Trustworthiness of Wikipedia

Bachelor Thesis Psychology

Koen Remmerswaal

Supervisors:

T. Lucassen, MSc

prof. dr. J.M.C. Schraagen

Date:

29 - 06 -2010

Contents

Abstract

Wikipedia is used increasingly often to find information on a wide variety of subjects. The quality and trustworthiness of Wikipedia articles is however questioned by several researchers. Despite this development however, Wikipedia keeps on growing; it is therefore important to explore how, why and when Wikipedia is used. We will examine 'the how', and will mainly focus on the aspect of trustworthiness of Wikipedia. We do this by manipulating the quality and familiarity of articles, in order to represent the differences users will usually find when reading articles on Wikipedia. A reader will often subconsciously decide how trustworthy information is when examining an article. We will try to tap into this processes of judging the trustworthiness of information on Wikipedia by using the think aloud method, requiring readers to verbalize their thoughts. Using these verbalizations we will show the different features used by readers when judging trustworthiness of Wikipedia articles. Three of the features most used by readers are textual features, references and images.

Introduction

With the introduction of Web 2.0 there has been an increase in use of websites with collaborative, user-generated content, including sites like Wikipedia, Digg, flickr and Youtube. All these sites rely heavily on users to add content to their webpages. Any user can add content on these sites, regardless of location, standing, age and any other factor some websites might have as a requirement. Such a wide range of content creating users has the advantage of enabling these sites to have a wide variety of content available appealing to a large user base. It is however difficult to properly moderate the information and content, thus leading to issues of quality control. For some sites this may be a non-issue, Digg has enabled users to advice other users as to which articles to read, thereby introducing quality control by peers. Wikipedia however, does not have this luxury. Since Wikipedia is an encyclopaedia, it is crucial that information on the site is correct, and to this end Wikipedia has a moderating team that monitors the quality of articles, repairs vandalised articles, and corrects simple factual mistakes. Some mistakes do slip through however, and because the users that add information are anonymous, it may be hard to judge the correctness of information presented. The sometimes questionable quality of articles has its effects on the trust users have in these articles. Trust in articles of Wikipedia is what we focus on in this experiment.

Wikipedia is a large and comprehensive online encyclopaedia, which can be viewed and edited by anyone. Wikipedia has been founded in 2001 and has since grown to more than thirteen million articles, of which more than three million are in English (Wikipedia statistics, n.d.). Every internet user can read articles on Wikipedia for free and has the option to edit articles or create new ones. Because of the amount of articles on Wikipedia and the difficulty of tracking down the source of the information presented in the articles it is often hard to say just how reliable the information in the articles is. This fact has sparked a lot of research into the correctness of information on Wikipedia. In an article by Giles(2005) the quality of Wikipedia is compared with that of Encyclopaedia Britannica, a traditional encyclopaedia in book form. Giles concludes that Wikipedia has only slightly more errors than the respected Encyclopaedia Britannica, this conclusion led to a lot of discussion. Not everyone agreed that Wikipedia could be of high quality (Encyclopaedia Britannica, 2006). Chesney (2006) also came to the conclusion that Wikipedia is of high quality. He let experts judge articles in their field of expertise. They deemed the articles of high quality, suggesting they are indeed factual and correct. Wikipedia keeps growing in popularity amongst regular internet users (Rainie & Tancer, 2007).

Priedhorsky et al.(2006) and Denning, Horning, Parnas & Weinstein (2005) looked into the errors that occurred on Wikipedia and at how they were corrected. They concluded that Wikipedia suffers from factually incorrect information and vandalism; users change, remove or add information that is incorrect on purpose. It is however very possible to combat vandalism and incorrect information with specific tools designed for Wikipedia. Adler, B.T., Benterou J., Chatterjee, K., de Alfaro, L., Pye, I. , and Raman, V. (2007) have suggested that Wikipedia add the option to display the reliability of information with the help of colour coding, that way users can see how trustworthy certain information is and make their own informed decision whether to use it or not. The colour coding has been implemented over the course of this study, but it has not seen enough usage to judge its usefulness.

It is important to define trust and trustworthiness in the context of Wikipedia. The Oxford Dictionary states that trust is “the firm belief in the reliability or truth or strength of an entity”. If the entity in question is a Wikipedia article, then the reader will assign a large amount of trust to an article if he thinks that the information presented is factually correct.

Kittur, Sun and Chi(2008) have shown that quality directly influences perceived trustworthiness of Wikipedia articles. Other methods have been used to rate trustworthiness, which can then be correlated to the quality Wikipedia assigns to articles. In a case study by Dondio, Barrett, Weber and Seigneur (2006) a model was developed in order to predict trustworthiness of an article based on criteria a computer program could easily retrieve. Such as article length, number of edits and the amount of discussion about an article. They found that the model could quite easily make the distinction between high quality and low quality articles. McGuinness, Zeng, da Silva, Ding, Narayanan and Bhaowa (2006) have suggested an option to increase the visible trustworthiness of Wikipedia similar to that of Adler et al. (2007), with the help of colour coding, but calculated in a different manner. They use the number of times an article is linked to by other Wikipedia articles and the times that the topic of an article is mentioned but not linked to. With these two measures they could calculate 'link-ratio'. When comparing link-ratio among articles in a similar category, like countries, or food, it is possible to calculate relative levels of trust between two or more of these articles.

Trust is highly related to quality, if one could be trained to rate trustworthiness of Wikipedia articles perfectly, one would in fact measure mostly quality. Trust mostly aims to predict quality; therefore we use the quality of Wikipedia articles as a baseline for expected trustworthiness. Our experiment isn't concerned with predicting trust readers have in articles, but focuses instead on how readers decide how trustworthy an article is they are reading. Participants will be allowed to use any method they want to come to their conclusion about the trustworthiness of the articles.

The Wikipedia Editorial Team (WET) assesses quality in Wikipedia articles, the WET judges quality of articles manually and does this according to certain criteria (Wikipedia's WET rating, n.d.). This leads to it being ranked into one of seven categories (see Table 1). Table 1 describes the experience a reader will have when browsing an article of a certain quality, it gives a good idea of what an article will include. Wikipedia uses a more detailed and specific set of requirements for each ranking, looking at style, length, proper use of images, among other factors (Wikipedia's WET rating, n.d.).

Article Status	Reader's Experience of Article
Featured Article	Professional, outstanding, and thorough; a definitive source for encyclopaedic information.
A Class Article	Very useful to readers. A fairly complete treatment of the subject. A non-expert in the subject matter would typically find nothing wanting.
Good Article	Useful to nearly all readers, with no obvious problems; approaching (although not equalling) the quality of a professional encyclopaedia
B Class Article	Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.
C Class Article	Useful to a casual reader, but would not provide a complete picture for even a moderately detailed study.
Start Class Article	Provides some meaningful content, but the majority of readers will need more.
Stub Class Article	Provides very little meaningful content; may be little more than a dictionary definition.

Table 1. Wikipedia Editorial Team Article Rating with Reader's Experience of Article

Users however do not use this exact set of criteria when choosing to trust an article or not, but quality will be a huge influence as shown earlier (Kittur et al. 2008), it is therefore expected that readers will at least use some of these aspects when judging articles. It remains unclear how users of Wikipedia judge the trustworthiness of articles, which is what this research paper will focus on. The main research question is: *Which aspects, and to what degree, influence the perception of trustworthiness of a Wikipedia article?*

We assess this using the think aloud method (Ericsson & Simon, 1984), using the think aloud method participants are asked to rate several Wikipedia articles on trustworthiness. The think aloud protocol requires them to verbalize their thoughts as much as possible. These verbalizations represent the information being processed in working memory and how this is used to comprehend and apply this information (J. P. Trabasso & Magliano, 1996a). Several research papers have shown that the verbalizations by subjects are indeed a valid

representation of the information used while comprehending what is being read (Cote & Goldman, 1999).

These verbalizations enable us to find aspects often used by readers and compare them to aspects the Wikipedia Editorial Team(WET) uses for quality control. We will be comparing how the WET rates articles versus how lay readers rate articles. Lay readers are expected to use some of the aspects the WET uses in their judgment of trustworthiness of an article, but they will not use all of them and lay readers will also use different criteria. Revealing the differences between lay people's ratings and the WET ratings helps us understand how lay readers judge Wikipedia articles. The WET can be considered experts on rating Wikipedia articles.

Hypothesis 1: *The features used by lay Wikipedia readers overlap with those of the Wikipedia Editorial Team, but different features will also be used.*

Articles of poor quality are rated significantly different from those of high quality (Kittur et al., 2008). Articles of lesser rating usually include less information and the information is of lower quality, they are also lacking in several aspects that featured articles have included. It is therefore likely that Wikipedia readers will use a different set of features depending on the WET rating of articles.

Hypothesis 2: *The features used by lay Wikipedia readers differ for articles of good and poor quality.*

A difference in features used for positive and negative comments is expected, some features will appear more often for positive comments, verifying information as being correct will happen more often than concluding information is incorrect. Some negative comments might also be used more often.

Hypothesis 3: *The features used by lay Wikipedia readers differ for positive and negative comments on an article.*

Because a participant is familiar with information presented in familiar articles he will be able to quite directly gauge the trustworthiness of an article; he is able to asses whether

information is correct more directly. This is not the case for unfamiliar articles; he has to rely on a different method of rating trustworthiness. Verifying correctness of information presented will often not be possible for unfamiliar topics. The difference in method will result in a different set of features being used.

Hypothesis 4: *The features used by lay Wikipedia readers differ for articles on familiar and unfamiliar topics.*

Verifying information present in an article will take time and the subjects will tend more specifically to familiar information. It is expected that assessing a familiar topic will consume more time.

Hypothesis 5: *People take more time to assess articles on familiar topics than on unfamiliar topics.*

When the familiarity of a topic is high it is expected that lay persons will rate the article higher on trustworthiness. When reading a familiar article readers will be able to verify the information presented with the knowledge they have available in memory. If confirmed the information will result in estimating the trustworthiness of the article as higher compared to an unfamiliar article. The quality and correctness of information on Wikipedia is high Chesney(2006). Confirming correctness is therefore more likely then concluding information is incorrect. This results in higher trustworthiness ratings for familiar topics. A study by Chesney(2006) has found results that leads to the same conclusion. Both experts and non-experts were asked to rate the same article. Experts rated the article significantly higher on credibility than non-experts. According to Chesney these results are due to the high quality of articles on Wikipedia and the fact that experts are able to notice the low amount of errors in the articles. These results suggest that the same will be true for our experiment. Instead of using experts versus non-experts we will use articles that participants are familiar with and articles that participants are not familiar with. The familiar topics are most likely similar to the expert condition used by Chesney (2006), leading to higher ratings than the unfamiliar topics.

Hypothesis 6: *The trustworthiness ratings are higher on articles on familiar topics than on unfamiliar topics.*

Methods

Subjects

Twelve participants' data was used for the data analysis(aged 20 to 44), of which 5 were male, all from the faculty of behavioral sciences at the University of Twente, participating for credits required to complete their education. No specific demands were set for subjects to participate. Three subjects were excluded from analysis, two due to poor performance on the think aloud method, one due to technical issues with the audio recording equipment.

Task

In the experiment a subject was given ten Wikipedia articles with the instruction to rate them on trustworthiness. The participant was required to verbalize their thoughts using the Think Aloud method. When the subject finished with the article he was required to fill in a questionnaire with their final rating of the article.

Design

A 2 (familiarity) x 6 (article quality) design was used. Both familiarity and article quality are a within subjects factor. The order of familiarity was alternated, beginning with a familiar article. The order of article quality was randomized.

Procedure

After a subject had enlisted to participate in the experiment he was contacted by phone. He was then asked for a short list of subjects he was familiar with and a short list he was unfamiliar with. The participant was coached as little as possible concerning subjects, to eliminate any bias relating to the chosen topics. This list of topics was then used to find articles on Wikipedia, to be given to the participant to rate on trustworthiness.
When the participant arrived for the experiment he filled in a short questionnaire. After this initial questionnaire he was given a sheet of paper on which the experiment was explained,

after he has read this information he was asked if he fully understood it and if he had any questions or objections concerning the experiment.

He then practiced with two practice articles, to make sure he understood the Think Aloud Method and how to fill in the questionnaire. If required the subject was coached in using the Think Aloud Method. Coaching was kept to a minimum.

After the participant had finished with the two practice articles the participant started with the actual experiment articles. Ten articles were be presented to be rated on trustworthiness. The order of the articles was alternated between familiar and unfamiliar topics, starting with a familiar one. The order of articles within the familiar group and unfamiliar group is randomized. After each article the participant was required to fill in a questionnaire with their rating of trustworthiness of the article, this was the same questionnaire as the one for the practice articles.

When the participant was finished with judging all twelve articles he was given a final questionnaire. After he has filled it in the experiment was concluded and he was given the option to ask any questions he might have related to the experiment. The experiment lasted approximately 90 minutes.

Materials

The two articles used for practice were the same articles for every participant. The ten other presented articles were picked on the basis of the list of topics they had given on which they were either familiar or unfamiliar. Five articles are picked for the familiar condition and five for the unfamiliar condition. The quality of articles varies. The quality rating of the Wikipedia Editorial Team (WET) (Wikipedia's WET rating, n.d.) was used, there are seven levels of quality on Wikipedia, there is no overlap in quality within the familiar or unfamiliar conditions for each subject. All articles were adapted in such a way that there is no direct reference to quality, such as requests from editors to change articles to include or exclude specific information, or 'citation needed' after a piece of information. See image 1 for an example of a request to change an article.

Image 1. Cleanup template, requesting a review of a specific part of text

Preceding the experiment the participants were given a sheet of paper with information (Appendix A) about what to do in the experiment, this sheet includes a short explanation of the Think Aloud Method and explains participants that they are required to rate articles on trustworthiness.

Three different questionnaires were used, one before the experiment starts (Appendix B), one after every article (Appendix C) and one upon conclusion of the experiment(Appendix D). The first questionnaire looked into demographic information and asks the participants about their familiarity with Wikipedia and how often and in what way they used Wikipedia. After every article they are asked to fill in how trustworthy they find the article and how familiar they are with the subject of the article on a seven point Likert scale. They are also given the option to fill in which aspects positively or negatively influenced their rating of the reliability. The final questionnaire included several control questions about the manipulations, quality and familiarity of the participants with the topics.

A microphone is used to record everything participants say once they have started reading the first article. The articles are presented on a 17” monitor; participants use a mouse to control the computer.

Analyses

Think-aloud Protocol

Subjects were instructed on how to perform the think aloud method before commencing the experiment and were given two articles with the purpose of them practicing with the think aloud method. Everything the participants said during the experiment was recorded. The sound file was later hand-coded. Everything the participant and the experimenter said was included in this protocol. This was all coded as literal as possible.

Independent variables

Familiarity has two categories, familiar articles and unfamiliar articles. The articles were selected after a short interview with each subject, which took place a few days before the actual experiment.

Quality has six categories. Six out of seven quality ratings of the WET were used (see Table 1). Articles of A-quality were not used. Since too few of this class exist on Wikipedia(only 0.03% of all rated articles).

Dependent variables

Five dependent variables are measured: Protocols generated after each trial, trustworthiness ratings, motivations for the trustworthiness, familiarity ratings and trial duration, the time it takes to complete a single article.

The experiment was recorded on audio. This audio was later typed into plain text. This text was coded, with the coding scheme being decided upon using pilot experiments and refining it based on participants' usage of features.
Protocols were analyzed after averaging percentages for each category in favour of using absolute numbers for each category. This was done to ensure so that each subject had the same amount of influence on the total results, some subjects use more features then others.

After each trial the subject was handed a questionnaire to fill in, this questionnaire included three variables.
In order to obtain the trustworthiness ratings the subject was asked to fill in how trustworthy he thought an article was on a seven-point Likert scale.
Subjects were also requested to fill in motivations they had for the trustworthiness of the judged article, they could write out features they thought had influenced their judgment on an article, this could either be positive or negative.
In order to ensure that an article was indeed familiar or unfamiliar the subjects were asked how familiar they were with the topic they had just read about. This was filled in on a seven-point Likert scale, ranging from very unfamiliar to very familiar.

Trial duration was also recorded for each trial. Familiar articles can then be compared to unfamiliar articles on the time it takes to complete them.

Inter-rater reliability

In order to ensure that both raters had an acceptable amount of agreement when rating the protocols both raters rated a single protocol the other rater had already rated. These protocols were then compared to determine in what aspects the rating differed, so that the raters could then decide which aspects of the rating went wrong, and how to change the way of rating so that it would be more consistent. Comparing the protocols also improved the definitions used in the coding scheme.

With the method of rating agreed upon the protocols could be re-rated in order to determine Cohen's kappa(Cohen, 1960). Cohen's kappa was used because we used two raters and because all the data was highly categorical. Three different kappa's were calculated, one which included valence and subcategory, one which included subcategories without valence and the last one looking only at the main category's. This resulted in a kappa of 0.427, 0.491 and 0.579, which considering the large amount of categories, 44 subcategories and 131 when direction is taken into consideration, should be taken as moderate to substantial agreement between raters.

		All	Familiar	Unfamiliar	Positive	Neutral	Negative
A	**Appearance**	**4,97%**	**4,99%**	**4,95%**	**8,08%**	**2,66%**	**4,95%**
A1	General	45,61%	53,57%	37,93%	68,57%	18,18%	0,00%
A2	Structure	54,39%	46,43%	62,07%	31,43%	81,82%	100,00%
C	**Table of Contents**	**4,62%**	**5,53%**	**3,75%**	**7,62%**	**4,84%**	**0,00%**
C1	General	79,25%	83,87%	72,73%	72,73%	90,00%	0,00%
C2	Length	11,32%	12,90%	9,09%	12,12%	10,00%	0,00%
C3	Structure	7,55%	0,00%	18,18%	12,12%	0,00%	0,00%
C4	Contents	1,89%	3,23%	0,00%	3,03%	0,00%	0,00%
F	**First alinea**	**5,06%**	**5,17%**	**4,95%**	**5,08%**	**6,30%**	**4,50%**
F1	General	43,10%	37,93%	48,28%	4,55%	92,31%	0,00%
F2	Length	13,79%	10,34%	17,24%	9,09%	0,00%	60,00%
F3	Clarity	20,69%	20,69%	20,69%	45,45%	0,00%	20,00%
F4	Contents	22,41%	31,03%	13,79%	40,91%	7,69%	20,00%
H	**History section**	**3,57%**	**3,92%**	**3,24%**	**2,54%**	**6,30%**	**1,80%**
H1	General	65,85%	63,64%	68,42%	9,09%	96,15%	25,00%
H2	Length	12,20%	22,73%	0,00%	27,27%	0,00%	50,00%
H3	Clarity	7,32%	4,55%	10,53%	27,27%	0,00%	0,00%
H4	Contents	14,63%	9,09%	21,05%	36,36%	3,85%	25,00%
I	**Infoboxes**	**1,39%**	**1,07%**	**1,71%**	**0,92%**	**2,91%**	**0,00%**
I1	General	75,00%	83,33%	70,00%	25,00%	91,67%	0,00%
I2	Relevance	6,25%	0,00%	10,00%	25,00%	0,00%	0,00%
I3	Clarity	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
I4	Overview	18,75%	16,67%	20,00%	50,00%	8,33%	0,00%
L	**Lists/Tables**	**2,70%**	**2,50%**	**2,90%**	**1,85%**	**4,84%**	**1,35%**
L1	General	87,10%	92,86%	82,35%	62,50%	100,00%	66,67%
L2	Relevance	3,23%	0,00%	5,88%	12,50%	0,00%	0,00%
L3	Clarity	3,23%	0,00%	5,88%	0,00%	0,00%	33,33%
L4	Overview	6,45%	7,14%	5,88%	25,00%	0,00%	0,00%
P	**Pictures**	**12,55%**	**11,23%**	**13,82%**	**10,39%**	**18,89%**	**9,46%**
P1	General	40,28%	41,27%	39,51%	17,78%	64,10%	0,00%
P2	Relevance	19,44%	25,40%	14,81%	26,67%	7,69%	47,62%
P3	Captions	1,39%	0,00%	2,47%	2,22%	1,28%	0,00%
P4	Quality	26,39%	19,05%	32,10%	48,89%	11,54%	33,33%
P5	Number of pictures	12,50%	14,29%	11,11%	4,44%	15,38%	19,05%
R	**References**	**26,07%**	**23,71%**	**28,33%**	**21,94%**	**35,59%**	**25,68%**
R1	General	34,45%	39,10%	30,72%	18,95%	51,02%	17,54%
R2	Relevance	4,01%	6,02%	2,41%	9,47%	0,68%	3,51%
R3	Quality	24,75%	18,05%	30,12%	30,53%	20,41%	26,32%
R4	Number of references	36,79%	36,84%	36,75%	41,05%	27,89%	52,63%
IL	**Internal links**	**5,84%**	**5,88%**	**5,80%**	**4,62%**	**9,93%**	**2,70%**
IL1	General	64,18%	60,61%	67,65%	50,00%	75,61%	33,33%
IL2	Relevance	10,45%	18,18%	2,94%	15,00%	4,88%	33,33%
IL3	Quality	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
IL4	Number of internal links	25,37%	21,21%	29,41%	35,00%	19,51%	33,33%
T	**Text**	**26,33%**	**31,37%**	**21,50%**	**36,95%**	**7,75%**	**49,55%**
T1	General	0,33%	0,57%	0,00%	0,63%	0,00%	0,00%
T2	Scope	4,97%	6,25%	3,17%	3,13%	6,25%	7,27%
T3	Writing style	5,63%	5,11%	6,35%	3,13%	9,38%	8,18%
T4	Neutrality	4,64%	3,98%	5,56%	4,38%	0,00%	6,36%
T5	Clarity	9,93%	5,11%	16,67%	10,63%	3,13%	10,91%
T6	Comprehensiveness	24,17%	18,18%	32,54%	18,13%	12,50%	36,36%
T7	Correctness	37,75%	51,70%	18,25%	55,00%	12,50%	20,00%
T8	Length	12,58%	9,09%	17,46%	5,00%	56,25%	10,91%
XX	**Other**	**6,89%**	**4,63%**	**9,04%**			
	Percentage of total	**100,00%**	**50,31%**	**49,69%**	**40,19%**	**33,94%**	**18,80%**
	Number of counts	**1147**	**561**	**586**	**433**	**413**	**222**

Table 2. Coding Scheme and feature usage distribution as split for each condition

It was felt that correlation could be improved, so protocols were reviewed once more, and agreed upon very specific ways to rate conditions and valence, using this all protocols were looked over again, and possibly changed, to ensure they were properly rated. Over these final protocols another Cohen's kappa was calculated, this time it resulted in a kappa of 0.792, which is very high considering the amount of categories available.

Results

Coding scheme

The final coding scheme that was used can be seen in Table 2, as well as the percentages within each category. Percentages for each main category can be seen in Figure 1.
Textual features, references and pictures stand out as much used features, ranging from 12% to 26% usage.

WET features versus Subject features

The Wikipedia Editorial Team uses a list of features to decide whether or not an article is of high enough quality to be a featured article (Wikipedia's WET rating, n.d.), Wikipedia takes into account the following features: An article should be well-written, comprehensive, well-researched, neutral and stable. It has a lead, appropriate structure and consistent citations. Images are relevant and used when necessary, the length of the article is also appropriate for its topic.
Most of these features are also used by subjects, although they are not necessarily named the same as the WET features. Most features can be found in the text subcategories, although the stable feature can't be found anywhere in the features used by subjects, subjects didn't appear to use that feature. Two subjects mentioned the edit date of an article a few times, which is arguably similar to stability, because of the low frequency of usage of this feature it has been classified as 'other'. Subjects looked at several specific areas which the WET does not consider specifically when rating an article on quality, the history section is often mentioned as is the table of contents.

Features used, poor versus high quality

There appear to be no significant differences between the six different levels of Wikipedia Editorial Team article quality used, χ^2 (40, N = 1069) = 43.52, p = .324.

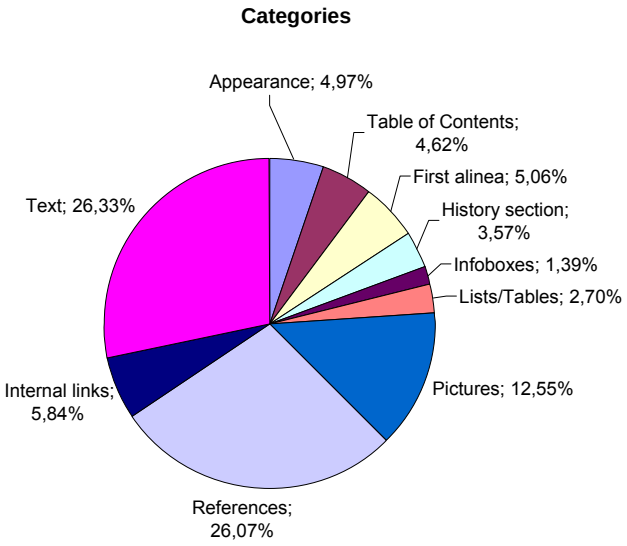

Figure 1. Percentages feature categories used

Features used, positive versus negative comments

A significant difference between positive and negative comments about features used was found, χ^2 (40, N = 655) = 111.80, p = .000. Further manual inspection of the data is used to compare categories(see figure 2 and 3). There are two features that stand out, correctness and comprehensiveness. Correctness accounts for 55% of all textual features mentioned for positive comments versus only 20% for negative comments. Comprehensiveness is used in 36% of negative textual features while it only accounts for 18% of positive textual features. Several features are almost exclusively used as neutral comments, mainly infoboxes and table of contents, while textual features are almost always either positive or negative.

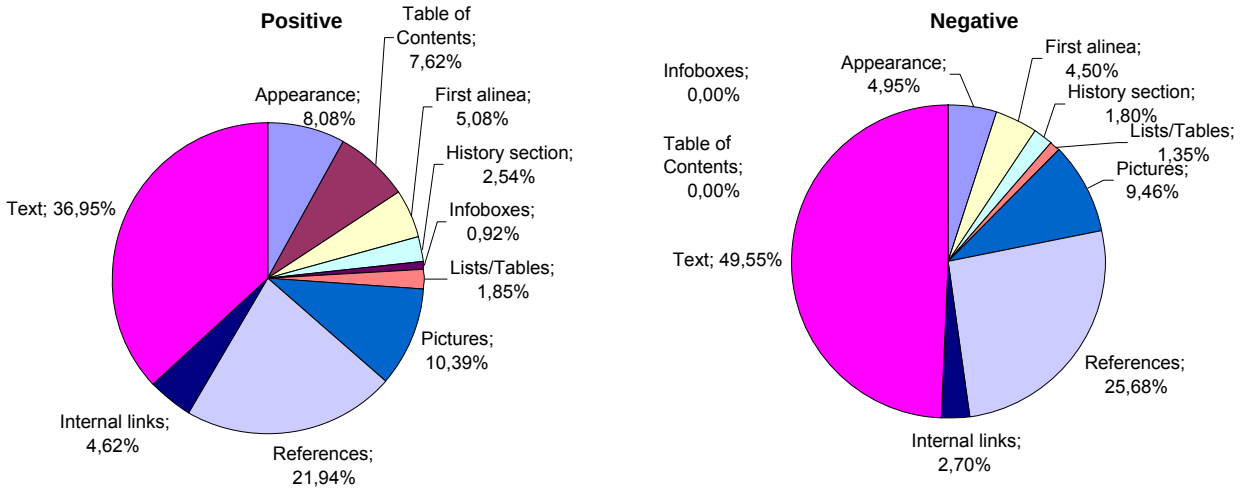

Figure 2 and 3. Features used for Positive(left) and Negative(right) comments

Features used, familiar versus unfamiliar articles

There is significant difference in features used for familiar and unfamiliar articles, χ^2 (40, N = 1068) = 98.81, p = .000. Familiar articles have a higher percentage of textual features mentioned(see Figure 4 and 5), mainly thanks to the correctness feature, although the comprehensiveness is used more for unfamiliar articles, the effect is almost the same as the effect found for positive and negative comments.

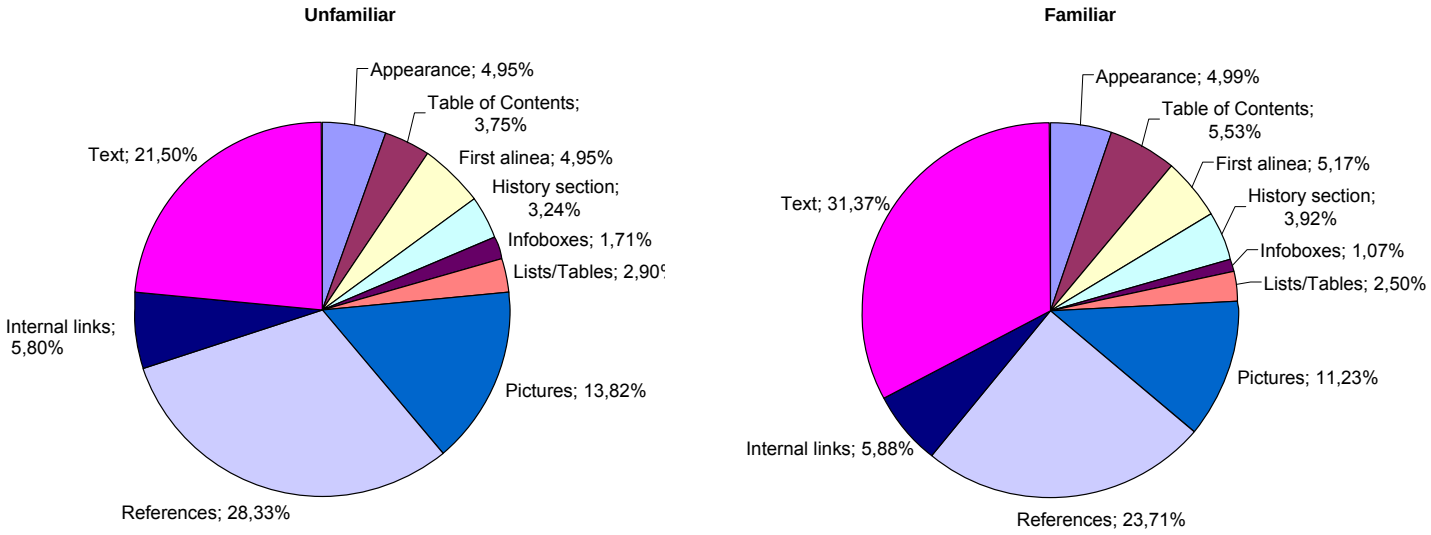

Figure 4 and 5. Features used for Unfamiliar (left) and Familiar (right) articles

Trustworthiness ratings

There is no significant difference in trustworthiness ratings of articles on familiar topics and unfamiliar topics; Z = 1.26; p = 0.21. There is a significant difference between high and low WET quality ratings on trustworthiness, Z = 2.67; p = 0.008.

Trial duration

The duration in seconds of the trials with articles on familiar topics (M = 263; SD = 124) was not significantly different from those of unfamiliar topics (M = 250; SD = 117); t(116) = 0.62; p = 0.54.

Discussion

The current study investigates which aspects, and to what degree, influences the perception of trustworthiness of a Wikipedia article. We did this by letting participants rate articles on trustworthiness while using the think aloud method. The resulting protocols were then analyzed and features used by participants for each article were calculated.
From the results we can conclude that the three most important feature sets are text (26%), references (26%) and pictures (13%). Text being very important is expected, it is a broad category that includes a wide variety of features such as writing style, comprehensiveness and length. Especially correctness is often mentioned. Correctness is mainly used in a positive manner, the participant recognizes information and is able to verify it as being correct. The Wikipedia editorial team (WET) also looks at a lot of textual features when rating quality. Images refer to participants finding the images helpful, useless, good looking or anything else related to the images in an article. Images are included in almost every article we presented to participants. Since articles of class B quality or higher are expected to have pictures, articles of lesser quality will also often include images. Most participants that notice an image will mention it, so it is guaranteed to be a fairly often used feature. References includes any mention of the participant of a reference, or the lack thereof in an article. A 26% share is on the high side, most university students have learned that references are important and that they should be included if you introduce information that the reader might want to verify. Most other people on the other hand, or high school students for example, will probably not pay as much attention to references, because they are not trained to do so. Future research could focus on the difference between secondary school students and university students, to see if a different set of used features emerges. We suspect that references especially will show a large difference, being used far less by the secondary school students compared to the university students.

There is a huge amount of overlap with a few minor differences in the features used by lay Wikipedia readers and the WET, as can be found in the introduction. It is unclear if the same features to judge trustworthiness are used if a reader is not actively deciding trustworthiness. The effort our participant put in to deciding trustworthiness results in refined strategies comparable to those being used by the Wikipedia Editorial Team. The ‘stable’ requirement is not on the list of features used by participants and well written is never explicitly mentioned

by subjects. Length is not always used in the same way as it is by the WET, participants notice the length of the text and seem more impressed by a long article than by a short one, regardless of content. The WET however requires that an article does not contain unnecessary information, while participants did not pay special attention to this, an explanation could be because they are unable to judge whether information is required or not. The first hypothesis is therefore accepted.

Significant differences between quality levels haven't been found, and the minor insignificant trends that are found, such as the high amount of textual features found for stubs, are hard to explain. Further research into the differences between quality levels will be required if the hypotheses is to be answered. Research focused specifically on the differences between the features in different quality levels, with a different design setup to allow for this, could lead to results that this study was not able find. The second hypothesis is rejected.

There is a significant difference between features used for positive comments and features used for negative comments; a significant difference between familiar and unfamiliar articles has also been found. Features used are mostly similar for positive and negative comments, as well as for familiar and unfamiliar articles, except for two notable exceptions. First, positive comments on familiar topics include a large amount of the correctness feature. This is because of subjects verifying information they find in a familiar article which due to the low amount of errors on Wikipedia will usually result in a positive comment about the correctness of information presented. Secondly, negative comments on unfamiliar topics include a relatively large amount of comprehensiveness features, this is probably due to subjects having trouble interpreting or understanding information presented in an article they are not familiar with, resulting in them criticizing the information presented, or mentioning the lack of what they deem to be proper explanations of information presented. Manipulating the correctness and comprehensiveness of articles in an experiment can lead to interesting results, participants might not see any problems with comprehensiveness when it is changed for articles they are familiar with and vice versa for correctness. The third and fourth hypotheses are both accepted.

Participants do not take significantly longer to assess familiar articles, contrary to what our hypothesis suggested. Correctness did appear a lot in familiar articles, so they might spend

more time verifying information, but this might be offset by the time comprehending information for unfamiliar articles. The fifth hypothesis is rejected.

There is no significant difference between trustworthiness ratings on familiar and unfamiliar topics. No difference in rating suggests there is no positive bias due to familiarity. The sixth hypothesis is rejected.

With the hypotheses answered we can look at the main research question: '*Which aspects, and to what degree, influence the perception of trustworthiness of a Wikipedia article?'*

Both hypotheses related to the quality manipulation used in this experiment have been rejected. Quality however does correlate with trustworthiness ratings, so quality does influence a reader's perception of trustworthiness. It is unclear how the quality influences trustworthiness ratings, since there is no evidence it does so via features.

The reason for this is unclear, more research into this area could lead to answers regarding the effects of quality. The lack of effect could be because quality of an article doesn't necessarily directly the factual correctness of an article, merely the way in which the correctness is conveyed. Participants might be able to consciously, or unconsciously, be able to not take judge information on the way it is conveyed. They might also not be able to judge the quality of an article.

A difference in familiarity with the topic of a given article leads to different features being used for deciding how trustworthy the article is. Familiarity does not appear to affect the amount of time it takes to complete an article. A different strategy is likely used depending on familiarity with the information presented, if it is familiar a participant will try to verify the truthfulness of the presented information. If it is unfamiliar a participant tries to comprehend the information, looking at the way information is displayed, whether the information is presented in such a manner to make it understandable. If the familiar information appears to be correct he will most likely increase his trust in the article; if the unfamiliar information is presented in a comprehensible manner he will also increase his trust in the article.

A lot of different features are used in deciding the trustworthiness of Wikipedia articles, especially textual features, references and pictures. All these features combined lead to a

rating of trustworthiness of an article, which features are used is dependant on the article itself, and on the familiarity of the participant with the article topic. Quality appears to have no effect on the way features are used, it does have an effect on the amount of trust readers place in an article.

References

Adler, B.T., Benterou J., Chatterjee, K., de Alfaro, L., Pye, I. , and Raman, V. (2007). Assigning trust to wikipedia content. *Technical Report UCSC-CRL-07-09*, School of Engineering, University of California, Santa Cruz, CA, USA, 2007.

Chesney, T. (2006). *An empirical examination of Wikipedia's credibility*. First Monday, 11(11). Retrieved May 25, 2009, from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1413/1331

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* Vol.20, No.1, pp. 37–46.

Coté, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 169-193). Mahwah, NJ: Erlbaum.

Denning, P., Horning, J., Parnas, D.,& Weinstein, L. (2005).Wikipedia risks. *Communications of the ACM*, 48(12), 152.

Dondio, P., Barrett, S., Weber, S., and Seigneur, J. (2006). Extracting trust from domain analysis: A case study on the Wikipedia project. *Autonomic and Trusted Computing* 362–373.

Encyclopedia Brittannica. Fatally flawed: Refuting the recent study on encyclopaedic accuracy by the journal nature, March 2006.

Ericsson, A. K., and Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data.* The MIT Press, 1984.

Giles, J. (2005). *Internet encyclopaedias go head to head*. Retrieved 15th December, 2009, from http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

Kittur, A., Suh, B., Chi, E. H. (2008). Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia, *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, November 08-12, 2008, San Diego, CA, USA

McGuinness, D. L., Zeng, H., da Silva, P., Ding, L., Narayanan, D., and Bhaowal, M. (2006). Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study. In the *Proceedings of the WWW2006 Workshop on the Models of Trust for the Web (MTW'06)*, Edinburgh, Scotland.

Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen,L., and Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. *Proc GROUP 2007*, ACM Press (2007), 259-268.

Rainie, L., & Tancer, B. (2007). *Wikipedia users*. Washington, DC: Pew Internet & American Life Project. Retrieved May 25, 2009, from http://www.pewinternet.org/□/media//Files/Reports/2007/PIP_Wikipedia07.pdf.pdf

Stvilia, B., Twidale, M., Smith, L.C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59, 983–1001.

Trabasso, T., and Magliano, J. P. (1996a). Conscious understanding during comprehension. *Discourse Processes*, 21, 255-287.

Wikipedia: Statistics (n.d.). Retrieved January 25th, 2010, from http://en.wikipedia.org/wiki/Special:Statistics

Wikipedia:Version 1.0 Editorial Team/Assessment(n.d.). Retrieved January 25th, 2010, from http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

Appendix A

Uitleg van het experiment

Gedurende dit experiment krijg je een aantal Wikipedia artikelen te zien. Stel jezelf voor dat je deze artikelen hebt gebruikt bij het schrijven van een essay. Hiervoor is het natuurlijk van belang dat je een beeld hebt van de betrouwbaarheid van de artikelen.

Er wordt van je gevraagd om een oordeel te geven over de **betrouwbaarheid** van de artikelen. Hoe je deze beoordeling maakt is aan jou. Je hebt onbeperkt de tijd om de artikelen te bestuderen. Let op: Er wordt niet gevraagd om de relevantie van de artikelen te beoordelen. Daarnaast krijg je ook geen inhoudelijke vragen over de artikelen. Alleen de betrouwbaarheid is in dit onderzoek van belang.

De experimentleider zal vertellen welk artikel op welk moment geopend mag worden. Het is niet toegestaan om te klikken op de pagina of om naar een andere pagina te gaan. Je mag wel scrollen om het hele artikel te kunnen bekijken. Je mag zelf aangeven wanneer je klaar bent met een artikel. Na elk artikel krijg je een korte vragenlijst over jouw oordeel over de betrouwbaarheid ervan. In totaal krijg je tien artikelen te zien. Indien gewenst kan er nog een korte pauze worden ingelast.

Terwijl je de taak uitvoert, zeg je alles wat je denkt, leest of doet hardop. Tijdens het invullen van de vragenlijsten hoef je dit niet te doen. Praat tijdens het experiment zo min mogelijk met de experimentleider.

Het hele experiment wordt opgenomen. Het verkregen materiaal is alleen voor analyse doeleinden bestemd. Bij rapportage van de resultaten van het experiment wordt je privacy beschermd in die zin dat het niet mogelijk zal zijn op enigerlei wijze jouw identiteit te achterhalen.

Het experiment zal ongeveer een uur duren. Je krijgt nu eerst de gelegenheid om even te oefenen met het uitvoeren van de taak en het hardop denken. Je krijgt hiervoor twee voorbeeldartikelen te zien. Deze oefening zal niet worden meegenomen in de resultaten.

Appendix B

PP#
……..

Questionnaire vooraf

Voordat we aan het experiment beginnen worden enkele vragen gesteld over jezelf, Wikipedia en vertrouwen.

Geboortedatum:	……
Geslacht:	M / V
Nationaliteit:	……
Opleiding:	Bachelor / Master, jaar:……

1. Hoe lang geleden heb je Wikipedia leren kennen?

…… jaar

2. Hoe vaak maak je gebruik van Wikipedia?

Iedere dag	Iedere week	Iedere maand	Ieder jaar
☐	☐	☐	☐

3. Indien je bij vraag 2 "Iedere dag" hebt gekozen: Hoeveel uren besteed je per dag aan het gebruik van Wikipedia?

Meer dan 4 uur	Meer dan 2 uur	Meer dan 1 uur	Minder dan 1 uur
☐	☐	☐	☐

4. Leg zo goed mogelijk in eigen woorden uit wat Wikipedia is en hoe het werkt.

……

……

……

……

……

……

5. Met welk doel zoek je doorgaans informatie op Wikipedia?

……

……

……

……

……

6. Heb je zelf al eens informatie toegevoegd of veranderd op Wikipedia?

ja / nee

7. Welke versie van Wikipedia heeft jouw voorkeur?

a. De Nederlandse
b. De Engelse
c. De Duitse
d. Anders, namelijk: ……

8. Heb je informatie van Wikipedia wel eens rechtstreeks gebruikt in een opdracht of een onderzoek?

ja / nee

Indien ja, geef een voorbeeld:……

……

……

……

……

9. In hoeverre vind je informatie van Wikipedia normaal gesproken betrouwbaar?

Zeer onbetrouwbaar			Neutraal			Zeer betrouwbaar
☐	☐	☐	☐	☐	☐	☐

10. Hoe moeilijk vind je het om een inschatting te maken van de betrouwbaarheid van artikelen op Wikipedia?

Zeer moeilijk			Neutraal			Zeer gemakkelijk
☐	☐	☐	☐	☐	☐	☐

11. Is het al eens voorgekomen dat je vertrouwen in informatie van Wikipedia onterecht bleek te zijn?

ja / nee

Indien ja, geef een voorbeeld:....……

……

……

……

……

Appendix C

PP# ……..

Art# ……..

Questionnaire na artikel

Op deze vragenlijst laat je jouw oordeel over de betrouwbaarheid van het voorgaande artikel weten. Wees opnieuw zo eerlijk mogelijk.

1. Hoe betrouwbaar kwam dit artikel op jouw over?

Zeer onbetrouwbaar			Neutraal			Zeer betrouwbaar
☐	☐	☐	☐	☐	☐	☐

2. Waarop is je oordeel gebaseerd?

Positief:…..……

……

……

Negatief:……

……

……

3. Hoeveel wist je van te voren al over dit onderwerp?

Zeer weinig			Neutraal			Zeer veel
☐	☐	☐	☐	☐	☐	☐

Appendix D

PP#
……..

Questionnaire achteraf

Hartelijk dank voor je deelname aan dit experiment. Als laatste willen we je nog wat vragen stellen over je deelname aan dit experiment.

1. In hoeverre kwam de taak van het beoordelen van de betrouwbaarheid die je tijdens dit experiment hebt uitgevoerd overeen met de manier waarop je normaal gesproken informatie op Wikipedia zou behandelen?

Zeer anders			Neutraal			Zeer gelijk
☐	☐	☐	☐	☐	☐	☐

2. Merkte je veel verschil in hoeveel je over het onderwerp van de verschillende artikelen wist?

Zeer weinig			Neutraal			Zeer veel
☐	☐	☐	☐	☐	☐	☐

3. Merkte je veel verschil in betrouwbaarheid tussen de artikelen?

Zeer weinig			Neutraal			Zeer veel
☐	☐	☐	☐	☐	☐	☐

4. Werd de taak bemoeilijkt door het feit dat er Engelse artikelen gebruikt werden?

Zeer weinig			Neutraal			Zeer veel
☐	☐	☐	☐	☐	☐	☐

5. Opmerkingen.

……

……

……

……