A New User Interface for a Text Proofreading Web Portal in a Digitization and Crowdsourcing Context

Dr Pietro Murano, Department of Computer Science, The Universal Design of ICT Research Group, OsloMet - Oslo Metropolitan University, Oslo, Norway.

Abstract

Purpose

This paper presents a new user interface design for text proofreading portals in a digitization and crowdsourcing context. Several of the current proofreading portals lack usability in their user interfaces. The aim of the new design is to increase user performance and satisfaction.

Approach

An empirical experiment was conducted to evaluate the new user interface as a comparison with 18thConnect – TypeWright proofreading portal. Two of the main measures involved times and errors and this approach was considered to be good for these kinds of measures allowing a good degree of control. Nevertheless, personal opinions were also very important and these were elicited by means of a post-experiment questionnaire.

Findings

The data was statistically analysed and overall the new user interface helped users to perform better in terms of task time. Errors were also better with the new user interface, but the differences were not statistically significant. Furthermore, users were more satisfied with the new user interface. User satisfaction measures were mostly statistically significant.

Originality

As far as has been ascertained, there have been no systematic studies evaluating a new design with an existing design of a proofreading portal. Therefore, this research is considered to be very original and if implemented widely would be very valuable to the mass digitization aims.

Keywords

Digitizing old documents, proofreading, crowdsourcing, usability, user interface design, evaluation, cognitive load.

Introduction

The aim of the research presented in this paper is to design and evaluate a new user interface (‘front-end’) prototype for crowdsourcing portals in the context of proofreading (the process of reading texts with the aim of finding and correcting errors) digitized text which was originally in printed paper form. Specifically, there is an ongoing effort at trying to digitize old documents such as newspapers, books and other old texts (Coyle, 2006). Coyle (2006) reasons that mass digitization aims ‘…to digitize everything, or in this case, every book ever printed’.

The research described in this paper does not deal with the ‘back-end’ aspects of optical character recognition (OCR), segmentation or other automated aspects that are involved in producing an output that can be proofread by a human. However, one perspective and discussion about the process of converting documents to online access can be seen in a paper by Yacoub, Burns, Faraboschi, Ortega, Peiro and Saxena (2005).

Pietro Murano, (2018) "A new user interface for a text proofreading web portal in a digitization and crowdsourcing context", International Journal of Web Information Systems, Vol. 14 Issue: 1, pp.107-122, https://doi.org/10.1108/IJWIS-05-2017-0042.

A further aim of the research is to try and make the new user interface more usable than existing crowdsourcing portal user interfaces. In this research, usable is intended as improving performance, eliciting more positive feelings from the users and lowering cognitive load. Coyle (2006) states that the user interfaces that allow access to digitized materials are the 'weakest point' in the mass digitization realm. Although this research is not dealing with the user interface that allows access to the digitized content in its final format, it is linked with it, because crowdsourcing portals are allowing access to pre-final versions of digitized content via a user interface. These also are weak in terms of usability and therefore require more effort to improve them. The result of such an improvement should be a faster and more accurate turnaround of digitized materials to final versions made available to the public. Another result should be the proofreading community being more efficient whilst achieving more personal satisfaction in the proofreading tasks they carry out.

In addition, Rahmanian and Davis (2014) are very clear in suggesting that the usability of such portals (e.g. MTurk) is lacking and that they can cause an increase in cognitive load (see Research Approaches section below for a definition of cognitive load).

This paper is divided into a number of sections. The sections will discuss some of the main relevant literature, the new user interface, the experimental evaluation carried out for the new user interface including the main statistical findings and the conclusions and ways forward.

Literature Review

Digitizing old texts, such as books or other documents poses many challenges. Drira (2006) discusses some of the problems in digitizing old local government documents which are considered important to continue to access. Drira discusses how old documents are affected by how they were stored and by the original materials used to produce the documents. This affects the ultimate readability of a digitized version of such a series of documents. Their research contributes by providing a 'typology for different types of degradation of old document images' and a non-segmentation approach for dealing with 'document degradation'.

In another study by Bulacu, van Koert, Schomaker and van der Zant (2007), very specific characteristics of old handwritten correspondence between the Queen of the Netherlands and the government of the Netherlands is dealt with, including layout issues of the original materials. They built a 'layout analyzer' 'as a first step towards the automatic content-based retrieval of document images'. Although the authors seem to have achieved good outcomes with their work, they do also acknowledge that dealing with the analysis of the layout in terms of handwritten texts continues to be a challenge.

Furthermore in Gao, Rusiñol, Karatzas, Antonacopoulos and Lladós (2013) some of the problems involved with digitizing old newspapers are discussed. The authors discuss that 'newspapers are digitized and OCRed in bulk'. Then 'in order to organize the image database by date, newspaper, location, etc. a user has to manually segment the flow of digitised images into newspaper issues and label them with multiple metadata' (Gao, Rusiñol, Karatzas, Antonacopoulos and Lladós, 2013). In line with these aspects, the authors developed a system that tried to help human users to segment the 'incoming flow of scanned images' with some automation. The aim was to be able to determine a particular issue of a newspaper by being able to ascertain which page was the first page of the newspaper. Furthermore the authors suggest a future possibility of using 'active learning' and crowdsourcing in this context.

According to Lang and Ross (2011) Crowdsourcing is explained as '… the process by which a task is outsourced to an undefined group of people (the crowd) rather than contracting professionals to accomplish that task.' For a more detailed and systematic definition of crowdsourcing see the paper by Estellés-Arolas and González-Ladrón-de-Guevara (2012).

In Zaidan and Callison-Burch (2011) some research was carried out into using crowdsourcing for translation and a ‘quality control model’. They concluded that good translations, close to professional quality, were possible using their approach. Furthermore, they concluded that translations done via crowdsourcing were a lot cheaper than using professional translators.

Kobayashi, Ishihara, Itoko, Takagi and Asakawa (2013) did a Japan-specific study where they tried to leverage the linguistic skills of older citizens in a crowdsourcing context. They found that older users are willing to contribute their time to help proofread text. Furthermore, the authors concluded that older users can have good linguistic skills. They had the idea of using older users for tasks involving linguistic issues and younger users for more ‘technical’ aspects. They also developed a prototype which aimed to help with the proofreading process, taking into consideration issues of older users using the system. One example was that the authors developed the prototype with larger fonts with the aim of helping older users. The authors mention carrying out an initial evaluation, but there are very few concrete results presented in the paper, other than suggesting that the proofreading that took place was faster than the typical proofreading time for proofreading books.

Lang and Ross (2011) worked on transcribing and proofreading handwritten text. They used Google docs and Amazon’s Mechanical Turk (MTurk) along with a crowdsourced workforce. They found that costs were a lot less than using a professional transcription service.

In Chrons and Sundell (2011) a gamification approach was devised for proofreading text which had been through the optical character recognition (OCR) process. The authors found that the crowdsourced effort produced an accuracy greater than 99%. Further, their paper states that the typeface of the original printed works was difficult to read, thus showing the 99% accuracy level to be remarkable.

The above examples show that crowdsourcing is being used in various ways and with some good degrees of success. However, to the author’s knowledge there has been very little research done on trying to design a usable user interface for carrying out proofreading tasks in the context of digitizing old documents. There are some portals currently in use for this purpose, e.g. 18thConnect - TypeWright (2017), Distributed Proofreaders (2017) and Trove - National Library of Australia (2017).

Although some of the digitization process described in this brief literature review involves several automated aspects, when crowdsourcing is used, there is clearly a group of humans at work and they will use a user interface of a dedicated web portal (e.g. 18thConnect - TypeWright (2017), Distributed Proofreaders (2017) and Trove National Library of Australia (2017)) to achieve the stages of proofreading. This indicates that the web portal user interfaces should be as usable as possible for all users.

Web usability overall is a very important topic for all concerned with the usage and functioning of a web site. A lot of effort has been dedicated to devising methods for evaluating web sites.

One example concerns using Item Response Theory (IRT) to aid the evaluation of web sites. ‘IRT describes a set of mathematical models aimed at measuring latent traits (that is, individual profile characteristics that cannot be measured directly)’ (Tezza, Bornia, de Andrade, 2011). This approach is not intended to replace existing methods used in evaluation.

Another example shows protocol analysis (think aloud method) being used to evaluate commercial web sites. The author’s work indicates at the time of their writing, that basic usability mistakes are done in web site development (Benbunan-Fich, 2001).

Further, some references indicate that web designers are still making very basic mistakes when it comes to web design. For example, Loranger (2017) explicitly states that a home page should not have an active home link as it can cause confusion. This clearly affects usability etc.

Also the Web Accessibility Initiative makes an effort in fostering a web that is accessible to everyone (W3.org, 2017a). In line with this the Web Content Accessibility Guidelines (WCAG) have been written to try and have 'a single shared standard for web content accessibility (W3.org, 2017b)'. Linked to this is also the main principle of universal design (UD). Originally this did not concern computer systems or user interfaces, however in recent years the UD philosophy has been extended to computer systems and user interfaces. Universal design concerns 'the design of products and environments that can be used and experienced by people of all ages and abilities, to the greatest extent possible, without adaptation (Story, 1998 and The Centre for Universal Design, 1991)'.

Linked to the above, the literature indicates that at times there is a lack of usability in web portals for specific purposes. For example Oakley and Daudert (2016) in their paper in the context of a web portal for atmospheric data indicate that accessing scientific data can be frustrating. Furthermore, they found their attempts at usability evaluation to be beneficial.

Youngblood and Youngblood (2013) found that usability could be improved in 'county' web portals. From their paper, it can be seen that even basic approaches to usability can make improvements to a portal.

Also, Ismailova (2017) conducted an evaluation of several Kyrgyz Republic government web sites using some automated tools. One of the findings was that usability was lacking in several of the web portals evaluated.

Furthermore, a study by Selden and Orenstein (2011) in the context of government portals and electronic employee recruitment in the USA, indicated that usability had an effect on the number of job applications received for government type jobs. They observed that more usable job application portals tended to increase the total number of job applications for a particular open position.

In addition, a review done by Irizarry, Dabbs and Curran (2015) concerning patient portals for health records and engagement, reached various conclusions concerning patient engagement, where one of these concerned usability. One of the article's streams suggested that good usability can affect patient adoption of such portals.

This brief literature review has shown that there is a lot of effort going on in the world at finding good and efficient ways of digitizing old texts and that crowdsourcing is used in various contexts including aspects of proofreading digitized texts. This review has further shown that other researchers have also noticed usability deficiencies in various kinds of web portals. In addition, general web usability is very important for all web designers and developers. This all indicates that the work being presented in this paper is valuable as it aims to increase the overall usability of proofreading portals and thus foster better general web usability. The next section of this paper provides a brief description of the approach used for this research.

Approach to the Research

The research being done has currently been started with the hypothetico-deductive approach (Popper, 1934). This approach is empirical in nature. Further, the method for collecting data has been grounded in empirical experimentation in a laboratory.

While it is acknowledged that this approach has strengths and weaknesses, it was chosen because it aids in collecting more concrete data that is statistically analysable. However, one

weakness of this approach is that a laboratory-based experiment could be viewed as artificial and lacking in ecological validity.

Since the main questions being looked at as part of the topic for this paper concern performance and subjective opinions in the form of task times, errors and 'quantified' (scaled questionnaire) subjective opinions, it was felt that this was a possible suitable approach.

The main issue was to try and find out if the new prototype user interface was more usable than the 18thConnect – TypeWright proofreading user interface. As indicated above, usable in this instance concerns performance and user satisfaction.

These aspects are linked to the theory of cognitive load (Sweller et al, 2011). In terms of user interfaces, 'the cognitive load imposed by a user interface is the amount of mental resources that is required to operate the system' (Whitenton, 2013). Cognitive load 'is a global term, which refers to the mental resources a person has available for solving problems or completing tasks at a given time' (Oviatt, 2006).

Whitenton (2013) suggests that removing visual clutter from the user interface would help in reducing cognitive load. Whitenton further suggests to 'offload tasks' to the system – where possible. This suggests that confusing interactions should be minimised and at best completely redesigned to be non-existent.

Sweller et al (2011) document various techniques for measuring cognitive load. Some of these are to use a scaled questionnaire with participants immediately after performing some tasks. The questionnaire should contain questions that elicit opinions on the mental effort used to do the task(s). Furthermore, measuring performance and errors have given indirect indicators regarding cognitive load, where lower performance and more errors can suggest higher cognitive load in a user. (see the results and discussion and conclusions section below for further details).

Therefore, if the new user interface prototype is better in terms of usability, one would expect that performance and user satisfaction would be better with lower errors and in turn would suggest lower mental effort (cognitive load) on the part of the users.

The next section of this paper provides a brief description of the new prototype user interface for proofreading.

Prototype Design

Having done a basic Heuristic Evaluation (Nielsen, 1994) of the main portals currently in use for proofreading, e.g. 18thConnect - TypeWright (2017), Distributed Proofreaders (2017) and Trove -National Library of Australia (2017), it was ascertained that the portals' user interfaces could have been improved. Generally, the main aspects revealed by the evaluation indicated that some of the user interfaces could have been improved concerning their 'aesthetic and minimalist design' due to e.g. unnecessary clutter being present on the screen. Furthermore, 'error prevention' issues could have been improved as e.g. some of the interactions could have been confusing for a user. These are both areas that are known to potentially increase the cognitive load in a user as it requires more mental effort.

It was therefore decided to design a new user interface aiming to reduce clutter and adopt a clear and logical structure with no unnecessary features. This suggests in turn that it would lead to reduced cognitive load. The user interface was developed using an evolutionary prototyping approach. The overall aim was to improve the usability of the user interface when compared with existing portals.

The design of the user interface was guided by the relevant International Organization for Standardization (ISO) standards. This approach was considered to be suitable because to

the author's knowledge there are few published studies where specific design guidelines for proofreading portals are available. The ISO 9241-151:2008 specifically concerns web user interfaces (International Organization for Standardization, 2008). While the ISO 9241-151:2008 tries to cover the whole spectrum of a web site being designed and developed, the prototype presented in this paper concerns one specific aspect of interaction whilst proofreading a digitized text. Therefore some of the guidance in the ISO 9241-151:2008 is not currently relevant to this context. However, there are several points in the ISO 9241-151:2008 that are very relevant and the main ones are discussed in this paper.

The ISO 9241-151:2008 suggests to minimise vertical scrolling and to avoid horizontal scrolling. This can be seen from Figure 1 below, that there is no vertical or horizontal scrolling required in the interaction (Note: if one uses a small screen, some vertical scrolling becomes necessary). This has been achieved by not placing large numbers of editing fields on a single page.

The standard also suggests that designers need to be careful with the use of colour. As can be observed from Figures 1 – 3 below, colours have been kept to a minimum. Also colour is not used in isolation to convey meaning, e.g. the meaning of each respective field is textually labelled as well as colour coded and the 'Save' buttons are also textually labelled and colour coded, depending on their state.

ISO 9241-151:2008 suggests to not abuse the use of white space. This has been achieved, as one can see in Figures 1-3 below, by not having too much white space. Furthermore, the standard suggests that 'navigation links and transactions' should be distinguishable from one another. This has been designed into the prototype by the use of clear unique textual labels and different colours, e.g. the 'Save' button (a transactional button) is blue and the 'Proofread Next Page' button (navigational) is green. The use of colours here also conforms to the standard's guidance on the use of colour.

Another ISO standard guiding the design of the prototype was the ISO 9241-161:2016 (International Organization for Standardization, 2016). This standard deals with the 'visual user interface elements'.

This prototype does not use at this stage of the interaction many different user interface elements. However the ones used were designed to conform to the ISO 9241-161:2016 standard. One example concerns the input fields with a dialogue button. These conform to the ISO 9241-161:2016 by having a 'label, a field with a boundary, a cursor within a boundary and a dialogue button'. In this case the dialogue is a saving action based on some proofreading action.

All labels within the user interface conform to ISO 9241-161:2016, e.g. by being brief and descriptive. The user interface also features a progress indicator. This was designed to conform to ISO 9241-161:2016. The standard states that a progress indicator should have a 'label, a visualisation of progress status and data on extent of progress'. This last aspect is optional within the standard, but it was considered to be useful for this context. ISO 9241-161:2016 further suggests that the progress indicator should be used when the user should be informed about some ongoing process and when the progress of the process is of interest to the user. This fits completely with the context of the prototype because a proofreader will likely want to know how far on they are with proofreading an entire work.

The remainder of this section will now discuss in more detail how the user interface functions. The main user interface is shown below in Figure 1. The prototype assumes that one or more previous stages of digitization have already taken place including segmenting the original text into 'chunks' (a 'chunk' may be viewed as a section of text from a larger body of text) as can be seen in the image below. The user, likely a member of the crowdsourcing community would then perform proofreading with this user interface.

Figure 1: The main proofreading and editing user interface.[i]
([i.] The user interface designs in this paper are the author's own designs and all designs in this paper cannot be reproduced or used without permission from the author of this paper.)

The manner a user proceeds is very simple. The original text version appears above the digitized version, where each version is clearly labelled and appears in a different coloured field. The original text field is not editable, while the digitized version is editable.

A user would proceed by reading the first line of the original text and then comparing this with the digitized version. If no errors are noted, the user clicks 'Save'. The blue save button then changes to white and has a 'Saved' label. At this point the save button is not actionable any more. This is shown below in Figure 2.

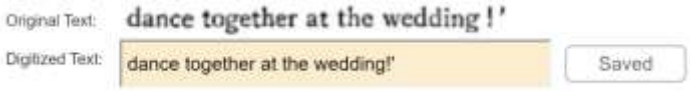

Figure 2: A saved state for a pair of digitized and proofread text.[i]
([i.] The user interface designs in this paper are the author's own designs and all designs in this paper cannot be reproduced or used without permission from the author of this paper.)

If an error is detected, the user corrects the error and then the user clicks 'Save'. As mentioned above, the blue save button then changes to white and has a 'Saved' label. At this point the save button is not actionable any more.

Once the whole page has been proofread, the user can click the 'Proofread Next Page' button. Near this button, one can also see a percentage of how much has been proofread for a particular text (e.g. an old novel).

Further, as can be seen in Figure 1 above, this prototype has the feature that at the beginning and end of a proofreading ‘page’, one can have a quick view of the previous or next three lines of text in the proofreading sequence. This is particularly helpful if context is needed for helping to make a correction. This is particularly so when there is the occasion of the original segmented text being illegible or near illegible. Sometimes the context can help to understand a particular word that may be damaged in some way. This feature is shown in Figure 3 below for viewing the previous three lines of text.

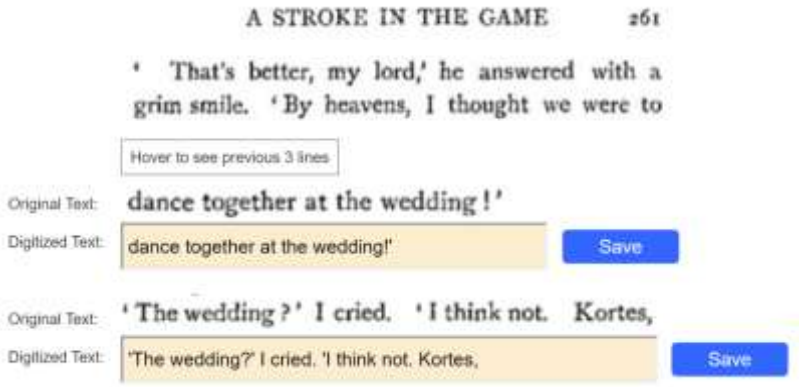

Figure 3: The result of using the hover feature at the top of the proofreading page.[i]
([i] The user interface designs in this paper are the author’s own designs and all designs in this paper cannot be reproduced or used without permission from the author of this paper.)

One aspect that has not been developed in the prototype concerns how a user should deal with the situation of finding an error in the digitized text and not being able to make the correction because, e.g. the part requiring correction is so unclear in the original part by perhaps being too faded with age.

A solution to this can be designed in various ways. One way would be to add another button next to the save button that flags this problem to a coordinator/administrator for further examination. The button would need to be meaningfully labelled. Another option would be to use ‘implication’ within the system and therefore no extra buttons would be needed for the user interface. This could simply be done by the system keeping a record of all the fields that have not been ‘saved’ by a proofreader. That record can then be outputted in a meaningful manner to a coordinator/administrator for further examination.

The second option suggested above, would need to be done anyway, since a user may not save a particular pairing of text for other reasons, e.g. forgetfulness or distraction etc. Therefore, this would be important for the proofreading process, since typically proofreading is done in iterations by different proofreaders.

However, with the first option of adding a further button, the output to a coordinator/administrator could be shown ready filtered, i.e. one output showing the non-saves which possibly suggests forgetfulness or distraction and another output that shows the explicit results of the proofreader having clicked a dedicated button to communicate that there is some unintelligible aspect at that point in the text.

A further aspect needing further testing and not covered by the evaluation described in the next section is the issue that once ‘Save’ is selected, it is not possible to alter this. In reality it may be better to have the option of making a further change(s) after ‘Save’ has been selected. This could be important in the situation where a proofreader has selected ‘Save’ but then realises that there was actually an error (or a further error) that needed correction. One simple solution to this would be to allow reactivation of the editing and saving states by clicking in the digitized text field, thus allowing more editing to be done.

Having designed the new user interface described above, it was then necessary to actually evaluate if this was better than any of the existing user interfaces used for such portals. Various approaches to evaluation could have been used. However, an empirical experimental approach was used, because two of the main measures involved times and errors and this approach was considered to be good for these kinds of measures (see previous section for further details). Nevertheless, personal opinions were also very important and these were elicited by means of a post-experiment questionnaire. The next section begins a series of sections dedicated to describing the evaluation and results obtained.

Experimental Evaluation

Hypotheses

Three hypotheses were used for this research and experiment where the 18thConnect – TypeWright proofreading user interface was compared with the prototype user interface described in the previous section:

H1: In comparing the 18thConnect – TypeWright user interface with the prototype user interface, there will be a statistically significant difference in terms of errors in the proofreading process.

H2: In comparing the 18thConnect – TypeWright user interface with the prototype user interface, there will be a statistically significant difference in terms of task time in the proofreading process.

H3: In comparing the 18thConnect – TypeWright user interface with the prototype user interface, there will be a statistically significant difference in terms of user satisfaction.

Users

A sample of participants was recruited from the student population within the university. These participants were experienced in computer and internet use.

In total 18 participants took part in the experiment. The age range of the participants ranged from 18-45 years. Overall, 13 males and 4 females took part in the experiment, with one participant not declaring their gender. All participants had a minimum of 6 years of active use experience with computers. Furthermore, all participants declared they had a broad range of experience in using different software applications and carrying out diverse internet activities.

Experimental Design

This experiment was carried out using a within users design. This was chosen because it was of interest to have the participants see both user interfaces and then make a comparison of these. Also in this manner users were essentially tested against themselves. This was an important consideration, because some of the measures used in relation to performance would likely differ from person to person. One example of this is that the tasks were timed and involved some typing. Participants would likely type at different speeds and the within users design would better deal with this issue, with the expectation that a slower typing speed on the part of a participant would likely be reflected across both conditions being tested.

Variables

The Independent Variables were the two user interfaces and the tasks associated with the interfaces. The Dependent Variables were performance and user satisfaction. The Dependent Measures were overall task time, errors and perceptual opinions elicited by means of a post-experiment questionnaire.

The errors that were recorded were defined as being an error if the participants did not make a correction within the text that should have been made (see description of the tasks below). Another error that was recorded was if the participant made a 'correction' when there was no correction to be made.

Apparatus and Materials

The materials used were a Windows desktop computer in a quiet laboratory specifically designed for HCI and Universal Design research. The specification of the computer was as follows: 24" LCD monitor (1920x1080 resolution), Windows 7 (SP1) 64 bit operating system, Intel i7 3.6 GHz chip and 32 Gb RAM. The keyboard and mouse used were of a 'standard' size and specification, i.e. not miniature/mini/small. Furthermore, a tablet stopwatch app was used for timing the tasks.

The materials included a recruitment and post-experiment questionnaire. The recruitment questionnaire asked participants to state their age group, gender and to give an idea of the kind of experience they had in using software applications. They were also asked to give an idea of the kinds of internet activities they engaged in. The post-experiment questionnaire asked a series of questions using a Likert-type (Likert, 1932) scale covering the general user interface, the process of proofreading text and the participants' feelings during the interaction.

Further, an information sheet giving the participants details of the study and expectations for participation etc. was given to each participant. An ethical consent form was also used. This outlined the participant rights and understandings for the experiment. Therefore, all participants gave informed consent.

The tasks were designed around the concept of proofreading text. Since this was a within users design, two different English texts of 117 words each were used from out of copyright publications of the kind that real proofreading activities could centre around. It was ascertained that both texts were approximately equivalent to each other in terms of difficulty.

Therefore, participants carried out one basic task under each system, which was to proofread the body of text in each system and correct any errors that they found. Unknown to the participants, each system and text had six errors deliberately introduced into it. The errors were of the kind that could be seen in proofreading text which had been automatically scanned by optical character recognition software. The introduced errors were of the kind where a character within a word was incorrect or missing and/or some element of punctuation was missing. The errors were placed at different positions in the text to avoid any possible learning behaviour from one system to the next.

Procedure

Each participant was treated in the same manner with the aim of minimising possible confounding variables. Initial contact with the participants was established by sending an email explaining the purpose of the study and what they would be expected to do. If in agreement, the participants were then given an appointment suitable to them. Upon physically meeting the participant at the laboratory, they were asked to read a formal

information sheet regarding the research being carried out and if satisfied with the information and still willing to take part, they were asked to read, complete and sign an ethical consent form.

When this stage was completed, they were introduced to the software and instructed to read the digitized text and compare it for correctness with the original version (at this point participants were shown on the screen which was which). Participants were informed that the text may or may not have had errors in it and if they found an error, they were to correct it and save the changes. Then they were to continue with the proofreading. Participants were also informed that navigation around the interface was to be done with the mouse and for any correction of errors, the keyboard was to be used to correct the text.

Whilst carrying out the tasks, the researcher observed and took notes of what was happening. During this time, the following aspects were manually recorded on paper: task times, errors, task completion and any particularly noticeable behavioural aspects on the part of the user (e.g. overt joy or frustration etc.).

Since this experiment deployed a within users design, participants experienced both user interfaces. Each participant was presented with the two experimental conditions in a different order, e.g. Participant 1 would experience and use the new prototype interface first and the existing interface second. Participant two then experienced the user interfaces in the opposite order etc.

When all tasks had been completed by each participant, they were then asked to immediately complete a post-experiment questionnaire which elicited their personal opinions on aspects of their experience with the two user interfaces.

Once the questionnaire was completed, the participants were thanked for their time.

Results

The data collected was firstly examined in a high level manner (details are not included in this paper for brevity). This initial examination showed that the data is not normal. Therefore, it was decided to not use a parametric test. However, the data is suitable for a non-parametric test. In this instance, the Wilcoxon Signed Rank Test is appropriate for the data and experimental approach used (Mayers, A. 2013). Therefore, the Wilcoxon Signed Rank Test was used with all the data collected involving task times, errors and subjective opinions using Likert-type scales. The Likert-type scales in all cases ranged from one to seven. One always represented the most negative opinion and seven always represented the most positive opinion. Further, this section will use abbreviations for the Mean (M) and Standard Deviation (SD).

The Wilcoxon Signed Rank testing showed the following results:

Errors based on the tasks given are not significantly different across the two conditions where the mean ranks are 5.00 and 6.83 (18thConnect-Typewright M = 1.28, SD = 1.45, New Proofreading User Interface M = 1.06, SD = 1.21), $z = -0.726$, $p = 0.47$ and Pearson's $r = 0.18$.

Task time in seconds, is significantly different across the two conditions, where the new proofreading user interface was significantly faster than the 18thConnect-Typewright user

interface where the mean ranks are 0.00 and 9.50 (18thConnect-Typewright M = 444.39, SD = 60.17, New Proofreading User Interface M = 205.56, SD = 66.05), z = -3.72, p < 0.001 and Pearson's r = 0.88.

The coherence of the layout of all the elements on the screen is significantly different across the two conditions, where participants felt that the new proofreading user interface was significantly more coherent than the 18thConnect-Typewright user interface where the mean ranks are 8.54 and 4.50 (18thConnect-Typewright M = 4.89, SD = 1.64, New Proofreading User Interface M = 6.39, SD = 0.70), z = -2.95, p = 0.003 and Pearson's r = 0.70.

The tidiness of the content on the screen is significantly different across the two conditions, where participants felt that the new proofreading user interface was significantly more tidy than the 18thConnect-Typewright user interface where the mean ranks are 8.69 and 14.00 (18thConnect-Typewright M = 4.39, SD = 1.69, New Proofreading User Interface M = 6.33, SD = 1.03), z = -2.98, p = 0.003 and Pearson's r = 0.70.

The pleasantness of the colours used for the user interface is not significantly different across the two conditions where the mean ranks are 7.88 and 5.60 (18thConnect-Typewright M = 5.22, SD = 1.52 New Proofreading User Interface M = 5.78, SD = 1.003), z = -1.26, p = 0.209 and Pearson's r = 0.30.

The readability of the on-screen text is significantly different across the two conditions, where participants felt that the new proofreading user interface's on-screen text was significantly more readable than the 18thConnect-Typewright user interface's on-screen text where the mean ranks are 8.65 and 7.83 (18thConnect-Typewright M = 4.67, SD = 2.03, New Proofreading User Interface M = 6.22, SD = 0.94), z = -2.32, p = 0.02 and Pearson's r = 0.55.

The amount of clutter on the screen is significantly different across the two conditions, where participants felt that the new proofreading user interface had significantly less clutter than the 18thConnect-Typewright user interface where the mean ranks are 8.32 and 3.50 (18thConnect-Typewright M = 4.06, SD = 1.86, New Proofreading User Interface M = 6.06, SD = 0.99), z = -3.24, p = 0.001 and Pearson's r = 0.76.

The overall ease of use of the system is significantly different across the two conditions, where participants felt that the new proofreading user interface was significantly easier to use than the 18thConnect-Typewright user interface where the mean ranks are 9.07 and 4.50 (18thConnect-Typewright M = 4.72, SD = 1.45, New Proofreading User Interface M = 6.56, SD = 0.98), z = -3.09, p = 0.002 and Pearson's r = 0.73.

The perceptions of how easy it was to lose one's place in the text during the process of comparing the digitised text with the original text is approaching significance, with mean scores for the new proofreading user Interface being overall more positive than those for the 18thConnect-Typewright user interface where the mean ranks are 9.71 and 7.30 (18thConnect-Typewright M = 4.78, SD = 1.26, New Proofreading User Interface M = 5.72, SD = 1.53), z = -1.92, p = 0.055 and Pearson's r = 0.45.

The perceptions of how easy it was to correct an error in the digitised text is approaching significance, with mean scores for the New Proofreading User Interface being overall more positive than those for the 18thConnect-Typewright user interface where the mean ranks are 5.67 and 7.50 (18thConnect-Typewright M = 5.72, SD = 1.32, New Proofreading User Interface M = 6.22, SD = 1.56), z = -1.65, p = 0.10 and Pearson's r = 0.39.

The perceptions of how easy it was to save the changes of one's correction of the digitised text is significantly different across the two conditions, where participants felt that the new proofreading user interface was significantly easier for saving changes made to the digitised

text than the 18thConnect-Typewright user interface where the mean ranks are 6.50 and 0.00 (18thConnect-Typewright M = 5.72, SD = 1.23, New Proofreading User Interface M = 6.94, SD = 0.24), z = -3.11, p = 0.002 and Pearson's r = 0.73.

The perceptions about the amount of mental effort required to do the comparison with the digitised text and the original text is significantly different across the two conditions, where participants felt that the new proofreading user interface required significantly less mental effort than the 18thConnect-Typewright user interface where the mean ranks are 8.13 and 7.50 (18thConnect-Typewright M = 4.67, SD = 1.65, New Proofreading User Interface M = 5.78, SD = 1.06), z = -2.21, p = 0.027 and Pearson's r = 0.52.

The perceptions about how easy it would be to remember how to use the system after a period of time of non-use of the system is significantly different across the two conditions, where participants felt that the new proofreading user interface would be easier to come back to than the 18thConnect-Typewright user interface where the mean ranks are 6.00 and 0.00 (18thConnect-Typewright M = 5.72, SD = 1.18, New Proofreading User Interface M = 6.78, SD = 0.43), z = -2.99, p = 0.003 and Pearson's r = 0.71.

The participants' feelings of stress during the interaction with the system is significantly different across the two conditions, where participants felt that the new proofreading user interface elicited less feelings of stress than the 18thConnect-Typewright user interface where the mean ranks are 7.55 and 4.00 (18thConnect-Typewright M = 5.28, SD = 1.74, New Proofreading User Interface M = 6.61, SD = 0.61), z = -2.67, p = 0.008 and Pearson's r = 0.63.

The participants' feelings of frustration during the interaction with the system is approaching significance, with mean scores for the New Proofreading User Interface suggesting overall less frustration than with the 18thConnect-Typewright user interface where the mean ranks are 6.30 and 7.50 (18thConnect-Typewright M = 5.17, SD = 1.78, New Proofreading User Interface M = 6.33, SD = 1.46), z = -1.90, p = 0.57 and Pearson's r = 0.45.

The participants' feelings of enjoyment during the interaction with the system is significantly different across the two conditions, where participants felt that the new proofreading user interface was more enjoyable than the 18thConnect-Typewright user interface where the mean ranks are 6.86 and 7.75 (18thConnect-Typewright M = 4.56, SD = 1.38, New Proofreading User Interface M = 5.50, SD = 1.54), z = -2.13, p = 0.033 and Pearson's r = 0.50.

Discussion and Conclusions

Overall, from the statistical analysis, it is concluded that the new user interface design is more usable than the user interface used in the 18thConnect-TypeWright portal.

For performance, the new user interface was better in terms of speed, where the process of proofreading and correcting errors was significantly faster. Therefore, the initial hypothesis is accepted. It stated that in comparing the 18thConnect – TypeWright user interface with the prototype user interface, there will be a statistically significant difference in terms of task time in the proofreading process. This result also suggests that the aim of producing a design that minimizes cognitive load is being met. Sweller et al (2011) discuss that performance measurements can indirectly indicate lower or higher cognitive load. In this experiment the tasks essentially involved solving the 'problem' of proofreading text and if errors were found to correct the relevant section of text. The new user interface prototype had faster task times.

Fewer proofreading errors were done with the new user interface too, however the differences were not significant. Therefore, the initial hypothesis is not accepted. It stated that in comparing the 18thConnect – TypeWright user interface with the prototype user interface, there will be a statistically significant difference in terms of errors in the proofreading process. It is unclear why the differences are not significant as the expectation would be that with higher cognitive load, the errors may be more than in a condition with lower cognitive load. One aspect that was observed during the experiment was that participants seemed to find the new user interface so easy to use, that they perhaps did not pay as much attention as they should have done to the details of the text. This could explain the lack of significant differences between the two experimental conditions. However as stated above, the new user interface did elicit fewer task errors.

For subjective opinions based on the post-experiment questionnaire which covered the topics of the general user interface, the process of proofreading text and the participants' feelings during the interaction, all the results show clearly that the new user interface design was preferred. Most results were either significantly different or approaching significance. There was only one result that was not statistically significant or approaching significance. This was to do with the pleasantness of the colours used. Therefore, given the overwhelming majority of significantly different opinions, the initial hypothesis is accepted. It stated that in comparing the 18thConnect – TypeWright user interface with the prototype user interface, there will be a statistically significant difference in terms of user satisfaction.

Further, several of the questionnaire items are directly related to cognitive load issues. According to Sweller et al (2011) asking users to rate the mental effort used is considered to be one approach. In addition, other researchers have asked participants to rate how difficult it was to learn a task, as a measure of cognitive load. The questionnaire used in this experiment contained one specific question asking participants to rate the metal effort they felt was required to do the tasks under each condition. In the statistical analysis the outcome was statistically significant showing less mental effort under the prototype user interface.

There were also four other questions in the questionnaire that one would suggest are related to perceptions of difficulty in relation to the experienced interaction and tasks. These were concerning how easy it was to lose one's place in the text during the process of comparing the digitised text with the original text, how easy it was to correct an error in the digitised text, how easy it was to save the changes of one's correction and how easy it would be to remember how to use the system after a period of time of non-use of the system. These questions, in all four cases, suggested the new user interface to be less difficult. Statistically the results were either significant or approaching significance (see previous section for actual results). This further suggests that the new user interface is likely to be incurring less cognitive load.

The approach that was taken in the redesign of the user interface was to remove as much clutter as possible and to have the original text as close as possible to the digitised text. The idea of this was to reduce possible errors, the time taken and cognitive load. While time savings are clearly shown from the data, errors seem to be similar in quantity for the amount of text used in the experiment. A longer test and/or a long-term test may show up more differences in terms of the errors committed in a proofreading session. The new design also aimed to make navigation as simple as possible by having as few controls as possible.

The participants' preferences were clearly in favour of the new user interface. The clutter-free and easy navigation of the user interface clearly elicited more positive feelings and less feelings of 'stress' whilst using the system.

Although intentions are always to strive for a 'perfect' experiment, there could have been some aspects that could have been improved. The sample of participants could have been perhaps larger. Also, although the concept of proofreading includes that users could be from

anywhere in the world and from all walks of life, it may have been interesting to have found actual users already engaged in online proofreading activities. This was attempted initially, but with insurmountable difficulties, leading to the use of university students as participants. Cognitive load has different facets and a future experiment could be designed to try and find which elements of cognitive load are at play with such user interfaces.

As mentioned earlier in the paper, to the author's knowledge there have been very few attempts at producing more usable proofreading portal user interfaces with formal evaluation. While this is just one experiment in this area, it would be good to investigate further the usability of this new design in relation to other portals that are currently in use. It would also be interesting to investigate the usability of the new design in perhaps a more long-term setting. In addition, more emphasis on the universal design aspects would be part of a next stage of work to be carried out.

Lastly, once more evaluation is carried out with the new design and if results continue to be positive, efforts at integrating the new design with existing or new proofreading portals would need to be undertaken. This would need to include persuasion of the relevant owners of such portals to adopt the new design.

Acknowledgements

Prof. Apostolos Antonacopoulos at the University of Salford, is thanked for his advice and many interesting conversations on the topic of digitising old documents.

References

Benbunan-Fich, R. (2001) Using Protocol Analysis to Evaluate the Usability of a Commercial Web Site, Information and Management, 39 (2001), p. 151-163, Elsevier.

Bulacu, M., van Koert, R., Schomaker, L. and van der Zant, T. (2007) Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen, Ninth International Conference on Document Analysis and Recognition, 23-26 Sept.

Chrons, O. and Sundell, S. (2011) Digitalkoot: Making Old Archives Accessible Using Crowdsourcing, Proceedings of the 11th AAAI Conference on Human Computation, P. 20-25, AAAI Press.

Coyle, K. (2006) Mass Digitization of Books, The Journal of Academic Librarianship, 32: 6, Nov 2006, P. 641–645.

The Distributed Proofreaders Foundation (2017) Distributed Proofreaders, https://www.pgdp.net/c/, Accessed March 2017.

Drira, F. (2006) Towards Restoring Historic Documents Degraded Over Time, Proceedings of the Second International Conference on Document Image Analysis for Libraries, 27-28 April.

Estellés-Arolas, E. and González-Ladrón-de-Guevara, F. (2012) Towards an Integrated Crowdsourcing Definition, The Journal of Information Science, Vol. 38: 2, P. 189-200.

Gao, H., Rusiñol, M., Karatzas, D., Antonacopoulos, A., and Lladós, J. (2013) An Interactive Appearance-based Document Retrieval System for Historical Newspapers, Proceedings of the Eight International Conference on computer Vision Theory and Applications, VISAPP13, 2013.

International Organization for Standardization (2008) Ergonomics of Human-System Interaction - Part 151: Guidance on World Wide Web user interfaces (ISO 9241-151:2008).

International Organization for Standardization (2016) Ergonomics of Human-System Interaction – Part 161: Guidance on Visual User Interface Elements (ISO 9241-161:2016).

Irizarry, T., Dabbs, A.D. and Curran, C.R. (2015) Patient Portals and Patient Engagement: A State of the Science Review, Journal of Medical Internet Research, 17: 6 e148.

Ismailova, R. (2017) Web Site Accessibility, Usability and security: a Survey of Government Web Sites in Kyrgyz Republic, Universal Access in the Information Society, 15:1, P. 257-264.

Kobayashi M., Ishihara T., Itoko T., Takagi H., Asakawa C. (2013) Age-Based Task Specialization for Crowdsourced Proofreading. In: Stephanidis C., Antona M. (eds) Universal Access in Human-Computer Interaction. User and Context Diversity. UAHCI 2013. Lecture Notes in Computer Science, vol 8010. Springer, Berlin, Heidelberg.

Lang, A.S.I.D. and Rio-Ross, J. (2011) Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents, The Code4Lib Journal, Issue 15, 2011-10-31.

Likert, R.A. (1932) Technique for the Measurement of Attitudes, Columbia University Press, NY, 1932.

Loranger, H. (2017) Homepage Links Remain a Necessity, https://www.nngroup.com/articles/homepage-links/ Accessed August 2017.

Mayers, A. (2013) Introduction to Statistics and SPSS in Psychology, Pearson.

National Library of Australia (2017), Trove, http://trove.nla.gov.au/, Accessed March 2017.

Nielsen, J. (1994) Enhancing the Explanatory Power of Usability Heuristics, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 24–28, P.152-158, ACM.

Oakley, N.S. and Daudert, B. (2016) Establishing Best Practices to Improve Usefulness and Usability of Web Interfaces Providing Atmospheric Data, American Meteorological Society, 14.3.16.

Oviatt, S. (2006) Human-Centered Design Meets Cognitive Load Theory: Designing Interfaces that Help People Think, Proceedings of the 14th ACM international conference on Multimedia, Santa Barbara, CA, USA, October 23-27, ACM.

Popper, K. R. (1934) The Logic of Scientific Discovery (as Logik der Forschung, English translation 1959).

Rahmanian, B and Davis, J.G (2014) User Interface Design for Crowdsourcing Systems, Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, Como, Italy — May 27-29, P. 405-408, ACM.

Selden, S. and Orenstein, J. (2011) Government E-Recruiting Web Sites: The Influence of e-recruitment Content and Usability on Recruiting and Hiring Outcomes in US State Governments, International Journal of Selection and Assessment, 19:1, March 2011.

Story, M. F. (1998) Maximizing Usability: The Principles of Universal Design, Assistive Technology: The Official Journal of RESNA, 10:1, P. 4-12.

Sweller, J., Ayres, P. and Kalyuga, S. (2011) Cognitive Load Theory, Springer.

Texas A&M University (2017) 18thConnect - Eighteenth Century Scholarship Online, http://www.18thconnect.org/, Accessed March 2017.

Tezza, R., Bornia, A.C., and de Andrade, D.F. (2011) Measuring Web Usability Using Item Response Theory: Principles, Features and Opportunities, Interacting With Computers, Elsevier.

The Centre for Universal Design (1991) Fact Sheet #6 Housing Definitions: Accessible, Adaptable, and Universal Design, HDFS.4.91, 1991, 3 pp.

W3.org (2017a) Introduction to Web Accessibility, https://www.w3.org/WAI/intro/accessibility.php, Accessed August 2017.

W3.org (2017b) Web Content Accessibility Guidelines (WCAG) Overview https://www.w3.org/WAI/intro/wcag.php, Accessed August 2017.

Whitenton, K. (2013) Minimize Cognitive Load to Maximize Usability, https://www.nngroup.com/articles/minimize-cognitive-load/, Accessed August 2017.

Yacoub, S, Burns, J, Faraboschi, P, Ortega, D, Peiro, J.A. and Saxena, V. (2005) Document Digitization Lifecycle for Complex Magazine Collection, Proceedings of the 2005 ACM Symposium on Document Engineering, P. 197-206.

Youngblood, N.E, and Youngblood, S.A. (2013) User Experience and Accessibility: An Analysis of County Web Portals, Journal of Usability Studies, 9: 1, Nov. 2013 P.25-41.

Zaidan, O.F. and Callison-Burch, C. (2011) Crowdsourcing translation: professional quality from non-professionals, HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, P. 1220-1229, Portland, Oregon — June 19 – 24.