	2	1	1	2	5	2	2	1	2	1
3	1									1	1		1	1			1		1						5	1	3	1	1	
4	1										1		1	1		1	1		1	1						1	1	1	2	
5				1			1					1		1		1					1			1		2		1	2	
6	1			1						1			1						1		1					2	1	1		
7	1					1				1		1								1	1				1	1				
8	1													1		1	1		1	1							1	1	1	1
9								1					1	1			1						1				1	1	1	
10				1			1			1											1		1		1					
11							1							1		1							1	1			3		1	
12	1						1										1		3				2						1	1
13										1		1					1								2	3			1	
14							1				1						1						1	2	1	1			1	
15	1						1												1				1	1						
16							1			1													1		1	2				
17	1															1			1										1	
18	1												1						1								2	2		
19				1							1			1									1	1	1		1			
20	1													1					1								1			
21							1				1												1		1					
22										1	1														2	3				
23										1	1															2				
24				1													1				2								1	
25	1				1														1			1								
26										1															1	1				
27				1																	1									
28					1						1											1			1					
29	1																		1											
30	1																		1	1										
31							1																	1						
32													1														2			
33													1														1			
34														1													1			
35																	1												1	
36														1													1			
37														1													1	1		
38		1	1	1	1	1	1							1		1			3	1	3	2	2	2			4	1	1	
39	1			1	1					1	1		1		1				1	1	6	1			2		1	1	2	
40	1							1		1				1		1	1			1			1		1				4	
41							1			1	1					1							1		3	2	2			
42	1	1						1											2	1				3						
43										1	1			1											1	2	1			
44	1			1															1		1	1								
45	1												1						2								1			
46	1						1												1					1						
47							1						1											1			1	1		
48	1												1						1	1								1		
49										1				1											2	1	1	1	1	
50								1	1														2	1						
51													1														1			
52	1																		1											
53				1																			1	1						
54				1																	3									
55				1																	1									
56							1																1	1						
57														1													1	1		
58														1														1		
59														1															1	1
60									1														1	1						
61															1													1		
62															1												2	2		
63		1		1	1	1	1	1		1										1	2	2	2	2		2				
64		1		1		1	1			1							1		1	1	1	1	1	2	2	1			1	1
65		1								1						1	1		1						4	3			3	3
66								1			1		1	1									2	1		1	1	2		
67		1				1						1						1		1	3	1			1	1			1	1
68				1			1							1							2	1	1	1			2	2		
69	1													1					1								2			
70				1				1													1		1							
71								1						1									1				1	1		
72										1															1			1		
73							1																1							
74				1																		1								
75	1																				1									
76										1															1					
77							1																	1						
78								1																1						
79																	1												3	
80			1																1	2										
81						1															2	2								
82						1															1	1								

Table E.4: Distribution of problems and their recordings in Peer Tutoring Tests.

					T1		T2		T3		T4		T5		T6		Frequency			T1		T2		T3		T4		T5		T6	
	Bug	Description	Improvement suggestion	Screen	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	TT	PT	Total	R1	R2	R1	R2	R1	R2	R1	R3	R1	R3	R1	R3
1	Updating changes is interrupted	At times the updating of poem changes is interrupted	The number of tries to update the system is not enough	Screen 2	1		1		1		1			1		1	4	2	6	1	2	2	2	2	2	1		1	1	2	
2	A test popup is opened during editing	Old popups for testing were opened when students accidentally clicked some elements that were not activated for the test	Popups were deactivated for the group testing	Screen 2					1	1	1			1		1	2	3	5					3	4	1	1	2	1	4	
3	An empty row appears in the trash-can	An empty row appears in the trash-can when the last word is removed	There is no need to put empty rows into the trashcan, if they are removed word by word	Screen 2		1											0	1	1	1											
4	Overflow in the speech-bubble-box	The box under the robot and the trashcan overflows with certain parameters	Need to make the content fit better	Screen 2	1	1	1		1	1					1	1	4	3	7	3		3	2	2						5	
5	Adding a word looks wrong	Adding a word looks wrong while the system is updating changes	Need to make the system more responsive	Screen 2				1														1									

Table E.5: Bugs and their frequency and recording during testing

Appendix F

Observation Forms and Their Results

F.1 form for analysing use during Peer Tutoring. Its results are shown in Table F.1.

F.2 form for analysing gestures during Peer Tutoring. Its results are shown in Table F.2

F.3 form for analysing comments during Peer Tutoring. Its results are shown in Tables F.3 and F.4.

F.4 form for analysin important activities during Group Testing.

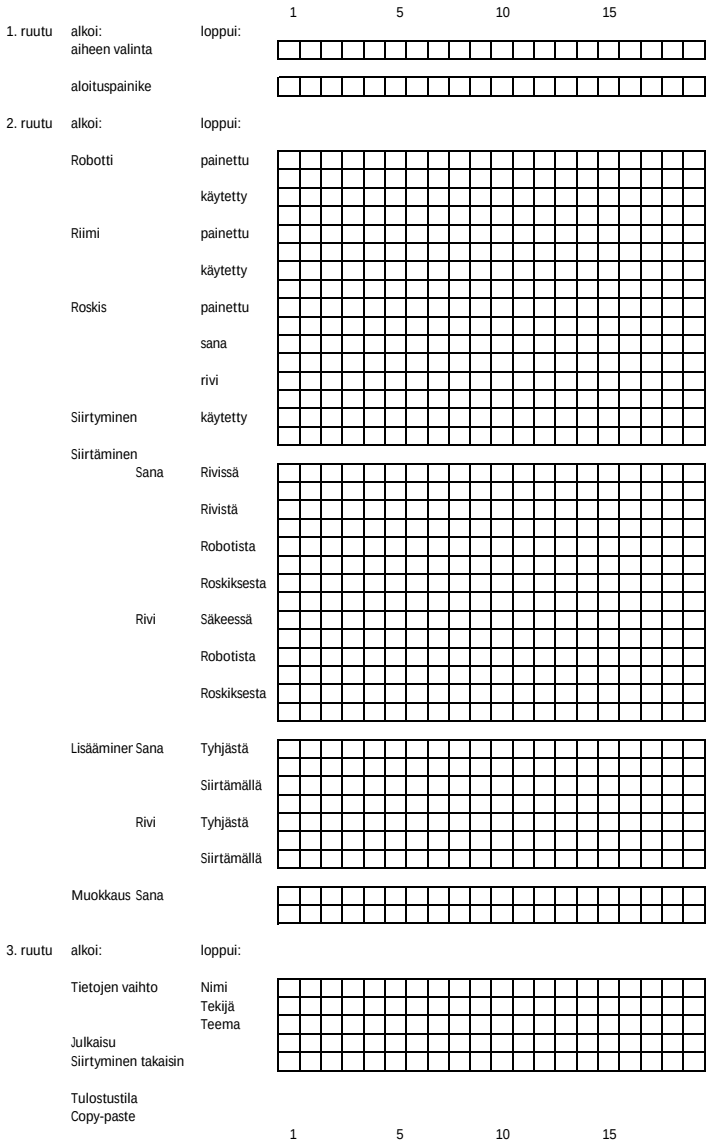

Figure F.1: Form for analysing use during Peer Tutoring

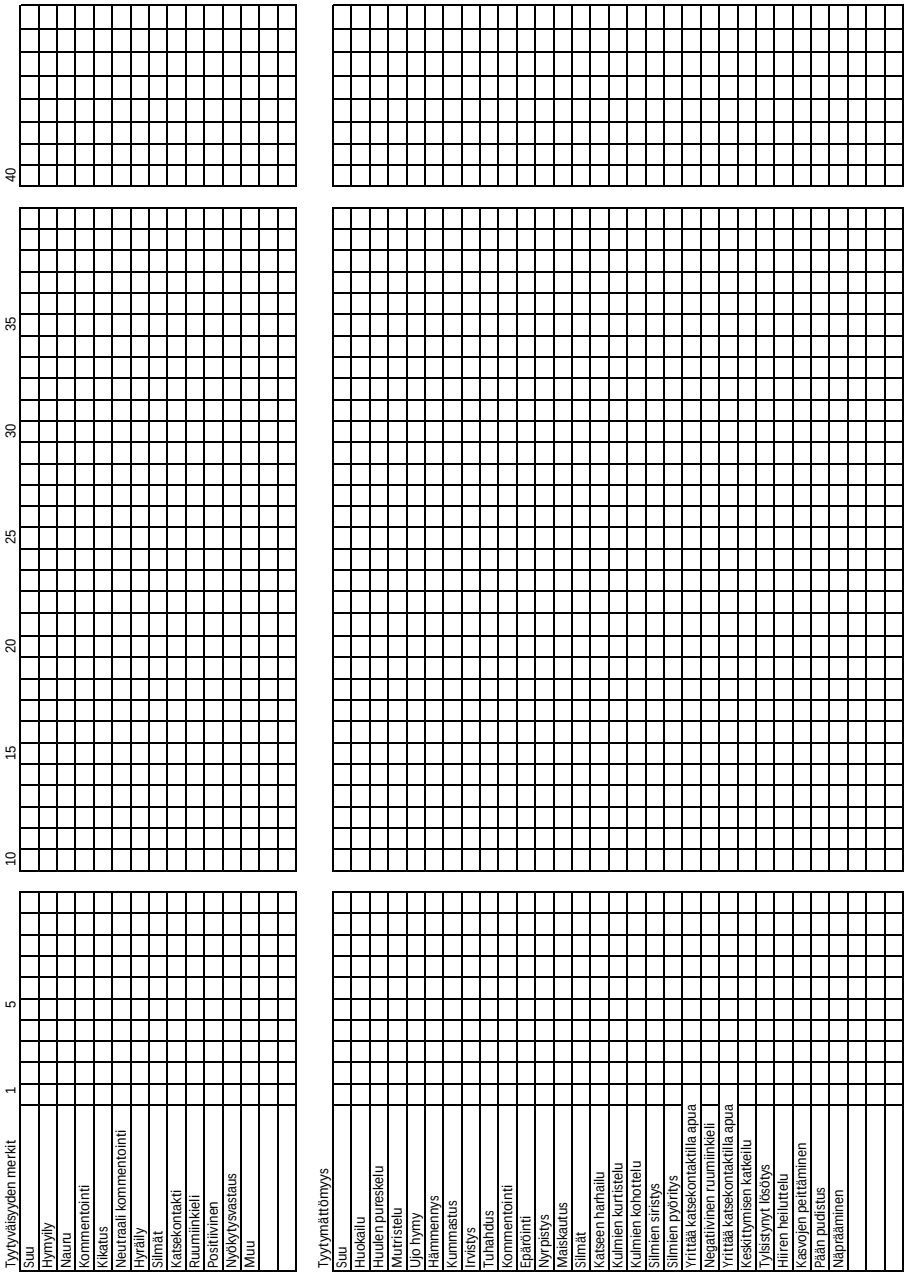

Figure F.2: Form for analysing gestures during Peer Tutoring (final version)

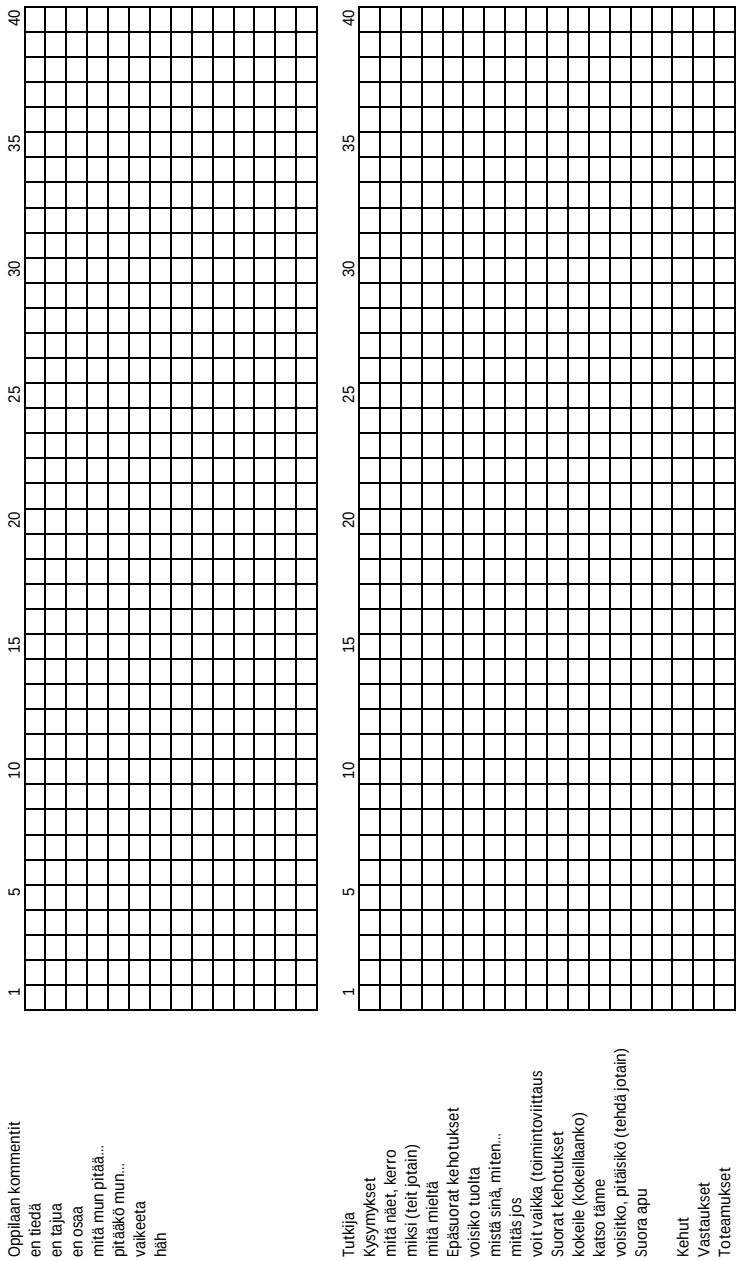

Figure F.3: Form for analysing comments during Peer Tutoring

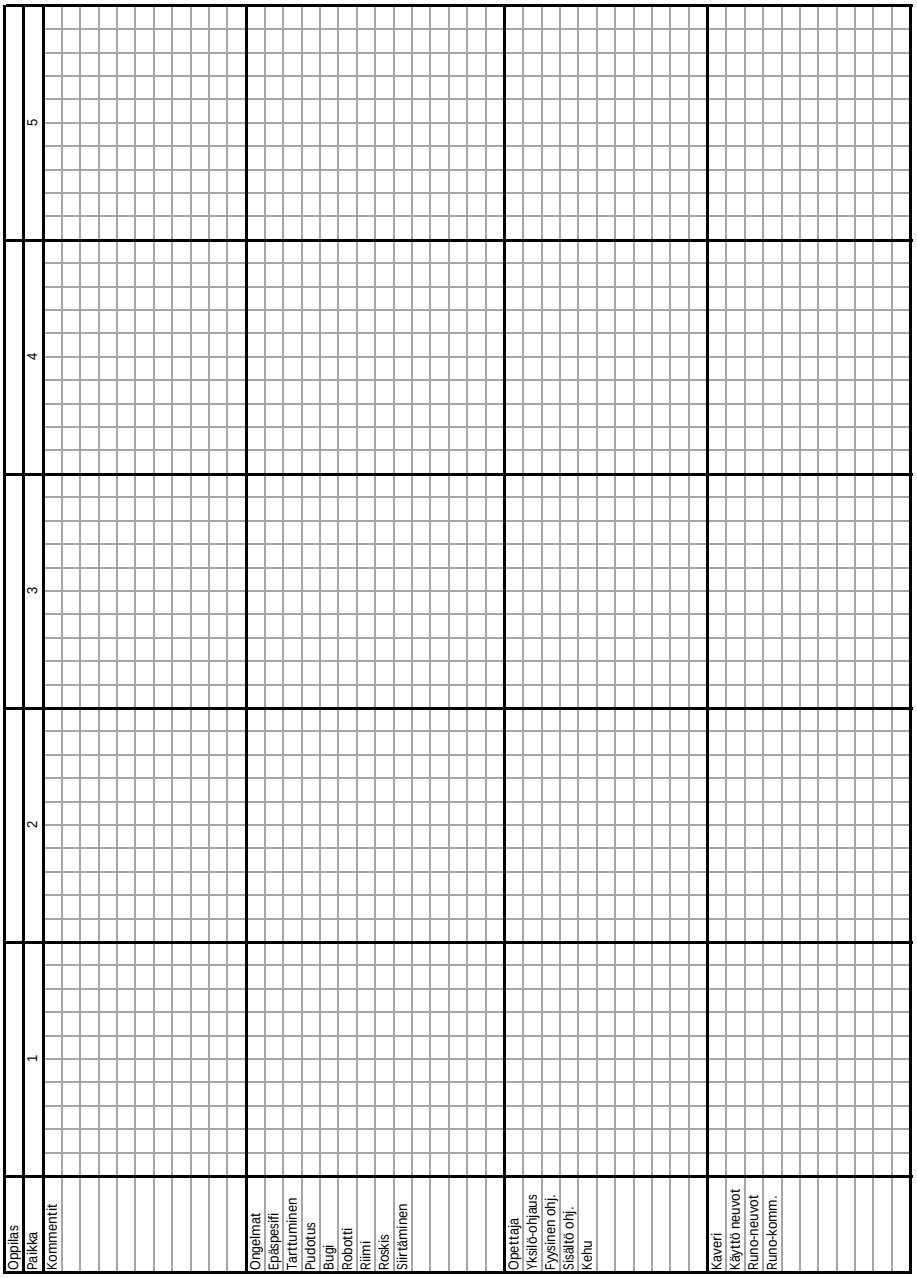

Figure F.4: Form for analysing activities during Group Testing

		Pupil												Total count		
		Pupil 1	Pupil 2	Pupil 3	Pupil 4	Pupil 5	Pupil 6	Pupil 7	Pupil 8	Pupil 9	Pupil 10	Pupil 11	Pupil 12	Tutors	Tutees	All
Screen 1	Theme selection	1	2	1	1	1	1		1	1		2	1	6	6	12
	Start button	1	2	1	1	1	1	2	1	1	1	1	1	7	7	14
Screen 2	Robot: opened	1		1		1	2			1		1		5	2	7
	Robot: Words	3	1	4		1	1			15	1	5	1	28	4	32
	Robot: Rows	3								1	1			4	1	5
	Rhyme-tool: Opened	1					1	1		1				3	1	4
	Rhyme-tool: Words	1					1				1			1	2	3
	Trashcan: Opened	1		1		4	1				1			6	2	8
	Trashcan: Words	17	12	10		10	3	21	2				13	58	30	88
	Trashcan: Rows		2	1	4	2		1	2		1	2	2	6	11	17
	Finish button	3	1	1	1	1	1	2	1	1	4	1	1	9	9	18
	Moving Word: Witing row	12						2	6	1	4		4	15	14	29
	Movin Word: Between Rows	13				1		11	6		4		17	25	27	52
	Moving Word: From the Robot	6				2								8	0	8
	Moving Word: From the Trashcan			2										2	0	2
	Moving Row													0	0	0
	Moving Row: From the Robot	1									3			1	3	4
	Moving Row: From the Trashcan													0	0	0
	Adding Word: From a target circle	4	19		1	8	12				6			12	38	50
	Adding Word: By moving	4												4	0	4
	Adding Row: From a focus circle	2									2		11	2	13	15
	Adding Row: By moving										1			0	1	1
	Modifying word	15	9	17	7	1		12	3	17	10	9	2	71	31	102
Screen 3	Changing Name		1		1	1	1	1	1		2	2	1	4	7	11
	Changing Author											2	1	2	1	3
	Changing Theme													0	0	0
	Publishing										1	1	1	1	2	3
	Returning to write-mode				1			2	1	1	3	1		4	5	9

Table F.1: Function usage per pupil during peer tutoring

		Test 1			Test 2			Test 3			Test 4			Test 5			Test 6			Total			
		TT	PT		TT	PT		TT	PT		TT	PT		TT	PT		TT	PT		TT	PT		
			Tutee	Tutor		Tutee	Tutor		Tutee	Tutor		Tutee	Tutor		Tutee	Tutor		Tutee	Tutor		Tutee	Tutor	Grand total
Positive	Smiling	11	5	1	6	2	5	1	1	1	12	17	1	25	25	14	9	0	1	64	49	24	137
	Laughing	4	0	3	7	1	2	1	6	5	2	3	2	1	11	1	0	0	0	15	20	6	41
	Snickering	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4	4	13
	Giggling	0	0	0	0	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Leaning in in concentration	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Enthusiastic nodding when answering	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
	Humming	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	2
	Relief	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
	Eye contact	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2
	Insight	0	0	0	0	0	0	0	1	1	0	0	0	1	3	0	0	0	0	1	4	0	5
	Grinning	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1
	Pointing	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1
	Talking with hands	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1
Grand total of positive gestures		20	5	5	14	7	12	3	8	7	14	20	3	27	43	15	10	1	1	88	82	37	207
Negative	Sighting	0	0	1	0	0	0	0	2	0	0	1	0	0	2	0	0	0	0	0	5	5	10
	Huffing	4	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	6	0	0	6
	Bored	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	3	2	4	9
	Losing concentration	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
	Wandering eyes	3	0	1	3	0	0	0	0	4	1	0	1	0	1	2	0	0	0	7	4	4	15
	Playing with hair	0	0	0	0	1	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	3
	Playing with ear	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Scratching	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	3
	Playing with clothes	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	5	0	0	6	1	0	7
	Playing with environment	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	3	2	5
	Knitting one's face	38	2	3	0	3	0	1	2	0	2	4	2	3	3	0	1	22	0	45	34	8	87
	Frowning	0	1	0	15	3	3	0	2	0	3	0	0	6	11	1	4	10	1	28	24	2	54
	Turning up one's nose	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	1	3	2	1	6
	Pouting	0	0	0			0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	2
	Hesitating	2	0	0	3	0	0	1	1	0	1	0	0	2	2	0	7	1	0	16	5	0	21
	Suprise	0	0	0	0	1	0	0	0	0	0	1	0	1	3	0	0	2	0	1	6	0	7
	Baffelment	0	0	0	0	0	0	2	3	0	0	0	0	1	0	0	0	0	0	3	3	1	7
	Rolling one's eyes	3	0	2	9	0	0	2	1	1	0	0	0	0	0	0	0	1	0	14	2	2	18
	Shy smile	0	0	0	6	0	0	0	1	0	0	0	0	0	0	0	11	0	1	17	1	1	19
	Asking help with eye contact	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2
	Raising eyebrows	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
	Squinting eyes	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
	Tsik tsiking	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	2
	Frustration	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	4	7
	Hiding face	1	0	3	4	2	3	0	2	1	0	1	3	1	11	9	0	11	0	6	25	15	46
	Biting lip	4	0	0	3	0	0	0	1	0	0	0	0	0	3	0	0	3	1	7	7	2	16
	Grimace	3	0	0	0	0	0	0	0	1	2	0	0	1	1	0	0	3	0	6	5	3	14
	Shaking head	6	0	0	0	1	1	3	0	2	0	0	0	0	0	0	1	4	0	10	4	0	14
	Worried	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1
Grand total of negative gestures		71	5	10	44	11	11	12	15	9	11	7	8	18	42	13	29	60	4	185	142	57	384

Table F.2: Pupil gestures and body language during peer tutoring. TT denotes the tutor training phase and PT the peer tutoring phase.

			Pupil 1	Pupil 3	Pupil 5	Pupil 7	Pupil 9	Pupil 11	Total
Tutor comments	Problems with program or task	I don't know.	7	1		1		1	10
		I don't get it.	5						5
		I can't think of anything!		2	1			2	5
		I'm not sure.					1		1
		I can't do it!	2						2
		It doesn't work				2			2
		This is difficult	3						3
	Questions	How do you...		1		1		1	3
		Can I...		1					1
		What should I...	3	1		1	1	3	9
		Should I...	5					2	7
		What happens if...?				3			3
		Other		1	5			1	7
	Answers			4	2	4	11	1	22
	Thinking aloud	Comments poem		1	1				2
		I got an idea		1					1
		I'm ready				1			1
		Other	1		1	10	1	1	14
Total			26	13	10	23	14	12	98
Facilitator comments	Prompting for think aloud	What are you thinking?					2		2
		What do you see?	6	2	3	4	2		17
		Why did you...?	1	1		1	1		4
		What do you think of?	6		2		6	1	15
	Indirect suggestions	Do you think that might...	2	3	2				7
		How would you...?	4	2	2	1	3	2	14
		How about...?	7	10	3	1	3		24
		You can, for example...	3	3	2	1	1	1	11
	Direct suggestions	Do ...					2		2
		Try it out!	10	7	7	3	3	4	34
		Look here	2	3	2	3	1		11
		Should you ...	7	7	10	1	1	1	27
		Shall we?		3					3
		Let's continue			1				1
	Direct help	Physical	3	5	2	2	2	2	16
		Non-Physical	4	1	2	3		1	11
	Praise		1	4	5	1	3	3	17
	Answers		5	3				2	10
	Counter question			1	2				3
	General comments		7	15	1	10	8	3	44
	Help with poem		7	5			3	1	16
Total			75	75	46	31	41	21	289

Table F.3: Pupil and facilitator comments during the tutor training phase.

			Pupil 2	Pupil 4	Pupil 6	Pupil 8	Pupil 10	Pupil 12	Total
Tutee comments	Problems with program or task	I don't know.					1		1
		I don't get it.					3		3
		I can't think of anything!	2		1	1		1	5
		It doesn't work						2	2
		I don't remember					1		1
		This is difficult		1					1
		It's not fun					1		1
	Questions	How do you...			2	1			3
		Can I...			1				1
		What should I...				1	1		2
		Should I...			1	1	1	3	6
		What is this...?			1		4		5
		What if?					1	1	2
	Questions for friend		5	2			1		8
	Questions for facilitator		1	1	1		4	2	9
	Answers to the facilitator			1				1	2
	Thinking aloud	Comments poem		1			6		7
		Now I see		1			1	3	5
		I would like to				1			1
		Other		2	1	1	11	2	17
	Improvement suggestion						2		2
	Total		8	9	8	6	38	15	84
Tutor comments	Direct instructions		7	12	6	12	4	7	48
	Direct intervention		3	1		1			5
	Indirect instruction	Try it	1			1			2
		Explains possibilities	8	5	9	3	6	7	38
		Describes own process			1		4	2	7
		Other				1			1
		Through questions	1						1
	Problems with program	I don't know	1		2				3
		I can't do it	1		1				2
	Poem	Comments	3				1		4
		Suggestions	1			2			3
		Corrects spelling		2					2
		Reads aloud				3			3
		Tells what a poem is					1		1
	Cheering						1		1
	Questions for friend					3			3
	Questions for facilitator						1		1
	Answers to facilitator			1					1
	Total		26	21	19	26	18	16	126
Facilitator comments	Answers		1				4	1	6
	Questions	for the tutee			1		7		8
		for the tutor		4	3	3	7	8	25
	Prompting	Tutee		3	2		2	1	8
		Tutor		3	2		5	1	11
		Both			1				1
	Direct physical help				2			3	5
	Comments						3	1	4
	Cheering for tutee						5		5
Total			1	10	11	3	33	15	73

Table F.4: Tutor, tutee and facilitator comments during the peer tutoring phase.