Transcription of the Bleek and Lloyd Collection using the Bossa Volunteer Thinking Framework

Ngoni Munyaradzi

Supervisor:
A/Prof. Hussein Suleman

Thesis presented for the Degree of Master of Science
in the Department of Computer Science
University of Cape Town

November 25, 2013

Plagiarism Declaration

1. I know that plagiarism is wrong. Plagiarism is using another's work and to pretend that it is ones own.
2. I have used the Association for Computing Machinery (ACM) as the convention for the citation and referencing. Each Significant contribution to, and quotation in, this thesis from the work, or works of other people has been citied and referenced.
3. This thesis is my own work.

SIGNATURE: ....

DATE: NOVEMBER 2013

Acknowledgements

First and foremost, it is only right to acknowledge the Lord for affording me this opportunity of studying for a Masters degree. This has certainly been one of my most challenging endeavours thus far. The past two and a half years have been extremely interesting, a lot of high and low moments, but without your strength, I would not have made it.

To my supervisor, Hussein Suleman, I say thank you. I could not begin to list all that you have done for me. Most importantly, your efforts and time over the last few years are appreciated and will never be forgotten. I am humbled that I got an opportunity to study under your supervision, I have learnt so much. The knowledge I have acquired will surely take me into the future a better student. I can proudly say I have learnt from one of the best minds.

To François Grey, thank you. It was an amazing opportunity collaborating with the Citizen CyberScience Centre for this project, not forgetting funding from the ShuttleWorth foundation. The visit to London opened my mind to the world of academia through meeting and interacting with other researchers with a passion for citizen science. Thank you for all your efforts to better my knowledge and open doorways for me, not forgetting the chance to present my research at the Muzeinberg Hackfest.

To all my colleagues from the Digital Libraries Laboratory, I say I love you all guys. Thank you to everyone who helped me with my research project. It has been such a wonderful learning experience interacting with young minds seeking knowledge and discussing thoughts on matters of life. I wish you all the best in life as you explore new avenues.

To the two most important ladies in my life, my mother and sister, thank you very much. You have supported me throughout my life, with all my plans of pursuing my academic career. I can personally say, I never knew that I would have reached this far when I first started Grade 1. I believe this is the beginning of greater opportunities in life. In everything, I did all this to make my mother proud.

Abstract

The digital Bleek and Lloyd Collection is a rare collection that contains artwork, notebooks and dictionaries of the earliest habitants of Southern Africa. Previous attempts have been made to recognize the complex text in the notebooks using machine learning techniques, but due to the complexity of the manuscripts the recognition accuracy was low. In this research, a crowdsourcing based method is proposed to transcribe the historical handwritten manuscripts, where volunteers transcribe the notebooks online. An online crowdsourcing transcription tool was developed and deployed. Experiments were conducted to determine the quality of transcriptions and accuracy of the volunteers compared with a gold standard. The results show that volunteers are able to produce reliable transcriptions of high quality. The inter-transcriber agreement is 80% for |Xam text and 95% for English text. When the |Xam text transcriptions produced by the volunteers are compared with the gold standard, the volunteers achieve an average accuracy of 69.69%. Findings show that there exists a positive linear correlation between the inter-transcriber agreement and the accuracy of transcriptions. The user survey revealed that volunteers found the transcription process enjoyable, though it was difficult. Results indicate that volunteer thinking can be used to crowdsource intellectually-intensive tasks in digital libraries like transcription of handwritten manuscripts. Volunteer thinking outperforms machine learning techniques at the task of transcribing notebooks from the Bleek and Lloyd Collection.

Glossary of Terms

- **BOINC** - Berkeley Open Infrastructure for Network Computing.
- **BOSSA** - Berkeley Open System for Skill Aggregation.
- **Cultural Heritage Collection** - a collection of historical artefacts that are either tangible or intangible like cultural beliefs.
- **Corpus** - a collection of written texts.
- **Digital Library System** - by [Oppenheim and Smithson, 1999]is an organised and managed collection of information in a variety of media (text, images, video or audio) all in digital form.
- **Digital Object** - any material encoded in a digital format.
- **Volunteer Thinking** - act of using one's brain and cognitive skills to solve a problem.
- **Volunteer** - Non-expert transcriber.
- **API** - Application Programming Interface.
- **GUI** - Graphical User Interface.
- **HIT** - Human Intelligence Task.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The speed and cost of scanning books has greatly improved over the years, to the extent of digitizing millions of books per year [Choudhury et al., 2006]. This has led many institutions to establish digital libraries for the preservation of ancient manuscript documents; most of the manuscripts are in image form. One such collection of interest to this research is the digital Bleek and Lloyd [Suleman, 2007] collection. It is a collection of scanned notebooks and artwork documenting the culture and beliefs of the hunter-gatherer people of Southern Africa, also known as Bushman, Basarwa, San or Khoisan people[1]. The notebooks specifically document the stories and languages of the |Xam and !Kun people. There is still need to transcribe these manuscripts into textual format, to enable indexing, searching, copying, editing[Alabau and Leiva, 2012] and possibly translation using text-to-speech tools. Manual transcription of the notebooks using experts is time-consuming and costly. Thus a crowdsourcing solution using volunteers was explored as an alternative.

1.1 Crowdsourcing

For the purposes of this project, crowdsourcing is defined as the process of outsourcing tasks to a distributed network of online volunteers to solve a problem using their intelligence and cognitive skills. Crowdsourcing has become a widely researched area in academia, and has gained significant popularity. Due to the success of crowdsourcing, numerous projects have emerged over the last few years; these projects or applications of crowdsourcing span multiple fields of study. The following list notes the various adopted forms [Eickhoff, 2011] of crowdsourcing across different disciplines:

- Wisdom of the Crowd - refers to aggregating the opinions of individuals in answering a question, rather than using a single expert on the subject.

[1] These terms are regarded as pejorative

- CrowdFunding - process where a collective group of people donate funding for an initiative.
- CrowdArt - is where individuals contribute a small portion of work in creating new art but none can do too much work.
- CrowdCreation - this is were crowds submit designs or ideas for an outsourced task.

An example of a successful CrowdArt initiative is Threadless[2]. This is an online community of artists and an e-commerce website where designers submit t-shirt designs for an incentive, which are later sold online. Kickstarter[3] is a profit-making company and said to be the worlds largest crowdfunding platform for creative projects. 99designs[4] is a marketplace for CrowdCreation. People come up with new logos, website and graphic designs including design contests. A classical example of the Wisdom of the Crowd is Wikipedia[5], a free online encyclopedia that is edited collaboratively by anyone using wiki software.

[Kittur, 2010] proposed that further research is needed to determine the full potential and limits of crowdsourcing. He suggested policy-making as a future research area. Since then, interesting applications of crowdsourcing have emerged. In Iceland, the crowd was used to approve a new constitution [Meyer, 2012a]. Finland, with the help of people, is creating new laws for the country [Meyer, 2012b] online. *Investigate Your MP's Expenses* was the first massively multi-player investigative journalism project. For the purposes of this research, the *Wisdom of the Crowd* is used on the problem of transcribing historical handwritten manuscripts.

Crowdsourcing has been applied to many domains of research and proven to produce good results that are comparable to those produced by experts. No work, to the best of our knowledge, has attempted to transcribe the |Xam or !Kun languages using this approach. This research shows that transcription of complex languages is possible using a distributed group of non-expert volunteers.

1.2 Problem Description

Within the last few years, there has been growing awareness for the need to digitize content among organizations like universities, museums and libraries. A number of reasons exist for this e.g, enhancing accessibility and ensuring

[2] www.threadless.com

[3] www.kickstarter.com

[4] http://99designs.com/

[5] www.wikipedia.org

preservation of the content. Through digitization of collections, cultural heritage institutions can improve dissemination of content to communities wider than the few individuals who have access to the artefacts. This promotes further inter-disciplinary studies and possibilities for knowledge creation. Most importantly, digitization facilitates the preservation of rare cultural heritage collections that contain human history, cultural beliefs and lifestyles.

With the continued increase of digital artefacts housed in Digital Libraries, various techniques are used to better expose these objects. Some of these activities include tagging, annotating and classification of images, and transcription and translation of texts. These activities are infamously task-intensive, repetitive and expensive to conduct. This project aims to investigate and evaluate the effectiveness of volunteer thinking for human-intensive digital library tasks, specifically transcribing the Digital Bleek and Lloyd collection[6]. The collection contains over 9800 images of handwritten pages of |Xam and !Kun languages of the earliest inhabitants of Southern Africa.

Transcription of the Bleek and Lloyd Collection poses a challenge, as some of the current well-known techniques like OCR are not adequate. What makes this task challenging is that the script used in the notebooks of the Bleek and Lloyd collection is not supported by Unicode, and not easily recognizable with OCR. The script has complex characters and diacritics that appear above and below characters; the diacritics can be stacked and span multiple characters. See Figure 3.9 on page 43 for an illustration of this. Once the images have been transcribed, metadata can be created that would allow for not only the images but the text in the notebooks to become searchable online.

1.3 Context

Preservation of cultural heritage artefacts is recognized as an urgent issue by UNESCO. In their Charter on the Preservation of Digital Heritage [Webb], under responsibilities, they urge hardware and software developers to work with heritage organizations to preserve digital heritage. This project aims to use software tools and crowdsourcing in creating new digital content for the preservation of the Digital Bleek and Lloyd Collection.

The research aims to determine how volunteers perform when presented with the task of transcribing complex handwritten manuscripts. This should give insight into which tasks volunteers are capable of performing in the digital libraries domain.

The digital Bleek and Lloyd project began in 2005 at the University of Cape

[6] http://lloydbleekcollection.cs.uct.ac.za/

Town. The first work done with this collection was by [Suleman, 2007], in which the goal was to make the digitized texts available online. Further research then focused on creating visual dictionaries for the languages [Williams et al., 2010], and using automatic machine learning techniques to recognize the text within the notebooks [Williams and Suleman, 2011b].

The Transcribe Bleek and Lloyd project is a collaboration between the University of Cape Town and Citizen Cyberscience Centre in Geneva. The aim of the collaboration is to promote the creation of citizen science projects in Africa where ordinary volunteers can contribute to scientific research.

Previous research used machine learning techniques [Williams and Suleman, 2011b] to transcribe the Bleek and Lloyd Collection but the accuracy of results was low. Hence a crowdsourcing solution is adopted for this project to determine how this compares with previous efforts.

1.4 Research Questions

This project aims to investigate:

- If volunteer thinking can be used to crowdsource intellectually-intensive tasks in digital libraries (like transcribing handwritten manuscripts).
- How volunteer thinking compares to machine learning techniques when applied to the problem of transcription.

1.5 Approach

This project proposed the development of a Bossa [Anderson, c] based digital library system that was integrated with the xoä'xoä transcription tool developed by [Williams and Suleman, 2011a] to transcribe the Bleek and Lloyd Collection. Bossa is an open-source framework for distributed/volunteer thinking, through the use of volunteers on the Internet to perform tasks that use human cognition, knowledge or intelligence.

The transcription tool was developed using an iterative process. The first of these iterative processes was the development of a prototype application on the Bossa framework. The purpose of the prototype was to obtain a conceptual understanding needed to develop Bossa applications. The next iterative phase included integrating the transcription tool into the Bossa framework. During the iterative implementation of the Bossa transcription tool, three experts were used for evaluating the system. Formal methods of recorded interviews were employed to note user feedback.

Once development was complete, the transcription tool was deployed online. The announcement of the project was done in phases - the initial target population was communities within South Africa who did research related to Bushman communities. This first phase was also meant for testing for any bugs, before announcement to the international community.

In the last phase, the project was announced internationally to interested volunteers. The project ran for 22 weeks. The results are detailed in Chapter 4. Volunteers were also asked to complete an online survey to better understand their experience when using the transcription tool.

1.6 Organisation of the Thesis

This thesis has five chapters and one appendix. Chapter two discusses the background and related work that serves as a foundation and motivation for the approach used in this research. Chapter three focuses on the technologies and tools used in the design and implementation of the transcription tool. Chapter four details the results and findings of the research. Chapter five concludes the work done in this research project and details future work. The user survey used for user experience evaluation is included in Appendix A.

Chapter 2

Background and Related Work

This chapter discusses volunteer thinking and its application in the area of natural language processing techniques, namely focusing on relevance judgements, question answering, translation and transcription of historical manuscripts. The discussion briefly touches on the factors that motivate volunteers to participate in volunteer thinking projects, and mentions popular tools available in setting up distributed thinking projects. The related work establishes the foundation of this research and serves as a guideline for the methodologies employed in implementing the transcription tool.

[Cohn, 2008] and [Silvertown, 2009] note that there is a realization amongst scientists that the public can provide free labour, computing resources and funding. Through open calls for public engagement in citizen science research, new and innovative projects have emerged [Trumbull et al., 2000]. Citizen cyberscience, a term coined by François Grey, is a spin-off from this realization. Three subcategories of citizen cyberscience exist, namely: volunteer computing, volunteer thinking and participatory sensing. This research focuses on volunteer thinking. The next section has more detail about volunteer computing and thinking.

2.1 Volunteer Computing and Thinking

Volunteer computing [Anderson and Fedak, 2006, Maurer, 2005] is a concept where the general public donate the idle time on their PCs to solve some scientific problem via the Internet. Volunteer computing is also referred to as peer-to-peer or global computing. The concept began in the mid 90's with the two projects, Great Internet Mersenne Prime Search (GIMPS)[1] and distributed.net[2], where thousands of computers were used to solve a single problem. To date these projects are still running. The GIMPS project has recently discovered the 48th largest known Mersenne prime. The initial software tools

[1] http://www.mersenne.org/

[2] http://www.distributed.net/

designed for distributed computing had problems, as they were too specifically tailored to individual projects. David Anderson developed the BOINC[3] middleware software, a generic software framework solution where multiple projects could take advantage of the distributed compute power from volunteers. To date over 50 distributed computing projects[4] utilize the BOINC software.

Volunteer/Distributed thinking [Quinn and Bederson, 2011] is the harnessing of human brain power on the Internet to solve problems that machines are not suitable to tackle. In volunteer thinking, users are tasked to solve some fundamental problem, reduced to a simplistic level that is easy to comprehend. Using their mental and cognitive abilities, volunteers actively attempt to solve the problem at hand. The types of problems vary in nature e.g. image tagging & classification, proof-reading documents and pattern recognition. The tasks are designed in such a manner that volunteers need no previous experience to solve the problem. Anderson also developed the Bossa [Anderson, c] crowdsourcing framework that manages distributed Web-based volunteer thinking projects.

Web 2.0 has made it possible to harness the computing power of non-experts in solving scientific problems, as current Web technologies support user generated content as opposed to passive viewing of content. What follows is a discussion of some popular crowdsourcing tools used in academia; the Bossa framework is the primary tool used for this research.

2.2 Crowdsourcing Tools

The crowdsourcing frameworks being used on the Web can be put into two categories: (1) Incentivised and (2) Non-Incentivised models. Task creators have varying reasons to choose either of these. For instance, if a project is running over a short period of time, then a task creator expects a fast turnaround for their project, so an ideal framework to use would be one that uses a payment model e.g the Amazon Mechanical Turk. On the other hand, if a project has unlimited running time or the task creator requires a framework that allows full flexibility to customize the project, one would opt for a non-monetary model like Bossa.

The Amazon Mechanical Turk [Ipeirotis, 2010] is a Web-based crowdsourcing platform that provides on-demand access to workers. A requester can set-up Human Intelligence Tasks (HIT) on Mturk - a HIT is a task a worker/turker can complete. Workers are paid for HITS completed. The range of payment

[3] http://boinc.berkeley.edu/

[4] http://boinc.berkeley.edu/projects.php

varies from \$0.10 to \$1.00. Mturk is a cheap way of carrying out tasks that are traditionally expensive when professionals are used. A requester can set-up a qualification test to filter out workers based on their skill level before they can participate in a project.

CrowdFlower [Finin et al., 2010] is a crowdsourcing platform that uses various channels to obtain their workforce, e.g their system interfaces with Mturk. CrowdFlower offers more control to the requester in managing and analysing their tasks. Another channel for workforce used by CrowdFlower is Gambit. The compensation model adopted by the Gambit marketplace is to pay workers using virtual currency, which rarely translates into actual money. The workers redeem the currency in online social games on Facebook like SportsBets or the Swag Bucks website[5]. Samasource[6] has similar functionality to Mturk and CrowdFlower.

Mturk uses a qualification test to filter out participants while CrowdFlower uses a gold standard. The gold standard is defined as a question to which the answer is already known. Amongst all these platforms, Mturk has emerged as the most popular crowdsourcing platform, mostly due to the fast turn-around time for tasks and the relatively low costs incurred in setting up a project.

Bossa[7] is an open source software framework for distributed thinking - where volunteers complete tasks online that require cognition skills, human knowledge and intelligence. Examples of such popular projects are GalaxyZoo[8] and Stardust@Home[9]. Bossa roughly works like the popular Amazon Mechanical Turk, but does not involve payment. The Bossa framework is implemented in PHP and can also be integrated with the Bolt[10] Web-based teaching tool for volunteers, which is a feature that can be used to assess volunteer skill.

PyBossa[11] is another open source software framework for distributed thinking; this was a spin-off from one of the Citizen CyberScience hackfests held in Cape Town, South Africa in November 2011. PyBossa is implemented in Python and its functionality is very similar to Bossa, but was designed to make the job creation tasks easier for task creators. To date, PyBossa is a live production system, and currently hosts a number of citizen science projects like Melanoma[12] and PDF transcription[13]. The following section highlights

[5] www.swagbucks.com

[6] www.samasource.org

[7] http://bossa.berkeley.edu/

[8] www.galaxyzoo.org

[9] http://stardustathome.ssl.berkeley.edu/

[10] http://boinc.berkeley.edu/trac/wiki/BoltIntro

[11] www.crowdcrafting.org

[12] http://crowdcrafting.org/app/melanoma/

[13] http://crowdcrafting.org/app/pdftranscribe/

fields in Natural Language Processing where crowdsourcing has been adopted.

2.3 Crowdsourcing Categories

In this section the discussion focuses on some applications of crowdsourcing in the fields of relevance judgements, translation, question answering, annotation and classification and lastly transcription.

2.3.1 Relevance Judgement

Relevance judgements are used in the field of Information Retrieval [Büttcher et al., 2010] where users have an information need. A query is usually performed on some retrieval system [Alonso et al., 2008] and based on the set of retrieved documents one can then judge if they are relevant or non-relevant with respect to the information need. The paper by [Mizzaro, 1997] gives a thorough discussion on the whole history of Relevance in an attempt to explain this fundamental topic that is not well understood.

[Lease and Yilmaz, 2012] point out that advances in stochastic evaluation algorithms have reduced the number of human judgement assessments required in the Cranfield tests [Cleverdon, 1997] for evaluating Information Retrieval systems. In spite of the advances, the assessment process is still slow and expensive. Fortunately, crowdsourcing offers an avenue for addressing these challenges, through the availability of distributed and on-demand workforce at a relatively low cost.

[Smucker and Jethani] research the judgement behaviour of crowdsourced workers versus university laboratory participants. They conclude that random crowsourced workers are not to be trusted, compared to university laboratory participants. [Alonso and Mizzaro, 2009] set out to access whether Mturk workers can replace Text REtrieval Conference (TREC) assessors in the task of relevance assessment. Their findings show that Mturk workers agree more with TREC assessors when the document is relevant, and less when the document is not relevant. In some instances, the turkers made better judgements than TREC assessors. Most importantly, they note that experiment design can adversely affect the outcome of results.

The SIGIR workshop report [Lease and Yilmaz, 2012] notes that studies in IR using crowdsourcing have been encouraging, but questions remain on how to effectively and efficiently employ crowdsourcing methods in practice. The work by [Alonso and Baeza-Yates, 2011] helps clarify some of these questions. Their research explores the design and execution of relevance judgements using Mturk and present a methodology for doing this. Their results show that the Mturk workers produce results comparable to TREC 8 experts. They also note that quality control should not only be implemented at worker level but different levels of the system e.g. user interface.

[Kazai et al., 2009] use a gaming model to encourage participants to contribute in the collective gathering of relevance assessments. The game makes provision for quality control and has incentives for users to follow a predefined review process. Their results show that incentives provide endurance for assessors, and the review process encourages truthful assessment.

2.3.2 Translation

Machine translation can be dated to the 17th century [Hutchins, 2004], where Renè Descartes proposed a universal language where one symbol can be used to represent different ideas. Machine translation (MT) is defined by Wikipedia[14] as the use of software to translate text or speech from one natural language to another. Harnessing the power of crowdsourcing for translation has many potential benefits [Kittur, 2010]. Some of the benefits that have been demonstrated are: translation of language in disaster relief [Hester et al., 2010]; and creation of training data and word alignment [Gao and Vogel, 2010] for machine translation.

Translations are mainly used in statistical machine translations, and these are obtained using various methods. Some of these include using comparable corpora [Hewavitharana and Vogel, 2011], using models that can be trained on monolingual data [Haghighi et al., 2008] and of late hiring human translators on Mturk. To fully realize the benefits of human computation, [Zaidan and Callison-Burch, 2011] point out that there has to be control measures used to obtain high-quality results. They recreate the NIST 2009 Urdu-to-English evaluation set on Mturk, and their models produce results within expert translator levels.

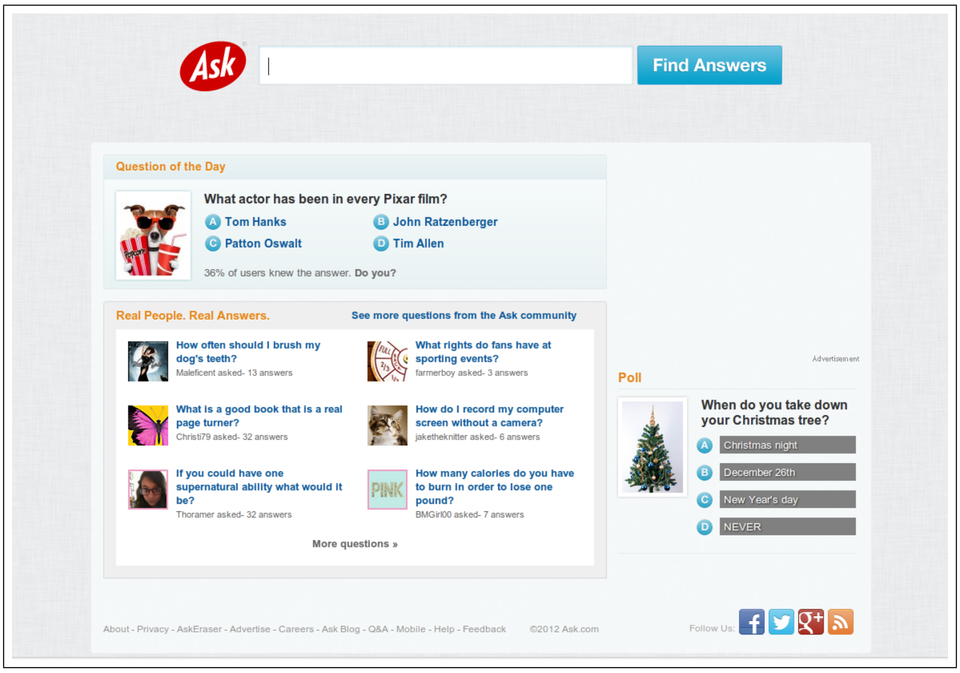
[Callison-Burch, 2009] states that the performance of machine translation depends of the size of the training data. Hence there is need for new and large training datasets, which are normally created by skilled experts and this is a time-consuming and costly process. The work by [Negri et al., 2011] attempts to address the issue of data scarcity for MT system training and evaluation. In a follow up work, [Negri and Mehdad, 2010] adopt a cost effective methodology of producing a bi-lingual Textual Entailment corpus. [Zaidan and Callison-Burch, 2011] estimated the cost of creating a small corpus with about 1.5 million words using an expert to be over $500 000. Negri and Mehdad showed that $100 was adequate to produce translations that are reliable. [Callison-Burch, 2009] findings show that non-experts produce results similar to experts in creating machine translation datasets. The challenge then that exists for

[14] http://en.wikipedia.org/wiki/Machine_translation

this project is to exploit the benefits of crowdsourced volunteer work and obtain results comparable to experts using volunteers.

2.3.3 Question Answering

The advent of Web2.0 has seen an increased growth in the use of Community-driven Question Answering (CQA) websites; Web2.0 has facilitated two way information exchange [Chua and Balkunje, 2012], allowing co-creation of content online. CQA websites have become competitors to traditional library reference services and machine learning techniques. Hitwise reports that the number of U.S visits to CQA sites between February 2006 and 2008 has increased nine fold [Chua and Balkunje, 2012]. CQA websites are viewed as knowledge hubs where users can post questions regarding any topic, and other members respond with an answer. The success of these websites relies on volunteer participation, assuming that everyone knows something [Adamic et al., 2008] and the combined wisdom of the crowds [Surowiecki, 2005]. Some of the popular CQA websites are: Askville[15], Yahoo! Answers[16] and Quora[17]. Figure 2.1 is a Web page for Ask.com.



Question of the Day

What actor has been in every Pixar film?

Ⓐ Tom Hanks
Ⓑ John Ratzenberger
Ⓒ Patton Oswalt
Ⓓ Tim Allen

36% of users knew the answer. **Do you?**

Real People. Real Answers.

See more questions from the Ask community

How often should I brush my dog's teeth?
Maleficent asked- 13 answers

What rights do fans have at sporting events?
farmerboy asked- 3 answers

What is a good book that is a real page turner?
Christi79 asked- 32 answers

How do I record my computer screen without a camera?
jaketheknitter asked- 6 answers

If you could have one supernatural ability what would it be?
Thoramer asked- 32 answers

How many calories do you have to burn in order to lose one pound?
BMGirl00 asked- 7 answers

More questions »

Advertisement

Poll

When do you take down your Christmas tree?

Ⓐ Christmas night
Ⓑ December 26th
Ⓒ New Year's day
Ⓓ NEVER

About - Privacy - AskEraser - Advertise - Careers - Ask Blog - Q&A - Mobile - Help - Feedback ©2012 Ask.com

Follow Us:

Figure 2.1: Ask.com - Question Answering Website

[Fichman, 2011] points out that issues are raised with regard to information quality and benefits of user generated content on Web2.0 platforms are con-

[15] http://askville.amazon.com/Index.do
[16] http://answers.yahoo.com/
[17] http://quora.com/

tested [Bandura, 1977]. He evaluates four Q&A websites (namely: Askville, WikiAnswers, Wikipedia Reference Desk and Yahoo! Answers) based on their answer qualities. The findings show that (1) the popularity of a Q&A site does not correlate to answer quality; (2) many answers to a question does not guarantee higher accuracy of answers, but improves verification and completeness and; (3) similar collaborative efforts lead to significantly different answer accuracy, completeness and verifiability.

The research by [Shachaf, 2009] investigates the quality of answers on the Wikipedia Reference Desk, and compares it with library reference services. Their aim is to determine if volunteers can outperform expert reference librarians. Their results show that the quality of the Wikipedia Reference Desk is similar to that of library reference services. Both systems provide reference services at the 55% accuracy level [Saxton and Richardson, 2002]. Wikipedia Reference Desk has better question responsiveness time and answer completeness than traditional library reference services. Overall, the volunteers outperform the expert librarians; this is significant because the volunteers are amateurs and not paid for the services. An important point to note is that the individual responses submitted by volunteers were comparable to those of librarians. Only the amalgamated responses from volunteers produced answers that were similar or better than those of expert librarians. Some of the findings by Shachaf nullify those of [Fichman, 2011] mentioned above.

The work by [Chua and Balkunje, 2012] does a comparative evaluation of six CQA websites based on a conceptual framework that considers information management, information quality and system usability. Information management is defined as information sharing and organization, while information quality is composed of content value, cognitive value and socio-economic value of answers. Their findings show that the usability features of CQA websites can be improved, and that people put more importance on the socio-economic value of answers than information quality. Lastly, the organization of information of CQA websites is more important to users than sharing content. Its also noted that uses express gratitude for responses to their questions, a finding consistent with that of [Kim and Oh, 2009].

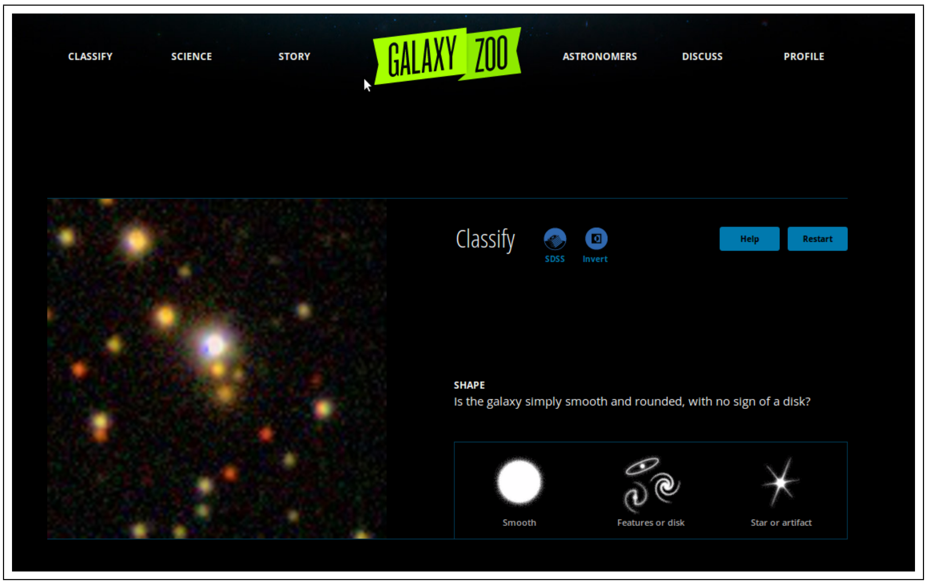
[Franklin et al., 2011] note that machines are not fully able to answer some questions. They developed a system (CrowdDB) that relies on human input to answer questions that machines cannot fully answer. [Bulut et al., 2011] investigate the feasibility of location-based query answering via Twitter and Foursquare; currently search engines perform poorly on this problem. Their findings show that latency in answering questions is low, with about 50% of questions being answered in 20 minutes. Library reference services on average take about 48 hours [Shachaf, 2009] to obtain a response.

2.3.4 Annotation and Classification

With the growing number of digital libraries and content stored within them, especially scanned images of historical importance, annotation plays an important role. Labelling of digital artifacts supports indexing, searching and browsing of the content online. [Nowak and Rüger, 2010] investigate whether annotations generated by non-experts via crowdsourcing are reliable to use as ground-truth data. They attain an accuracy of 92% on a small database when the ground-truth generated by non-experts is compared with the merged ground-truth by experts, similar to findings by [Snow et al., 2008] and [Sorokin and Forsyth, 2008]. Snow et al further show that large annotation tasks can be carried out at a fraction of the cost. Researchers need alternative viable methods of collecting data fast and at a low cost [Alonso and Mizzaro, 2009], while still attaining accurate results. [Rashtchian et al., 2010] show that the use of qualification tasks in creating image corpora on the Mechanical Turk produces the best improvement in the quality of results.

Stardust@Home [Westphal et al., 2006] is one of the first Web-based volunteer thinking projects started in 2006; it is a project in search of contemporary interstellar dust collected from space. The task of the "dusters" is to identify the interstellar particles before they can be analysed further. Within a period of eleven months, 20064 people had performed more than 30 million searches. GalaxyZoo [Lintott et al., 2008] is an astronomy project that was inspired by Stardust@Home. In this project volunteers are tasked with classifying the morphology of approximately 1 million galaxies obtained from the Sloan Digital Sky Survey. It is estimated that a graduate student would take between 3 and 5 years to complete the task, whereas more than 20,000 volunteers took approximately a month[18]. The project recruited over 100,000 public volunteers. Figure 2.2 shows the Web interface for the GalaxyZoo project.

[18] http://en.wikipedia.org/wiki/Galaxy_Zoo



SHAPE

Is the galaxy simply smooth and rounded, with no sign of a disk?

Figure 2.2: GalaxyZoo Project

Clickworkers [Kanefsky et al., 2001] is an experimental project set-up by NASA, where volunteers identify and classify the age of craters on Mars images. The volunteers rely on their perception and common sense to solve the task. It is estimated that it would take a graduate student or scientist months to complete the task. One of the objectives of this project was to answer two questions:

1. Determine if volunteers are ready and willing to contribute to science?
2. Determine if this new way of conducting science produces results as good as earlier established methods?

Research thus far [Callison-Burch, 2009, Lintott et al., 2008, Nowak and Rüger, 2010, Snow et al., 2008] has shown that both questions can be answered in the affirmative.

2.3.5 Transcription

Transcription is a widely studied research area, with most research focusing on the transcription of speech or text. The primary focus of this project is on the manual transcription of unconstrained handwritten historical manuscripts. Three approaches to the transcription of historical manuscripts are discussed in this section: (1) fully automated, (2) semi-automatic and (3) manual transcription.

The automatic transcription of handwritten manuscripts is a challenge because most handwriting recognizers perform poorly on these noisy texts [Rath, 2003]. Other than the noisy data, what makes this task more challenging is when the collection of documents being analysed are written by multiple authors. A study by [Williams, 2010] on the transcription of the |Xam language showed that this is feasible, with accuracies of 78% and 63% for two different authors.

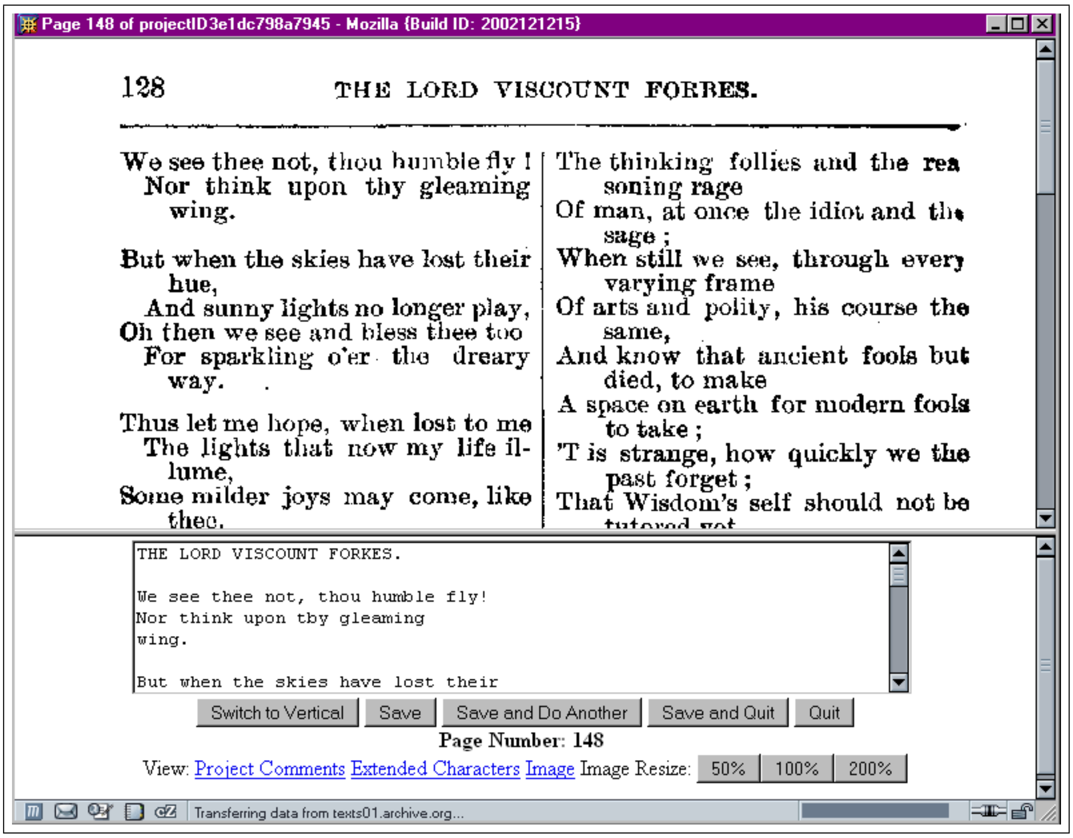
Unlike OCR, some semi-automatic transcription systems use Handwritten Text Recognition (HTR) modelled closely to Automatic Speech Recognition [Alabau and Leiva, 2012]. Three methods for HTR are currently being used: (1) post-editing [Plamondon and Srihari, 2000] (2) interactive-predictive [Toselli et al., 2010] and (3) active learning [Serrano et al., 2010]; all these require human-input. [Toselli et al., 2007] developed a hybrid system that takes advantage of the accuracy of human transcribers and speed of automatic handwriting recognition systems to complete highly accurate transcriptions. [Guichard et al., 2011] further propose a technique to reduce automatic recognition errors and the tedious human input required, by taking advantage of word redundancies over pages and considering documents from a collection level. They reduced the human effort required by 28% and achieved an annotation rate of 80%, showing an improved performance in their work. [Dahab and Belz, 2010] present a prototype game-based methodology of transcribing typed or handwritten text on images. [Alabau and Leiva, 2012] aim to change the tedious process of transcribing handwritten text into a fun enjoyable experience with a word soup game online.

Distributed Proofreaders[19] (DP) [Newby and Franks, 2003] is a Web-based project where volunteers proof-read OCR'ed text and compare with source images. Project Gutenberg aims to create and disseminate electronic books online from hard-copy versions, which are not under copyright. In 2002, over 250,000 unique pages were proofread. Their page rate increased to 110,000 in December 2002 due to coverage by Slashdot[20] in November 2002. Figure 2.3 is an image of the DP user interface. Old Weather[21] is a citizen science project that aims to collect data about temperatures from historical ship records. The records were captured by sailors on British Royal Navy ships between 1905 and 1929. Within three months, 202,904 pages had been transcribed. The task is fairly simple as volunteers only have to capture the following details: date, location, event and weather from the ship logs.

[19] www.pgdp.net/

[20] http://slashdot.org

[21] www.oldweather.org/



Page 148 of projectID3e1dc798a7945 - Mozilla {Build ID: 2002121215}

128 THE LORD VISCOUNT FORBES.

We see thee not, thou humble fly !
Nor think upon thy gleaming wing.

But when the skies have lost their hue,
And sunny lights no longer play,
Oh then we see and bless thee too
For sparkling o'er the dreary way.

Thus let me hope, when lost to me
The lights that now my life illume,
Some milder joys may come, like thee.

The thinking follies and the rea soning rage
Of man, at once the idiot and the sage ;
When still we see, through every varying frame
Of arts and polity, his course the same,
And know that ancient fools but died, to make
A space on earth for modern fools to take ;
'T is strange, how quickly we the past forget ;
That Wisdom's self should not be tutored yet

```
THE LORD VISCOUNT FORKES.

We see thee not, thou humble fly!
Nor think upon tby gleaming
wing.

But when the skies have lost their
```

Switch to Vertical
Save
Save and Do Another
Save and Quit
Quit
Page Number: 148
View: Project Comments Extended Characters Image Image Resize: 50% 100% 200%
Transferring data from texts01.archive.org...

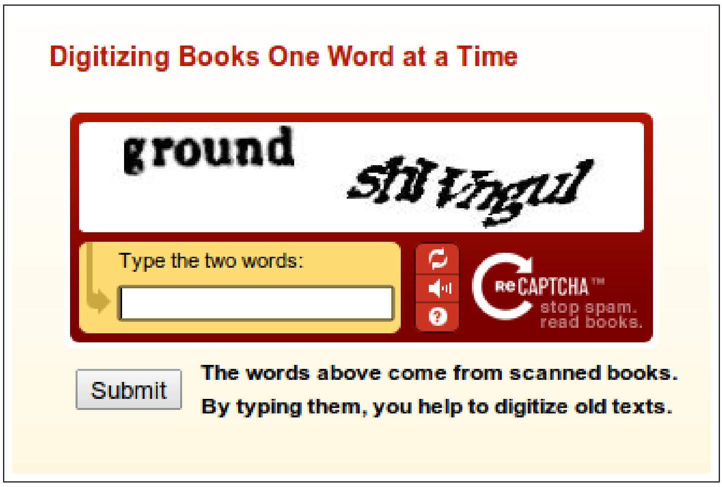
Figure 2.3: Distributed Proof Readers Web Interface

[Brumfield], a notable blogger on collaborative manuscript transcription, described 2010 as *The Year of Crowdsourcing Transcription*, as a number of new projects emerged. Some of these are: The North American Bird Phenology Program[22], Demogen[23], Family Search Indexing[24], World Archives Project [25] and Transcribe Bentham[26].

reCAPTCHA[27] is a tool used for security against online spam programs, (see Figure 2.4). reCAPTCHA is used to digitize books, newspapers and old radio shows. Humans are required to recognise some distorted text, which machines cannot, before they can gain assess to some Web service. This service is deployed in more than 44 000 websites and has been used to transcribe over 440 million books, achieving word accuracies of up to 99% [Ahn et al., 2008]. The work by [Causer and Wallace, 2012] in the Transcribe Bentham project gives an enlightening picture of the effort required to successfully create awareness about a transcription project and costs involved. Their research also discusses

[22] http://www.pwrc.usgs.gov/bpp/
[23] http://demogen.arch.be/
[24] https://familysearch.org/volunteer/indexing
[25] http://community.ancestry.com/awap
[26] http://blogs.ucl.ac.uk/transcribe-bentham/
[27] http://www.google.com/recaptcha

the successful methods employed. The approaches and works discussed are aimed at improving dissemination, accessibility and preservation of cultural heritage collections. The next section looks at a few cultural heritage collections were efforts have been made to digitize the collections.



Digitizing Books One Word at a Time

The words above come from scanned books. By typing them, you help to digitize old texts.

Figure 2.4: reCAPTCHA Web Interface

2.4 Cultural Heritage Collections

In line with the theme of this research, this section shall discuss some popular heritage collections and their use in academic research.

2.4.1 The Timbuktu Manuscripts

The Timbuktu manuscripts originated from the country of Mali [Brenner and Robinson, 1980] and are believed to have been written around 1300, making them some of the oldest manuscripts in Africa. The manuscripts were written in various languages, including Arabic and Syriac. Old British Commonwealth courses taught that Africa had no written languages before the arrival of European colonial powers; it was perceived that Africans were incapable of intellectual work. But all this was disproved by the discovery of the Timbuktu manuscripts. [Heng, 2007] notes that more than a million manuscripts exist, and of these only a fraction have been catalogued or digitized and stored in libraries, while the rest deteriorate in people's homes. [Hale, 2012] points out that some of the manuscripts have been looted by part of the Tuareg nationalists.

In 2004, a pilot academic project called Tombouctou [Farouk-Alli and Mathee] began at the University of Cape Town. This project's goal was to promote academic research with the manuscripts, and not only focus on conservation. The team digitized 100 manuscripts from the Mamma Haidara Library and

60 manuscripts from the Ahmed Baba Institute. Further studies done on the digitized manuscripts included translation to English and efforts to proficiently read the texts. The Ahmed Baba Institute has been microfilming and cataloguing the manuscripts they possess, estimated to be around thirty thousand [Hale, 2012].

[M'kadem and Nieuwenhuysen, 2010] conducted a survey to find out whether researchers in Moroccan universities are prepared to change from direct access to on/offline access of the manuscripts. Their findings reveal that there exists a resistance amongst researchers to adopt this technology as they prefer to interact with the owners of the private collections. This is attributed to availability of richer commentary possessed by the holders of the manuscripts. [Doumat et al., 2008] developed a prototype online archive application for management and collaborative annotation of ancient handwritten manuscripts. They use a number of collections for testing, including manuscripts from Timbuktu. Aluka[28] is a unique collection of manuscripts, reference works and many other artefacts from and about Africa. In 2006, Aluka collaborated with partners in Timbuktu and set up a lab to catalogue 600 manuscripts, and digitize 300 of these [Ryan, 2010], which was competed in 2007.

2.4.2 Oxyrhynchus Collection

The Oxyrhynchus Collection [Oxy, 1898] was discovered by British scientists near the city of Oxyrhynchus in Egypt at the end of the 19th century. This is considered to be one of the world's most valued treasure dumps. The collection of about a million papyrus records, roughly 2 000 years old, was well preserved. This was later moved to Oxford University for study by scholars who would transcribe and translate the texts. The collection included accounts of Greek daily life, the controversial Gospel of Thomas and other Greek practices.

After 100 years, scholars had only managed to analyse 15% of the collection. To quicken the pace, a website (Ancient Lives[29]) was set up for volunteers to help transcribe the ancient Greek texts. Within a short time, volunteers completed four million transcriptions. Amongst the text transcribed there are works by Thucydides and Aristophanes. Ancient Lives is one of many citizen science projects that are being run to help engage the community in new discoveries within academic research fields. Efforts like these show the power of crowdsourcing in advancing scientific research.

[28] www.aluka.org

[29] http://ancientlives.org/transcribe

2.4.3 George Washington Letters

The Library of Congress has approximately 65,000 documents of original writings by George Washington, former president of the United States of America [Andreassen, 1949]. These writings are a documentation of Washington's interests, activities and correspondence between the period of 1741 and 1799, also covering his two presidential terms, command of the American army during the revolutionary war and youth. The writings of Washington also provide knowledge about how the United States of America was established [Manmatha and Rothfeder, 2005].

The works by Washington have been used in several optical character recognition projects. [Lavrenko et al., 2004] propose a holistic word recognition approach for single author manuscripts and achieve 65% recognition accuracy. [Manmatha and Rothfeder, 2005] propose a novel algorithm for the segmentation of handwritten manuscripts into words. They achieved 17% recognition accuracy, which is far better than state of the art metrics for word segmentation. [Rath et al., 2004] developed the first known search engine system for the retrieval of handwritten images in large collections, based on statistical models.

[Kane et al., 2001] initially evaluated the possibility of indexing the George Washington handwritten manuscripts and noted that one main challenge was matching the word images and classifying them into classes to build a searchable index. The works discussed next aim to address this issue using various techniques. [Rath and Manmatha, 2007] proposed using word spotting for indexing historical documents; they obtained 2867 image labels with an error rate of 38.12%. [Feng et al., 2008] used Hidden-Markov models for assembling characters in alphabet soups. They used 20 pages and noticed a 20% error difference with two estimates, suggesting that more reliable estimates for recognition accuracy could be explored in future. [Adamek et al., 2007] used single closed contours for word matching and achieved 83% recognition accuracy on a set of 20 pages for the same task.

2.4.4 The Beowulf Collection

The 'Beowulf manuscript' or 'the Nowell codex', is a collection of individual medieval manuscripts [Heaney, 1999]. The only copy of this manuscript is found in the London British Library. The structure of the manuscript is difficult to decipher as it was damaged in a fire in 1731. The following texts are believed to make up the volume, and are written in Anglo-Saxon:

- Beowulf
- Judith

- The Marvels of the East
- Letter of Alexander to Aristotle
- A fragment of a Life of St Christopher

Amongst these texts, Beowulf is the most discussed [Lucas, 1990] manuscript and longest surviving Old English poem, with uncertainty regarding its compilation and make-up. The Beowulf collection is believed to be the work of two scribes, produced around the eleventh century.

[Brown and Seales, 2001, 2004] and [Graham, 1998] discuss techniques that can be used to preserve and restore deteriorated manuscripts. Some of these methods have been applied to the Beowulf collection, where digitizing using special lighting has made certain parts more readable [Kiernan, 1991] with implications on established knowledge about the manuscripts. In 1997, a CD-ROM containing transcriptions, essays and a glossary had been made of the collection [Porter, 2002]. In his paper, [Prescott, 1998] describes the process, technologies and strategies undertaken in constructing the Electronic Bewoulf project with partnership of Kieran, Szarmach and many other players. Kieran helped set up the Electronic Beowulf project[30] at the University of Kentucky, which is a digital image archive of the Beowulf [Prescott, 1997]. This collection is said to be radical, as it is described by images and not words, while possessing scholarly research material. Part of the Bewoulf Collection is used as educational material in schools [Conner, 1991, Klass, 2002].

2.4.5 Jeremy Bentham

Jeremy Bentham (1748-1832) was a philosopher and jurist [Bentham, 2000]. At the age of three he began studying Latin, a sign of his brilliance. He studied law at Queen's College, Oxford. During that time he began writing and would write about ten to twenty manuscripts daily, even into his old age. Bentham was an advocate for freedom of speech, gender equality and a critic of existing laws during his time.

The Bentham project was established at the University College of London in 1958, with intentions of publishing the works of Jeremy Bentham. In 2010, the Transcribe Bentham[31] project was launched with the goal of transcribing part of the collected works of Bentham and making this material accessible to the general public. Another aim of the project was to answer five questions [Causer et al., 2012] but only the ones relevant to this research are stated below:

[30] http://ebeowulf.uky.edu/

[31] http://blogs.ucl.ac.uk/transcribe-bentham/

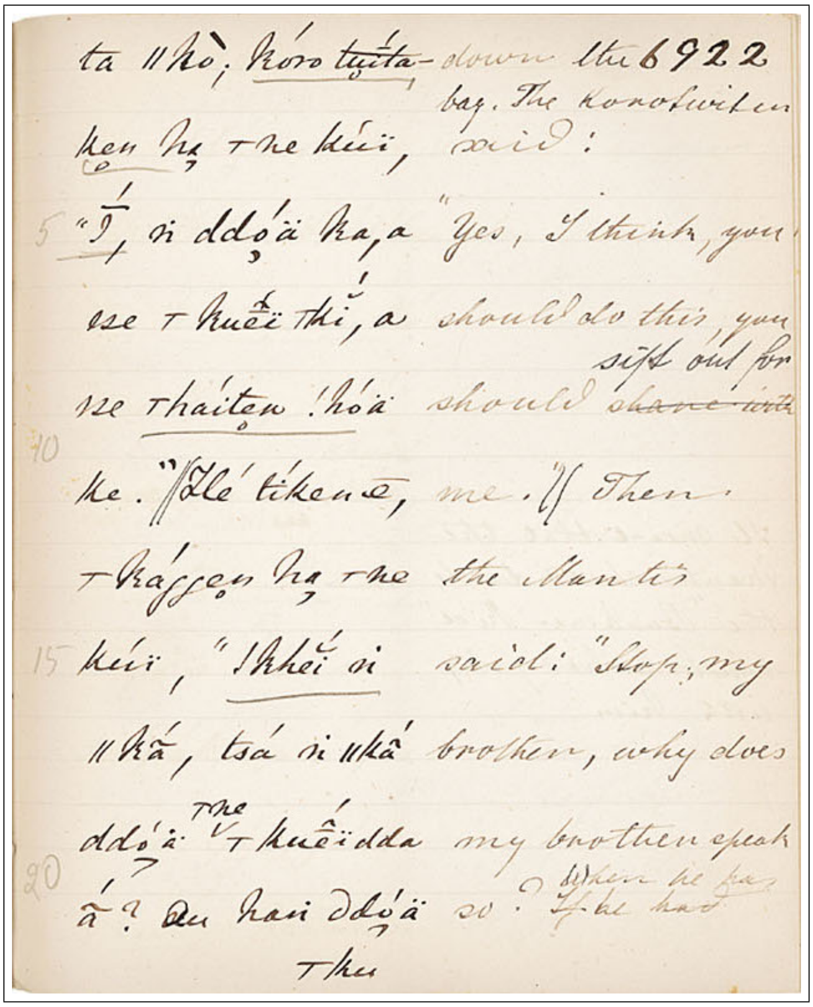
- How would the success of a project like Transcribe Bentham be measured?
- Are volunteer transcriptions of good quality for academic purposes?

A total of 1009 manuscripts were transcribed and, of these, 56% were regarded as complete. Out of the 1207 volunteers who registered on the transcription desk, only 21% were active within the first period of the project. The majority of the transcriptions were produced by a minority. Overall, the project was successful, based on public engagement, number of transcriptions obtained and sustainability. Most importantly, they showed that volunteers are able to produce transcriptions that can be used for scholarly purposes.

2.4.6 Bleek and Lloyd Collection

The Digital Bleek and Lloyd Collection [Suleman, 2007] is composed of dictionaries, artwork and notebooks documenting stories about the earliest inhabitants of Southern Africa. The notebooks were written by Wilhelm Bleek, his sister-in-law, Lucy Lloyd and Dorothea Bleek in the 19th century, with the help of a number of |Xam and !Kun speakers who were prisoners in the Western Cape region of South Africa at the time. Figure 2.5 is a sample page from the notebooks of the collection.

The notebooks were recorded in the |Xam and !Kun languages; English translations of these languages are available in the notebooks. The |Xam and !Kun languages are not represented in standard Unicode; the text contains complex diacritics that appear above a character, below it or both. This is a rare collection of original cultural heritage of the earliest inhabitants of Southern Africa who possessed a unique view of the world. The notebooks have been digitized and made accessible online by the University of Cape Town. [Suleman] discusses the importance and need for preservation of cultural heritage collections, and uses the Bleek and Lloyd Collection as an example. [Williams, 2010] explores the feasibility of automatically transcribing the notebooks from the Bleek and Lloyd Collection; he then develops a corpus of the texts for handwriting recognition [Williams and Suleman, 2011a]; and lastly uses Hidden Markov Models to transcribe the text [Williams and Suleman, 2011b]. These are all efforts made by researchers at the University of Cape Town to improve the accessibility of the collection. The following section gives a summary of the chapter.



ta ‖kō; !koro tuita- down the 6922
bay. The Korotuiten
ken ha ⊤ne kui, said:
"Ī, ŋ ddóä ka, a "Yes, I think, you
ūse ⊤kuēĩ thĩ, a should do this, you
sift out for
ūse ⊤haiten !hóä should ~~share with~~
ke." "‖kē tikenṵ, me." Then
⊤kaggen ha ⊤ne the Mantis
kui, "Ikhei ŋ said: "Stop, my
‖kā, tsá ŋ ‖kā brother, why does
⊤ne
ddóä ⊤kuēĩdda my brother speak
~~When he has~~
ā? Au hai ddóä so? If he had
⊤hu

Figure 2.5: A sample page from the Bleek and Lloyd Notebooks

2.5 Summary

This chapter has shed light on the idea of volunteer/distributed thinking (human computation) and its roots in volunteer computing. Next the chapter mentioned some of the popular crowdsourcing tools used within academia. The discussion also highlighted research areas where volunteer/distributed thinking is being applied. The literature published thus far has shown that volunteers are capable of producing results comparable to experts and in some instances superior results. Crowdsourcing has been shown to be a viable solution to use in research, and manual transcription appears to be a feasible solution for scholarly purposes. In light of the research theme of preservation of cultural heritage collections, the chapter discussed preservation efforts of some ancient heritage collections, similar to the Bleek and Lloyd collection used for this research. The following chapter describes the design and implementation of the Bossa-based transcription tool.

Chapter 3

Design and Implementation

This chapter mainly focuses on the design and implementation of the experimental Bossa-based transcription tool. Firstly, the chapter describes the Bossa, Bolt and Boinc software tools essential to the core functionality of the transcription tool. Next the chapter outlines the implementation specifics of the transcription tool, followed by a brief discussion of additional software tools used. The features and functionality of the transcription tool are then explained. Lastly, the chapter concludes by looking at the deployment of the transcription tool.

3.1 BOSSA

The Berkeley Open System for Skill Aggregation (Bossa) [Anderson, c] is an open source software framework for distributed thinking - where volunteers complete tasks online that require cognition, human knowledge and intelligence. Bossa was developed by David Anderson[1] from Berkeley University of California. On their project website[2] they provide thorough documentation to set up Bossa Web applications. A video tutorial and slides[3] from one of Anderson's presentations are available online, also discussing Bossa application design. This material was referenced for the design and implementation of the transcription tool.

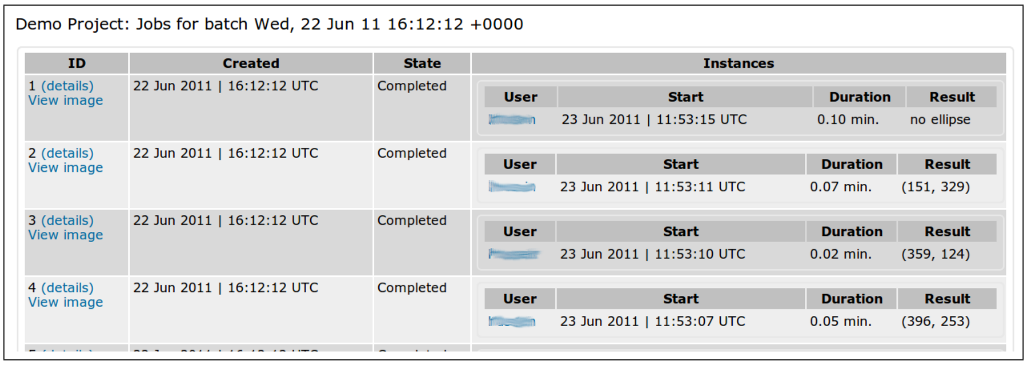
The Bossa framework is similar to the Amazon Mechanical Turk but gives the project administrator more control over the application design and implementation. Unlike the Mechanical Turk, Bossa is entirely volunteer work with no monetary incentives. The framework simplifies the task of creating distributed thinking projects; Stardust@home is an example of a popular crowdsourcing project that is run on the Bossa framework.

[1] http://boinc.berkeley.edu/anderson/

[2] http://boinc.berkeley.edu/trac/wiki/BossaIntro

[3] http://boinc.berkeley.edu/slides/bossa_intro.pdf

Bossa was designed in a manner that would allow a project administrator to easily set up a Bossa application. The framework provides a MySQL database with pre-populated tables of the important application details that need to be captured; one has the option to add more tables if they choose to do so. To set up an application, the administrator has to define a few PHP callback functions. These callback functions determine how the tasks are to be displayed, manage issuing of further tasks and what happens when a task is completed or has timed out. Additional job creation scripts have to be defined (see Appendix A.1). Each application that is created has a batch of jobs that are associated with it, and these can be viewed using an operator's administrative interface (see Figure 3.1).



Demo Project: Jobs for batch Wed, 22 Jun 11 16:12:12 +0000

ID	Created	State	Instances			
			User	Start	Duration	Result
1 (details) View image	22 Jun 2011 | 16:12:12 UTC	Completed	[redacted]	23 Jun 2011 | 11:53:15 UTC	0.10 min.	no ellipse
2 (details) View image	22 Jun 2011 | 16:12:12 UTC	Completed	[redacted]	23 Jun 2011 | 11:53:11 UTC	0.07 min.	(151, 329)
3 (details) View image	22 Jun 2011 | 16:12:12 UTC	Completed	[redacted]	23 Jun 2011 | 11:53:10 UTC	0.02 min.	(359, 124)
4 (details) View image	22 Jun 2011 | 16:12:12 UTC	Completed	[redacted]	23 Jun 2011 | 11:53:07 UTC	0.05 min.	(396, 253)

Figure 3.1: Administrators Interface to manage Jobs

Bossa provides two important Web pages, *bossa_get_job.php* and *bossa_job_finished.php*. The former displays a new job to a volunteer. This is displayed on the transcription interface (see Figure 3.12). For this project, a new job is defined as an image with |Xam and/or English text. The latter function is invoked when a volunteer has completed their assigned job, and that job's state is altered in the MySQL tables. Based on the project's policies, various actions can be executed; this is later discussed in section 3.4.3. Figure 3.2 shows the software structure[4] of Bossa. The application has a set of jobs associated with it; the job distribution policies determine the number of instances of a job that are sent out to users. Bossa projects can have diverse requirements, and implementation specifics depend on the project developer. Bossa provides mechanisms for dealing with such varying project requirements. A developer needs to have basic knowledge of PHP and has to provide implementations of the following callback functions:

- job_show() - displays the next job in the queue.
- job_issued() - changes the state of a job once it is issued.
- job_finished() - changes the state of a job once completed.

[4] http://boinc.berkeley.edu/trac/wiki/BossaOverview

- job_timeout() - changes the state of a job when the time limit is reached.

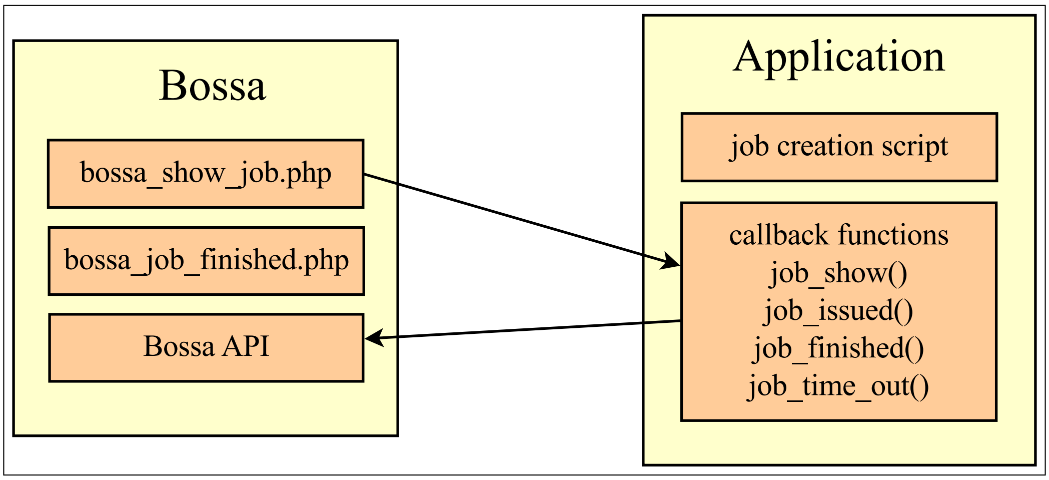


Figure 3.2: Software structure of Bossa

The naming of these functions is intuitive to function purpose. The Bossa framework allows for developers to run multiple projects at the same time, each governed by its own set of policies. An interesting use-case of this would be when the project initially runs a Bolt training course to determine volunteer skill, and based on the results volunteers are allocated to different Bossa projects.

Bossa supports calibration and non-calibration jobs. A calibration job is when a solution to a problem is known. Integration of this with the Bolt training tool would effectively help in filtering volunteers. Another interesting feature of Bossa is adaptive replication, where volunteers are assigned jobs in increasing order of complexity; the complexity level would depend on a volunteer's performance on previous jobs. In summary, to set-up a Bossa project, one needs to define policies for job distribution, representation, display and volunteer assessment (optional). Further details are given in the section 3.4 on how these policies and mechanisms were implemented for this project. The following section looks at the Bolt framework, a tool used for volunteer training in this research.

3.2 BOLT

Volunteer training is a common practice in crowdsourcing projects, to ensure that volunteers have sufficient knowledge to produce quality results [Le et al., 2010] for the task. Bolt [Anderson, b] is a software framework developed for volunteer training and assessment. It is a Web-based training and education software toolkit, which integrates with Bossa. A Bolt course is composed of a sequences of lessons, exercises and course document; illustrated in Figure 3.3. The toolkit can be used to display various HTML content, e.g. flash, videos. The tool is used to train volunteers on how to perform transcriptions tasks, using the custom transcription tool (see Figure 3.12).

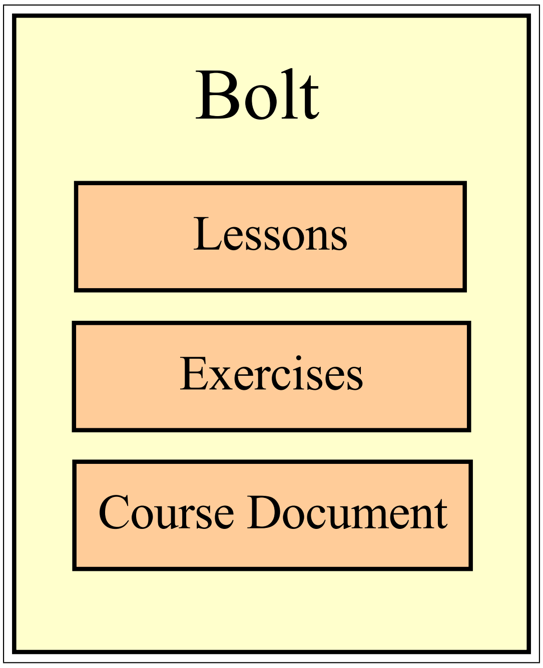
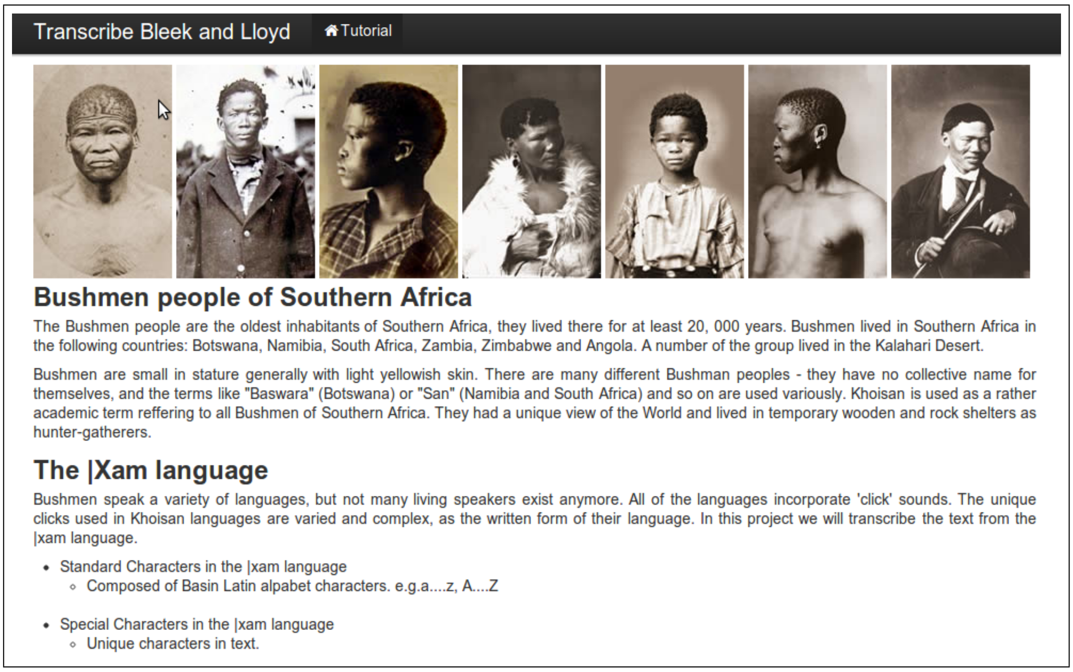


Figure 3.3: Bolt Course Structure

The Bolt framework allows the project administrator to implement mechanisms of determining individual volunteer skill and moderating access to users. The volunteer skill assessment can come from several different sources of information, for example, from pattern recognition with an application like Recaptcha [von Ahn et al., 2008] or a test written after completing a tutorial on a classification problem. Volunteer assessment can be implemented in one of two ways or a combination of both - designing a training course or creating an exercise with known solutions. Crowdsourcing challenges can either have a known or unknown solution to the problem. In the case that a solution is not known for the challenge, the best form of assessment would be to design a training course for the given task. The challenge posed in this research does not have a known solution to the problem. An ideal assessment for this project would be tasking volunteers with an exercise to recognize the characters and diacritics of the |Xam language.

Once a user has created an account for the project, they are expected to take a training course before they can start transcribing. The training course is composed of two parts. The first one is a short descriptive history of the earliest inhabitants of Southern Africa, the hunter-gatherer people. The first part is also aimed at motivating volunteers to contribute to the project. The second part is a short transcription tutorial video on how to complete transcriptions of |Xam and !Kun languages. Figures 3.4 and 3.5 show the descriptive history of the Bushman people and the video tutorial respectively. Next is a description of the Boinc software.



Bushmen people of Southern Africa

The Bushmen people are the oldest inhabitants of Southern Africa, they lived there for at least 20, 000 years. Bushmen lived in Southern Africa in the following countries: Botswana, Namibia, South Africa, Zambia, Zimbabwe and Angola. A number of the group lived in the Kalahari Desert.

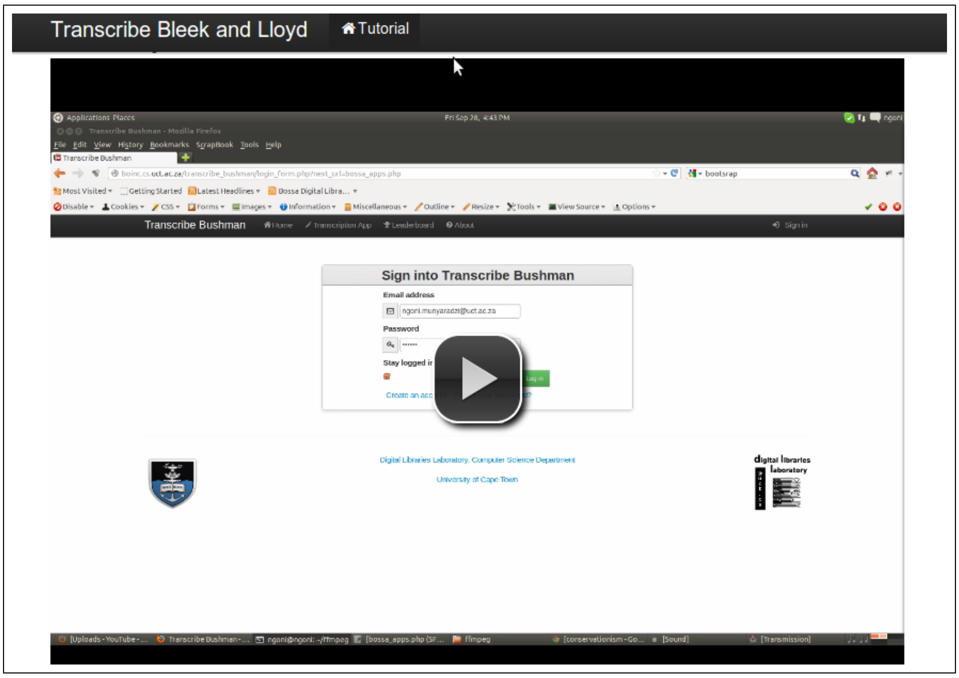
Bushmen are small in stature generally with light yellowish skin. There are many different Bushman peoples - they have no collective name for themselves, and the terms like "Baswara" (Botswana) or "San" (Namibia and South Africa) and so on are used variously. Khoisan is used as a rather academic term reffering to all Bushmen of Southern Africa. They had a unique view of the World and lived in temporary wooden and rock shelters as hunter-gatherers.

The |Xam language

Bushmen speak a variety of languages, but not many living speakers exist anymore. All of the languages incorporate 'click' sounds. The unique clicks used in Khoisan languages are varied and complex, as the written form of their language. In this project we will transcribe the text from the |xam language.

- Standard Characters in the |xam language
  - Composed of Basin Latin alpabet characters. e.g.a....z, A....Z
- Special Characters in the |xam language
  - Unique characters in text.

Figure 3.4: Descriptive History of the Hunter-Gatherer People


Transcribe Bleek and Lloyd
Tutorial
Applications Places
Fri Sep 28, 4:43 PM
Transcribe Bushman - Mozilla Firefox
File Edit View History Bookmarks ScrapBook Tools Help
Transcribe Bushman
Most Visited Getting Started Latest Headlines Bossa Digital Libra...
Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options
Transcribe Bushman
Home Transcription App Leaderboard About
Sign in
Log in
Digital Libraries Laboratory, Computer Science Department
University of Cape Town
digital libraries laboratory

boinc.cs.uct.ac.za/transcribe_bushman/login_form.php?next_url=bossa_apps.php

Sign into Transcribe Bushman

Email address: ngoni.munyaradzi@uct.ac.za

Password: ••••••

Stay logged in ☑

Figure 3.5: Transcription tutorial video

3.3 BOINC

BOINC [Anderson, 2004] (Berkeley Open Infrastructure Network Computing)[5] system software is an open-source cross-platform software tool for computing using resources harnessed from volunteers. The BOINC framework is developed by the Space Sciences Laboratory at the University of California at Berkeley. It was initially built to support the SETI@HOME[6] project, but later supported diverse research projects in fields like medicine, mathematics and climatology. BOINC was designed to be used by research scientists with the goal to advance the public-resource computing paradigm. This would be achievable by harnessing distributed volunteer computing power for computationally intensive projects. More compute power can be obtained through public-resource computing compared to the traditional central server supercomputing model. For example, SETI@home has a sustained processing rate of 70 TeraFLOPS in contrast to the NEC Earth Simulator, which provides 35 TeraFLOPS[Anderson, 2004]. Volunteers can participate in multiple BOINC-based projects by installing the BOINC client remotely, and linking these with the client software.

The BOINC scheduler manages distribution of work units and completed work. BOINC supports redundant computing, a mechanism for handling erroneous

[5] http://boinc.berkeley.edu
[6] http://setiathome.berkeley.edu/

results. It uses exponential backoff for clients connecting to a failed server once it recovers. The client software supports configuration of participant preferences e.g. the amount of disk space or CPU allocated to BOINC projects. Participants are credited for tasks completed in each project. Figure 3.6 illustrates how to Boinc software functions [Anderson, a].

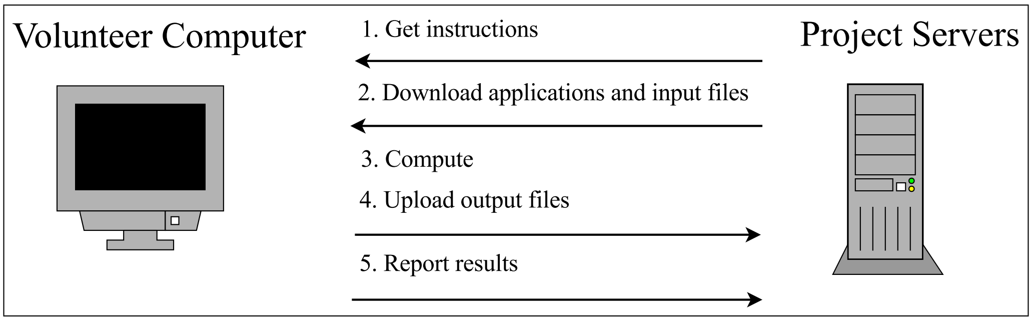


Figure 3.6: Boinc functionality

The BOINC software consists of several components, namely:

- The core client - manages network communication, scheduling, computation and the deployment and monitoring of applications.
- A client GUI - provides summary statistics of currently joined projects and has functionality to join and quit projects.
- An API - provides mechanisms to report system usage and reports on tasks completed.
- A screensaver program - interacts with the core client to provide screensaver graphics; this functionality is platform-specific.

The BOINC software manages user accounts, experimental data, groups and communication through message boards. BOINC interacts with the MySQL database to keep track of project details for both Bolt and Bossa; this includes the above-mentioned features. For the purposes of this project, when the BOINC server software was compiled only the Web application that includes Bossa and Bolt modules was enabled. The following section gives a system overview of the transcription tool.

3.4 System Overview

Figure 3.7 provides a cross-sectional view of the whole Bossa-based transcription tool, and shows how Bossa, Bolt and Boinc are integrated. The whole system is divided into three major layers, namely the back-end, middle-ware and front-end, all of which are modular. The MySQL database and experimental data for the project reside in the back-end. The database records the locations of the transcription images. The middle-ware layer handles user accounts, groups and job distribution. Lastly, the front-end handles the logic and layout of the transcription tool Web interface.

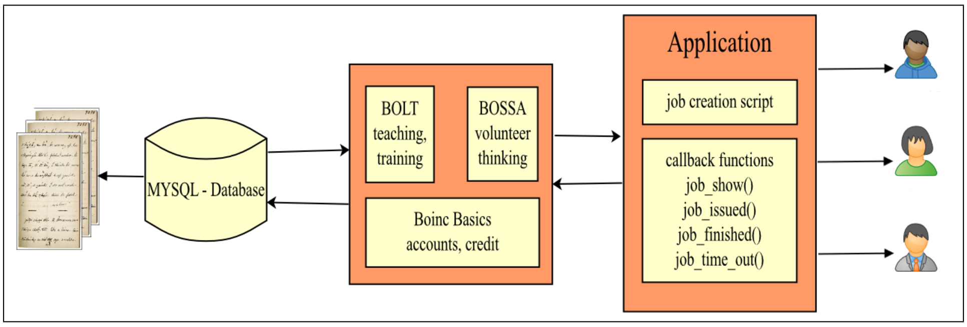


Figure 3.7: System Overview

3.4.1 Job Distribution Policy

A task in Bossa is defined into two parts, a job and an instance. A Bossa job is the assigned task the volunteers are supposed to complete; this can be likened to a HIT in Mturk. An instance is essentially the same as a job but instances represent the number of times that particular job has to be repeated by unique volunteers. An instance can be viewed as a subtask of job. In Bossa a job distribution policy defines how a project's jobs are managed. Factors to consider are: how many instances of a job should be distributed, what threshold values have been set for each job or which jobs have higher priority. In Bossa, applications have different job distribution policies, which are user-defined.

Below is a description of two projects with unique job distribution requirements:

- Project X - has a limited set of jobs and thousands of volunteers.
- Project Y - has an unbounded set of jobs but limited number of volunteers.

In the instance of Project X, where there is a limited set of jobs, the goal is to get all jobs completed the same number of times. The best job distribution policy would be to issue out all the jobs once; when completed, the jobs are issued out for a second or third time. More accurate results are obtained the longer the project runs.

Project Y has a targeted threshold of accuracy, and in this case each job is given out one at a time to a sufficient number of volunteers who can achieve this threshold. Once the threshold is reached, the second job is issued and so on. More jobs are completed the longer the project runs.

For this project a hybrid job distribution policy was used - a combination of the policies of Project X and Project Y. Like Project Y, which has an unbounded set of jobs, a dataset of 9800 pages was used for this project, but with no pre-determined threshold. As this project was Web-based, the expectation was to get thousands of volunteers online, hence all jobs were given once. A volunteer can transcribe as many pages as they like. A job replication policy was implemented to improve accuracy of results; this is explained in section 3.4.3.

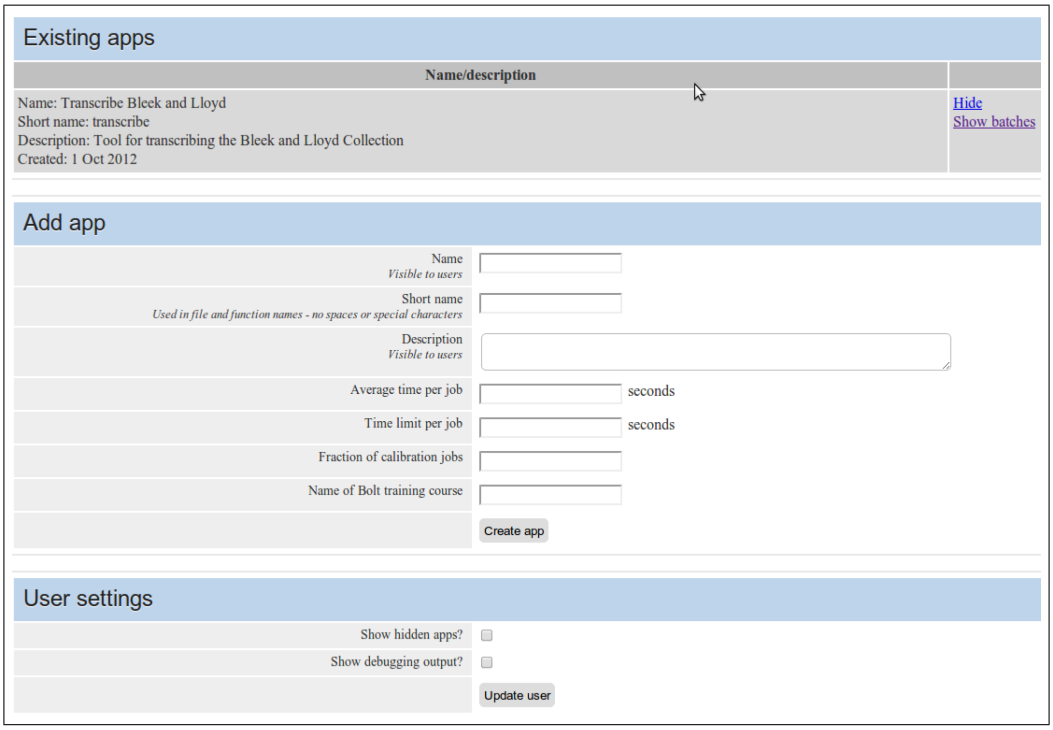
3.4.2 Bossa jobs and result Representation

Each job has a priority level and is defined in the project call back function. By default, Bossa distributes jobs based on decreasing priority level, but assigns the same priority to all jobs. This project implemented the default functionality. Bossa jobs have a number of states depending on the jobs' current progress. Below is a description of the different job states:

Bossa job states:

- Status 0: Job has been completed.
- Status 1: Job is still in progress but has not been issued to any user.
- Status 2: Job is still in progress and has been issued to a user.
- Status 3: No consensus was reached and job is classified to be inconclusive.
- Status 4: Job timed-out.

Bossa provides a Web interface where applications can be created - see Figure 3.8. Once the application was created, a job creation script was defined. The job creation script links the application registered in MySQL with the batch of transcription images. The four callback functions defined in section 3.1 were implemented to display the jobs on the transcription tool interface, and result representation within the database.



Existing apps

Name/description	
Name: Transcribe Bleek and Lloyd Short name: transcribe Description: Tool for transcribing the Bleek and Lloyd Collection Created: 1 Oct 2012	Hide Show batches

Add app

Name *Visible to users*	
Short name *Used in file and function names - no spaces or special characters*	
Description *Visible to users*	
Average time per job	seconds
Time limit per job	seconds
Fraction of calibration jobs	
Name of Bolt training course	
	Create app

User settings

Show hidden apps?	□
Show debugging output?	□
	Update user

Figure 3.8: Admininistrative Interface

Each image is represented as a single job. The name and file path of the image are stored in a PHP data structure called an *opaque object*. The results of each job are also stored within this multi-dimensional data structure. The transcription tool was implemented as a single Web page.

3.4.3 Replication Policy

Bossa supports the use of two replication policies: (1) Fixed and (2) Adaptive replication. Fixed replication has a set number of instances that are issued, whereas adaptive replication depends on whether the accuracy threshold for the job has been reached. This project adopts the fixed replication model because adaptive replication cannot be supported without a known solution for the problem or solution fitness function.

Each job is repeated three times, and any given instance is issued to a unique volunteer. In the research by [Lee and Hu, 2012] for music mood classification, they collect three relevance judgements from participants. [Lee, 2010] again collected three judgements for music similarity. [Marge et al., 2010] used two workers to produce transcriptions in the first phase of their experiment. For the second phase they collected three transcripts, making a total of five users for each transcription. This methodology is adopted based on the assumption that as multiple volunteers work on a transcript, they will likely produce an accurate transcription. For a particular job, if three volunteers reach consensus on how a page is transcribed, the job is classified as COMPLETED. If more than five instances of a job have been issued, and there is no consensus amongst the volunteers, it is classified as INCONCLUSIVE. No time limits were set for jobs, as this would deter volunteers from contributing to the project. The following section briefly discusses additional tools used in the implementation of the transcription tool.

3.5 Technologies and Tools

FFMPEG

FFMPEG[7] is an open source tool to record, convert, stream and play multimedia content. This tool was used to create the video screencast tutorial on how to transcribe the |Xam and !Kun languages of the hunter-gatherer people of Southern Africa.

Bootstrap and Fontawesome

Bootstrap[8] is a front-end toolkit for rapidly developing Web applications. It is a collection of CSS, Javascript and HTML conventions. Fontawesome[9] is a set of iconic fonts that were designed for use with Twitter Bootstrap.

Additional Technologies

PHP, CSS, Javascript and HTML were used in the development of the user interface. The xFenster library[10] from the Cross-Browser website was used to display the characters and diacritics on a palette.

3.6 Transcription Tool

This section describes the features of the transcription tool and the process of completing transcriptions.

3.6.1 Login, Registration and Qualification

So as to lower the barriers that hinder volunteers from participating in volunteer crowdsourcing projects, the process of signing-up and training volunteers is simple and short. Once a volunteer registers, they are required to watch a short transcription tutorial video first (see section 3.2). After the transcription tutorial, the user can begin transcribing. Other crowdsourcing projects

[7] http://ffmpeg.org

[8] http://twitter.github.com/bootstrap

[9] http://fontawesome.github.com/Font-Awesome

[10] http://cross-browser.com/x/examples/xfenster-demo.php

require users to complete an assessment exercise to determine volunteer skill. This was not done for this project, as the project aimed to assess the overall accuracy achieved by the volunteers.

3.6.2 Characters and Diacritics Panel

More than 300 diacritics of the |Xam and !Kun languages are used in the transcription tool. Still more diacritics are being discovered in the notebooks. These languages are not supported in standard Unicode representation. A specialized encoding tool was developed by [Williams and Suleman, 2011a] to represent this complex script. The custom encoding tool was developed using LaTeXand the TIPA package. The TIPA package has a limited set of similar diacritics but it supports the creation of new nested and stacked custom diacritics. See Figure 3.9.

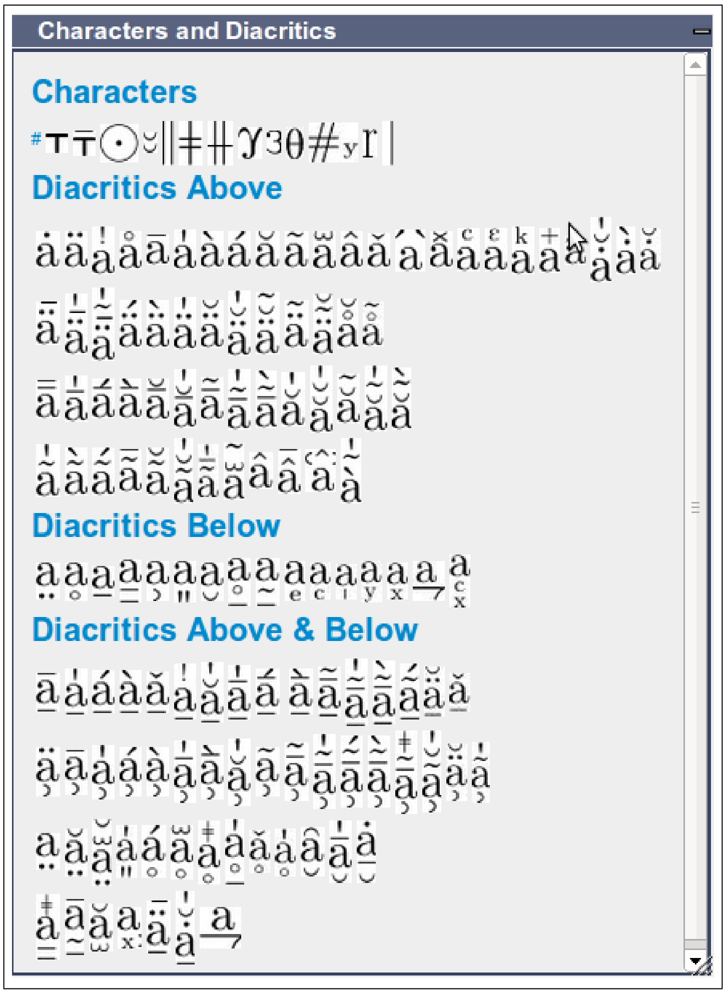
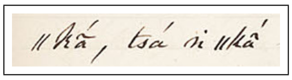


Figure 3.9: Characters and Diacritics Panel

The visual representation of the encoding is a near approximation of the text in the notebooks. Future work as suggested by [Williams and Suleman, 2011a]

Table 3.1: |Xam & !Kun Text Encoding and Representation

Text	*Encoding*

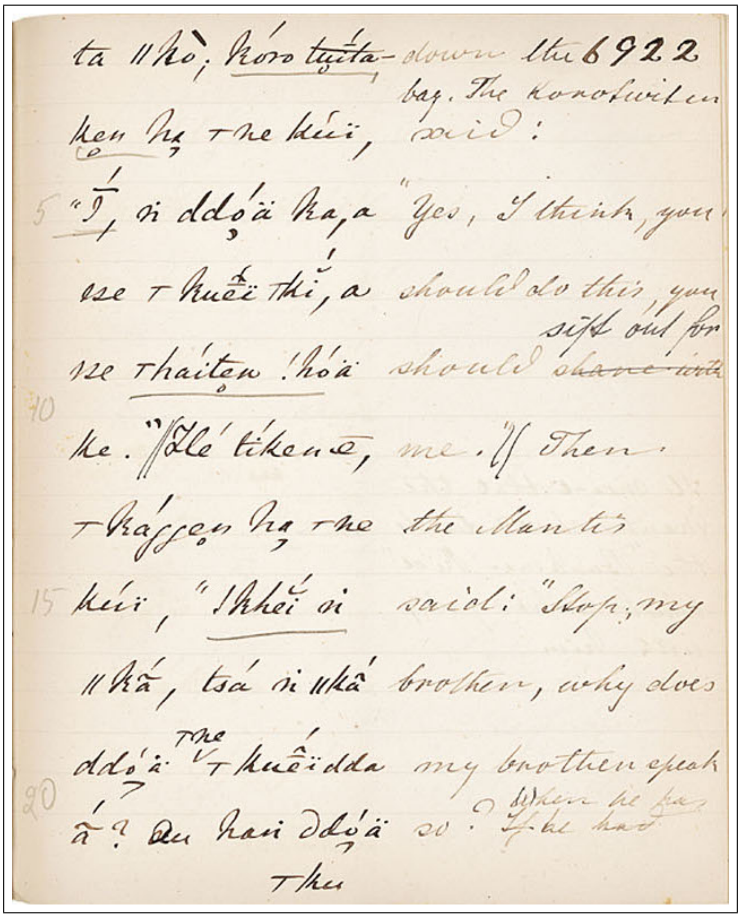

//kā́, tsá ṅ //kā́
	\textdoublepipe{}k\elipline{a},
	ts\dialine{a}\onedot{n}
	\textdoublepipe{}k\elipline{a}

Representation	*English Translation*
||kā́, tsá ṅ ||kā́	the Mantis

would be to develop a custom font for the languages Table 3.1 shows sample |Xam text, the equivalent encoding, visual representation and the English translation.

3.6.3 Transcription Task

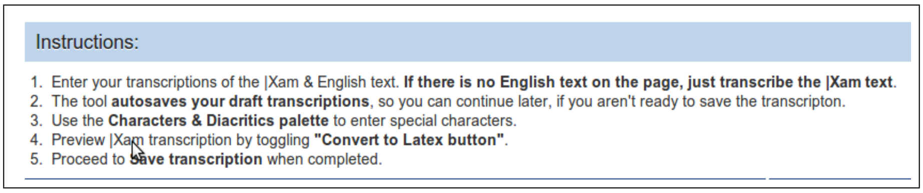
For the transcription task, volunteers are assigned an image from the Bleek and Lloyd collection with |Xam and English text. Volunteers are then instructed to transcribe the text that appeared on the right side page of the image, and include the most appropriate characters and diacritics for the |Xam text. The |Xam and English text are grouped into two columns, (see Figure 3.10.) Volunteers are also instructed not to transcribe the text that appears in the side margins or on the left side of the page. If an image could not be transcribed for some reason, volunteers are told to click on the *Cannot Transcribe Page* button (see Figure 3.12.) The |Xam and English text are supposed to be typed into the left and right textareas respectively. Once a volunteer completes transcribing a page, they would then click on the *Finish and Exit* button.



ta ||kò; kóro tuïta- down the 6922
bay. The Korotuïten
ken ha ≠ne kúï, said:
"Í, ñ ddóä ka, a "Yes, I think, you
ñe ≠kuẽï thĩ, a should do this, you
sift out for
ñe thaiten !kóä should ~~shave with~~
ke. "||Ke' tikenã, me." Then
≠kággen ha ≠ne the Mantis
kúï, "!khéï ñ said: "Stop, my
||kã́, tsá ñ ||kã́ brother, why does
ddóä ≠ne ≠kuẽïdda my brother speak
ã? Au kaň ddóä so? ~~When he has~~ If he had
≠ku

Figure 3.10: Layout of Sample Transcription Image

Further instructions on how to use the transcription tool were embedded above the transcription tool interface. Figure 3.11 is a snippet of the instructions, which are simple and short and emphasis is put on the important points.



Instructions:

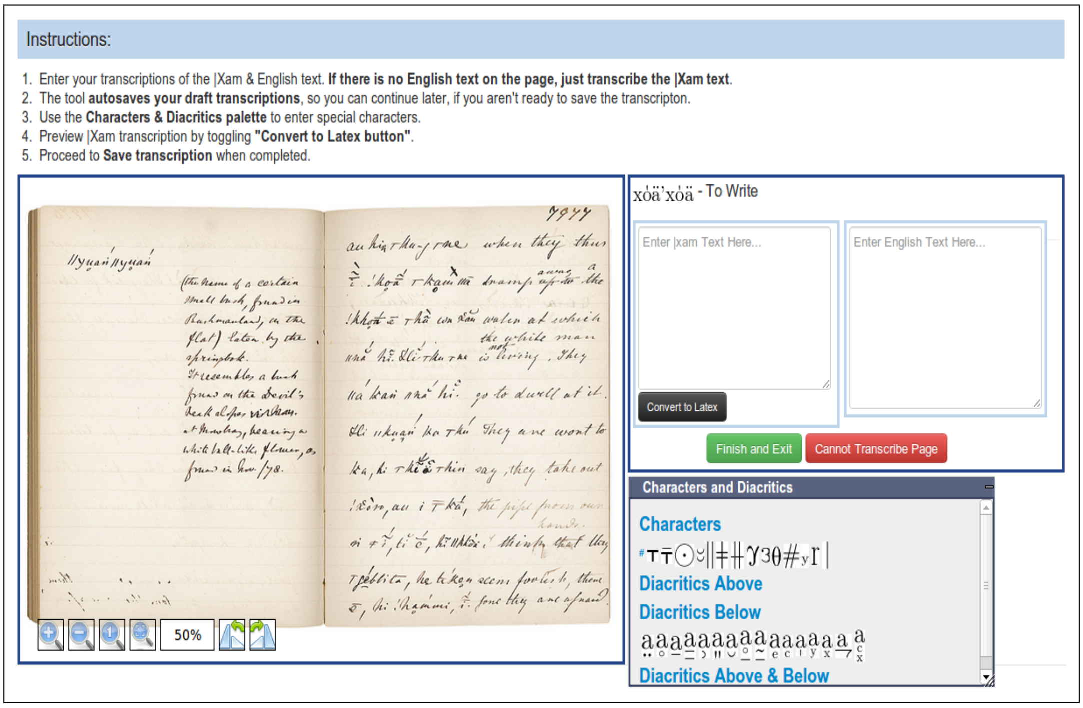
1. Enter your transcriptions of the |Xam & English text. **If there is no English text on the page, just transcribe the |Xam text.**
2. The tool **autosaves your draft transcriptions**, so you can continue later, if you aren't ready to save the transcripton.
3. Use the **Characters & Diacritics palette** to enter special characters.
4. Preview |Xam transcription by toggling **"Convert to Latex button"**.
5. Proceed to **Save transcription** when completed.

Figure 3.11: Transcription Instructions

3.6.4 Transcription Interface

During the design and implementation of the transcription tool, three experts from computer science were used to evaluate the layout of the project website and the transcription tool. The research on usability testing by [Nielsen and Landauer, 1993] showed that no more than five users are needed in evaluating a system. This evaluation was conducted in the form of a live demonstration of the system, and audio recordings were used to note feedback.

A simplistic design was used for the transcription tool interface, to cater for the varying volunteer skill. The affordance of the text inputs resembled the columns of text within the Bleek and Lloyd notebooks. The |Xam and English text would appear either in the left or right column of a page. The layout of the interface is illustrated in Figure 3.12 below.



Instructions:

1. Enter your transcriptions of the |Xam & English text. **If there is no English text on the page, just transcribe the |Xam text.**
2. The tool **autosaves your draft transcriptions**, so you can continue later, if you aren't ready to save the transcripton.
3. Use the **Characters & Diacritics palette** to enter special characters.
4. Preview |Xam transcription by toggling **"Convert to Latex button"**.
5. Proceed to **Save transcription** when completed.

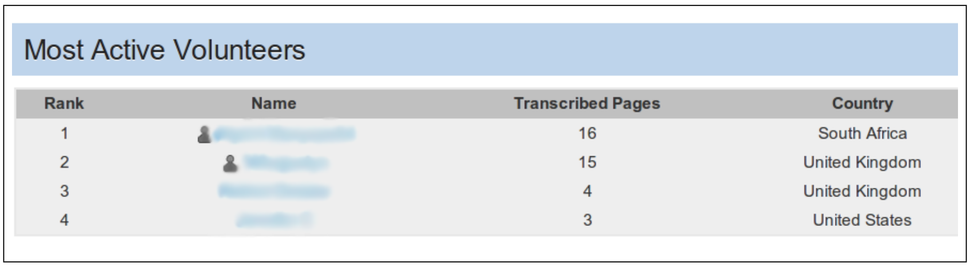
Figure 3.12: Transcription Interface

The red button in the image was an option to indicate whether a page could be transcribed or not. The green button was the Finish and Exit option once a volunteer finished transcribing a page. The black button was to preview the |Xam text. To better improve viewing the transcription images, zooming in and out features were included.

3.6.5 Motivational Features

Numerous studies [Budhathoki, 2010, Hossain, 2012, Kaufmann et al., 2011] has been conducted to understand why volunteers participate in crowdsourcing projects. The results suggest that people not only contribute to crowdsourcing projects for monetary reasons, but there are intrinsic benefits to be gained. Some of these are recognition for contributions made, interest in the research findings or because of the competitive aspect of the project. Three features were implemented in this project to motivate users to contribute.

1. Leaderboard - this is a ranked list of the top four active users in the project. (See Figure 3.13)
2. User of the Day - this is a feed that displays the profile of a recent user who joined the project.
3. Badges - these are earned depending on number of contributions made.



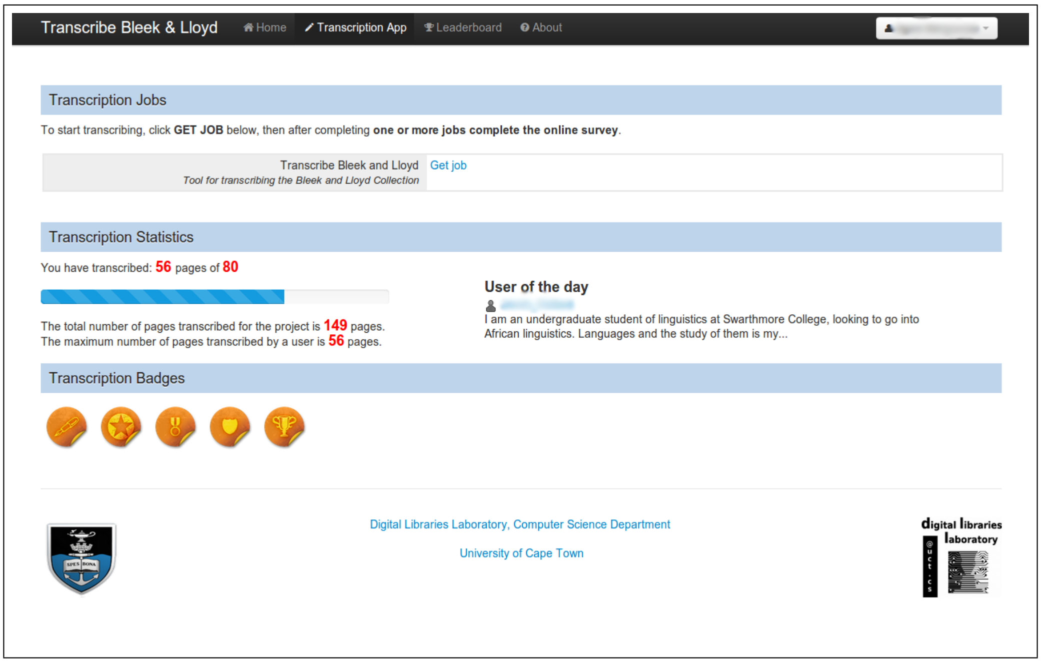
Most Active Volunteers

Rank	Name	Transcribed Pages	Country
1	[redacted]	16	South Africa
2	[redacted]	15	United Kingdom
3	[redacted]	4	United Kingdom
4	[redacted]	3	United States

Figure 3.13: Leaderboard

3.6.6 User Statistics

When designing a crowdsourcing project, it is important to give feedback and credit to users regarding their performance or project progress. A crowdsourcing project has to have an end goal or target set. This helps to motivate users to participate. Figure 3.14 below is an illustration of the user account page for the Transcribe Bleek and Lloyd project.



[redacted]

Transcription Jobs

To start transcribing, click **GET JOB** below, then after completing **one or more jobs complete the online survey**.

Transcribe Bleek and Lloyd
Tool for transcribing the Bleek and Lloyd Collection
Get job

Transcription Statistics

You have transcribed: **56** pages of **80**

The total number of pages transcribed for the project is **149** pages.
The maximum number of pages transcribed by a user is **56** pages.

User of the day

[redacted]

I am an undergraduate student of linguistics at Swarthmore College, looking to go into African linguistics. Languages and the study of them is my...

Transcription Badges

Digital Libraries Laboratory, Computer Science Department

University of Cape Town

Figure 3.14: User Statistics Interface

The user account page displays the following information:

- Number of pages transcribed by logged-in user.
- The user of the day.
- A progress bar.
- Highest number of transcriptions achieved by any single user.
- Badges currently earned by a user. Badges are earned for a certain number of pages transcribed.

3.7 Transcription Tool Deployment

The deployment of the transcription tool was executed in two phases. The deployment phases were part of the iterative design of the tool.

3.7.1 Phase 1

During the first phase a number of invitational emails were sent out to communities in Africa, conducting research based on the hunter-gatherer people of Southern African. Invitations were also sent to our collaborating partners from the Citizen CyberScience Centre in Geneva. The purpose of phase 1 was to get essential user feedback about the functionality of the transcription tool and identify any potential bugs. An online user survey was used to collect user feedback with regard to user experience and general comments regarding the transcription tool. The feedback regarding the user experience using the transcription tool is discussed in detail under section 4.7.2. In this section, only the feedback regarding the functionality of the transcription tool is addressed.

Useful feedback regarding the functionality of the tool was obtained, and some of the comments were taken into consideration in the following implementation phase. For example, one user suggested including a feature where a job could be paused and finished at a later time. This was essential because the transcription task would generally take about 15-20 minutes, and would avoid loss of data in the case that the volunteer chose to close the browser or any other scenario that could arise before completing the task. Another user suggested that Glyphs[11] be used to represent the diacritics within the Bleek and Lloyd notebooks. Glyphs can be used to create custom fonts for the languages and this would be ideal for the script used in the notebooks. This suggestion was not possible to implement as the Glyphs would not capture the complexity of the diacritics, as diacritics can span multiple characters and be combined in many orderings. A volunteer with linguistics background helped identify a

[11] http://en.wikipedia.org/wiki/Glyph

new diacritic that was then included in the Characters & Diacritics palette. Initially, the button meant to be used when submitting a transcription after completion was labelled “Save”; this was a cause of confusion to some volunteers, and they suggested that labelling it “Finish and Exit” would be better. Overall, phase 1 was a success, as some important issues were identified and resolved.

3.7.2 Phase 2

The second phase was targeted to the broader online volunteer community. Once the new features had been incorporated, and adequate system testing performed, more invitation requests were sent out. Multiple forms of communication media were exploited to create increased awareness online of the transcription project:

- Social Media - Facebook[12] and Twitter[13]
- Online Message boards - Boinc platform[14] and Slashdot[15]
- Online Science Magazines - MyScienceWork[16] and International Science Grid this Week[17]
- Blogs - Rosetta Stone project[18] and Ben Brumfield’s Blog[19]
- Micro -Volunteer Websites - Sparked[20] and Volunteer Match[21]

3.8 Summary

This chapter has described the design and implementation of the Bossa-based transcription tool. The chapter first discussed in detail the core technology tools used for the transcription tool, namely: (1) Boinc (2) Bolt and (3) Bossa. An overview illustrating the interaction of the three core tools is given for the transcription tool, followed by description of the job distribution and replication policy implemented for this project. Also mentioned are the third party tools used. Lastly, the chapter focuses on the transcription tool features and the deployment phases for the project.

[12] www.facebook.com

[13] www.twitter.com

[14] http://boinc.berkeley.edu

[15] http://slashdot.org

[16] http://www.mysciencework.com/

[17] http://www.isgtw.org/

[18] http://rosettaproject.org/blog/02012/oct/12/help_transcribe_historical_language/

[19] http://manuscripttranscription.blogspot.com/2013/02/ngoni-munyaradzi-on-transcribe-bleek.html

[20] http://www.sparked.com/

[21] http://volunteermatch.org/

Chapter 4

Experimentation and Results

The purpose of this chapter is to answer the project's research questions through experimental methods and analysis of the user experience with the transcription tool. The chapter begins by stating the research questions, then describes the experimental data and volunteer transcription challenges. A corpus analysis is performed, followed by a description of the experiments and user survey. Finally the chapter ends by summarising the findings of the research.

4.1 Research Questions

This project aimed to investigate:

- If volunteer thinking can be used to crowdsource intellectually-intensive tasks in digital libraries (like transcribing handwritten manuscripts).
- How volunteer thinking compares to machine learning techniques when applied to the problem of transcription.

4.2 Experimental Data

The pilot project has run for about 22 weeks. In that time 179 volunteers have registered and 233 transcriptions have been submitted. From the 233 transcriptions, a total of 1551 lines of |Xam and 1389 lines of English text were collected. Of note, the most active user has transcribed 62 pages, followed by a user with 41 transcriptions. A number of the most active volunteers who participated in the project have a background in linguistics or related research, while other participants have contributed in other crowdsourcing projects like GalaxyZoo [Cook, 2011].

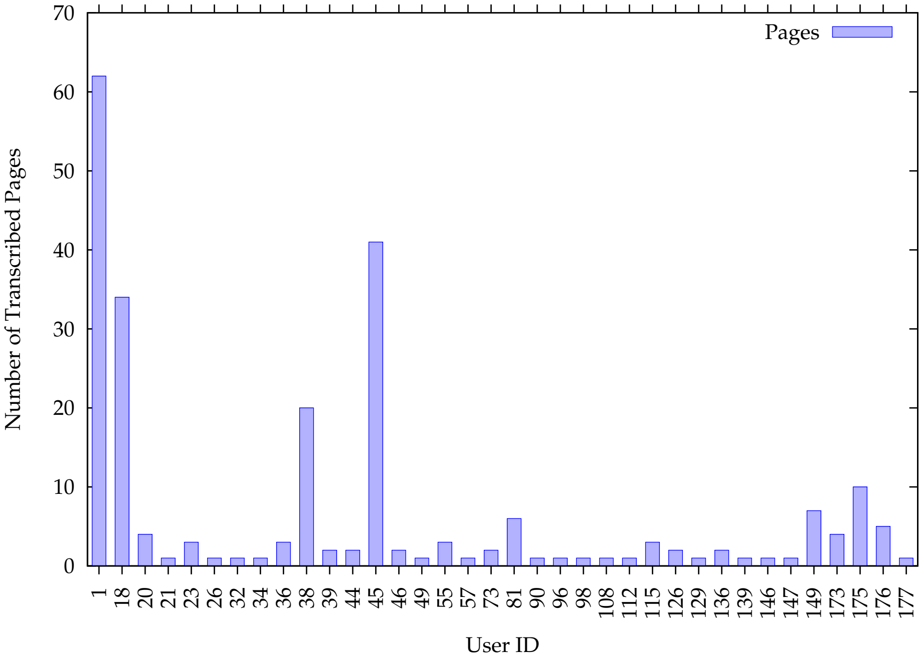


Figure 4.1: Distribution of transcriptions across volunteers

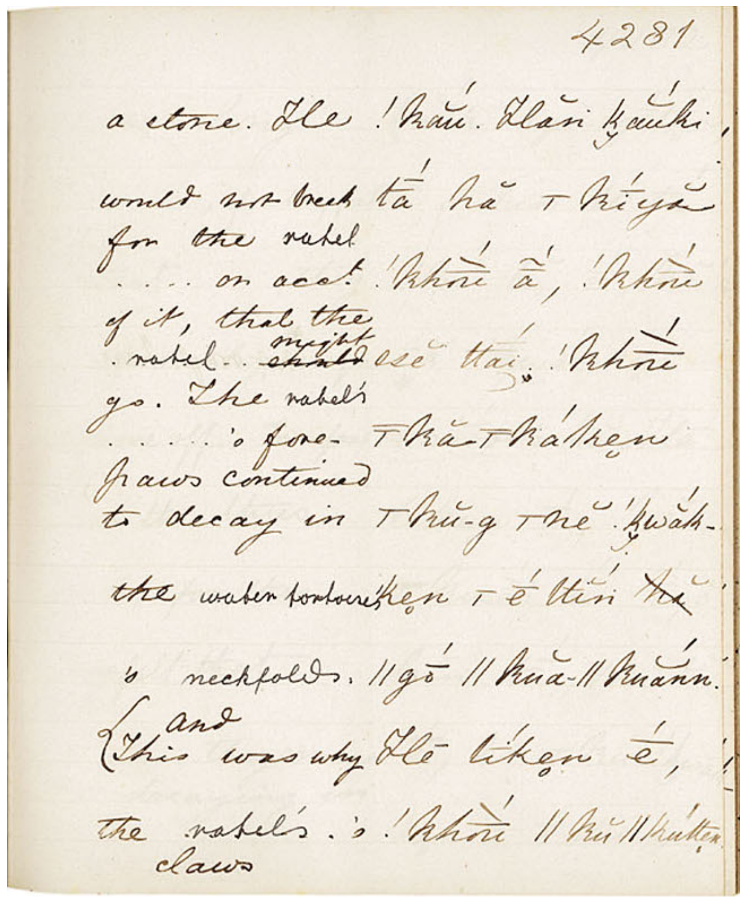
Figure 4.1 shows the number of transcriptions submitted by individual volunteers. A total of 36 volunteers made contributions to the project; the other 143 volunteers did not submit any transcriptions. A possible reason for this is that the transcription tool was advertised to a class of undergraduate linguistics students in the United States of America to view the transcription tool, who then registered for the project out of interest. Further analysis of the jobs distributed to volunteers showed a number of jobs assigned to volunteers had not been completed and were re-assigned. [Lee and Hu, 2012] noted that volunteers do not always complete tasks they start. Transcriptions from one volunteer were discarded as they were spam.

4.3 Transcription Challenges

Different challenges were experienced by volunteers while transcribing the manuscripts, the challenges are discussed below. Some of the challenges could potentially have a negative impact on the experimental results. Some of the challenges observed relating to confusion regarding the data were later resolved in section 4.4.1 under corpus pre-processing. As mentioned earlier in section 3.2, a video tutorial was used to train volunteers on how to transcribe images from the notebooks. Though this method was used, some transcribers still faced challenges.

Inconsistent layout

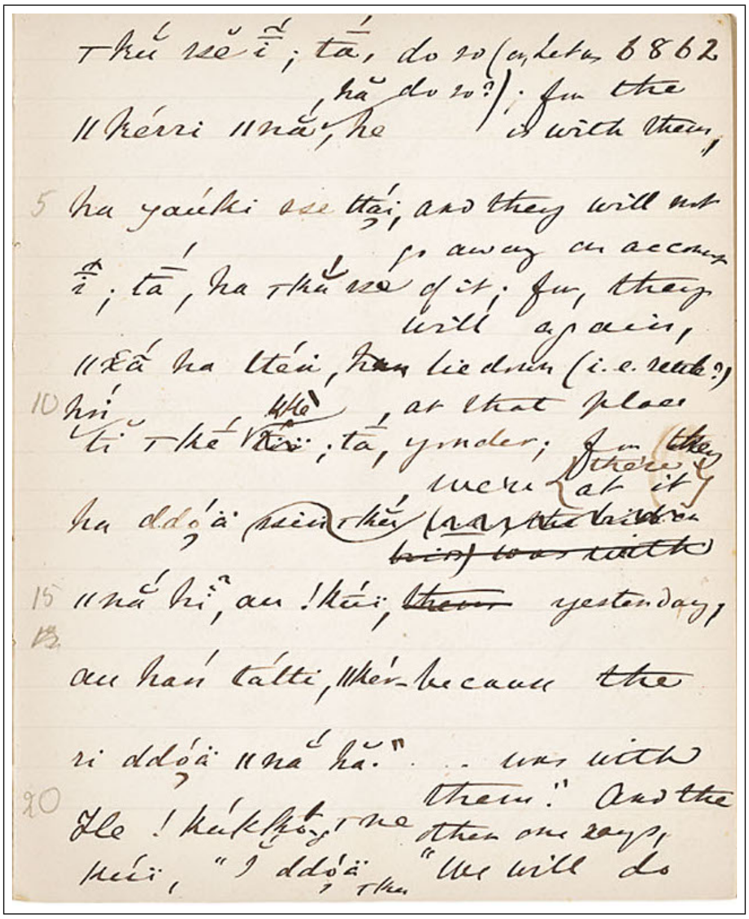
At times the |Xam text would appear on the left column on the image and then on the right side. Some of the users were confused about how to deal with the inconsistent layout of the text on the images. Figures 4.2 and 4.3 below demonstrate such a case. This point had been clearly explained in the video tutorial on how to transcribe this text.



4281

a stone. ǂle !kau. ǂlan kauki,
would not break tá há ǂkiya
for the ratel
. . . on acct. !khwe ā, !khwe
of it, that the
. ratel . . ~~shall~~ might ose thai, !khwe
go. The ratel's
. . . . 's fore- ǂkaǂkakken
paws continued
to decay in ǂkuiŋ ǂne !kwah-
the water tortoise ken ǂe ||kin ~~ha~~
's neckfolds. ||go ||kua-||kuann.
{and This was why ǂle likken e,
the ratel's . 's !khwe ||ku ||khutten
claws

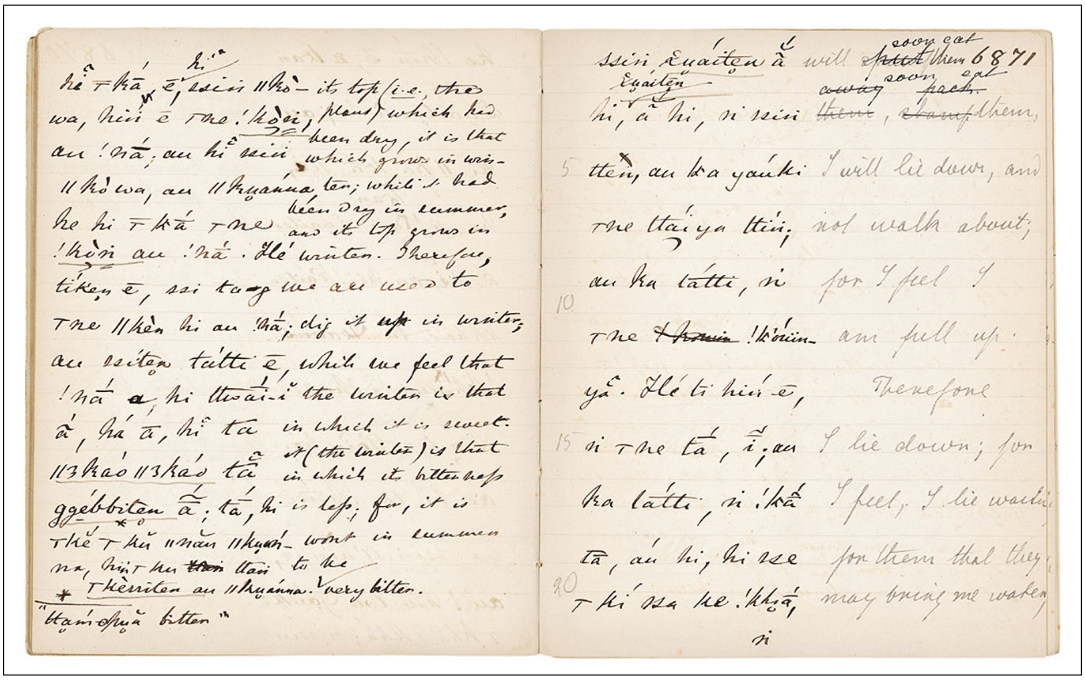
Figure 4.2: |Xam Text on Right Side



ᵀkù se ĭ; tā, do so (or, better, 6862
ha do so?); for the
||kerri ||nā, he is with them,
5 ha yauki sse ttai; and they will not
go away on account
ĭ; tā, ha ᵀkù se of it; for, they
will again,
||xā ha ttēi, ~~ha~~ lie down (i.e. settle?)
10 hĭ, at that place
ti ᵀkē ||kei; tā, yonder; for ~~they~~
there &
were {at it}
ha ddóä ~~(was the children~~
~~being with~~
15 ||nā hi, au !kúi; ~~them~~ yesterday,
au hau tatti, ||kei - because the
ri ddóä ||nā hā." ... was with
20 them." And the
ꞁle !kukkō, ᵀne other one says,
kúi, "I ddóä ᵀkù "We will do

Figure 4.3: |Xam Text on Left Side

Some of the volunteers were confused about which text to transcribe in cases where the transcription image had text on both sides of the image, (see Figure 4.4). This was observed during data cleaning, and the additional irrelevant transcriptions were filtered out. This inconsistency had been explained in the transcription tutorial video. Volunteers were expected to transcribe only the |Xam and English text that appeared on the right side of the page.



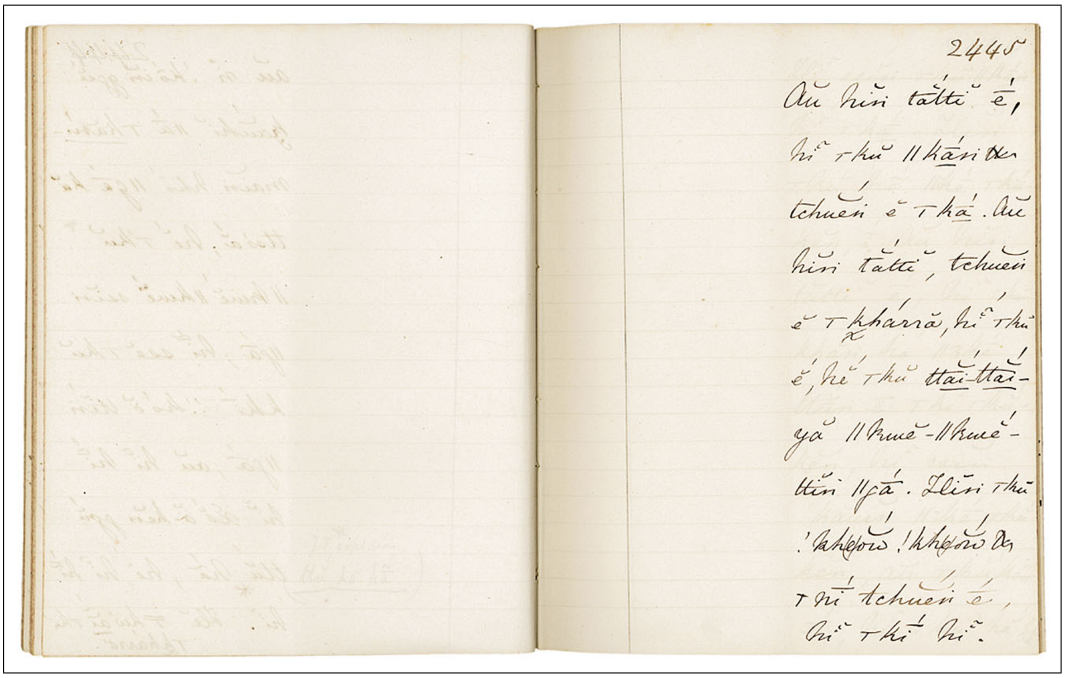
he ǂkā ē, ssin ||kō – its top (i.e. the
wa, hin ē ǂne !kōi, plant) which had
au !nā; au hi ssin been dry, it is that which grows in win-
||kōwa, au ||kuanna ter; while it had been dry in summer,
he hi ǂkā ǂne and its top grows in
!kōri au !nā. ||é wainten. Therefore,
tiken ē, ssi ta we are used to
ǂne ||kēn hi au !nā; dig it ~~up~~ in winter,
au ssiten tatti ē, while we feel that
!nā ~~a~~, hi ||wai-i the winter is that
ā, há ā, hi ta in which it is sweet.
||3kaō ||3kaō ta it (the winter) is that
ggebbiten ā; ta, hi in which its bitterness is less; for, it is
ǂhé ǂkei ||nāu ||kuan– wont in summer
na, hiŋ ǂkei ~~tten~~ ||an to be
* ǂkerriten au ||kuanna. very bitter.
"||kanixua bitter"

ssin ẑuaiten ā will ~~spatt~~ soon eat them 6871
ẑuaiten
hi, ā hi, n ssin ~~them~~ ~~away~~ soon, ~~stamp~~ ~~pack~~ eat them,
tten, au ka yauki I will lie down, and
ǂne ||taiya ||tin; not walk about;
au ka tatti, n for I feel I
ǂne ~~ǂkouin~~ !kouin- am full up.
yā. ||é ti hin ē, Therefore
n ǂne ta, ī; au I lie down; for
ka tatti, n !kā I feel, I lie waiting
tā, au hi, hi sse for them that they
ǂhi ssa ke !khoā, may bring me water,
n

Figure 4.4: Notebook Inconsistencies

Missing Information

Most pages of the scanned notebooks had two columns of text inscribed in them; one column was for the |Xam language and the other column contained the English translation. A few pages contained just one column of |Xam text. To avoid confusion amongst the volunteers, specific post-video tutorial instructions were posted onto the website on how to deal with this type of inconsistency. An example instruction used for this situation stated that, volunteers were to still transcribe the |Xam text into the appropriate box on the transcription tool. Figure 4.5 illustrates the point.

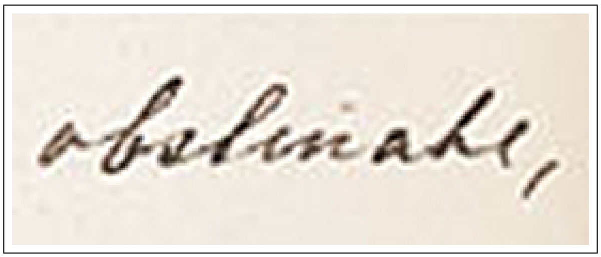


2445
Aŭ hiñ tátti ē,
hĭ ⊤kŭ ||kárri tē
tchuēn ē ⊤ká. Aŭ
hiñ tátti, tchuēn
ē ⊤kkhárrā, hĭ ⊤kŭ
ē, hĭ ⊤kŭ ttái-ttái-
ya ||kuē-||kuē-
ttiñ ||ā. Hiñ ⊤kŭ
!khóu !khóu ta
⊤nĭ tchuēn ē,
hĭ ⊤kĭ hĭ.

Figure 4.5: Missing English Text

Text Legibility

With some of the images (see Figure 4.6), the legibility of the text was very poor due to faint ink, inconsistent writing or squashed up words due to limited space on the page. This proved to be challenging for some volunteers; they ended up making their best guess or used their own understanding of the language to decipher the meaning. Some of the comments obtained during the survey from users support this point: *"The text is really difficult to read"*. Figure 4.6 below illustrates this point. The word inscribed is *obstinate.* In this case, auto-correct was used to decipher the word.

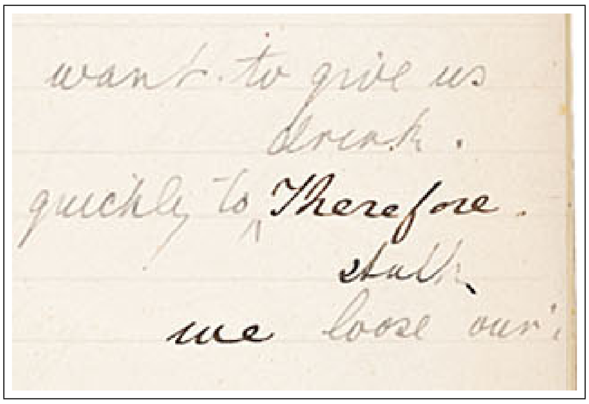


obstinate,

Figure 4.6: Challenges in Deciphering Text

Figure 4.7 shows an example where the legibility of the inscribed text is poor. Some volunteers did not know how to transcribe these lines. The transcription of this image should read: *"want to give us quickly to drink Therefore shall we loose our:"*. One volunteer transcribed this as: *"want to give us drink.*

Therefore shall we quickly loose our:". They failed to see that the word *drink and shall* had been inserted. This case could not be resolved as this was part of determining the accuracy of volunteers. Other users included deleted text in their transcriptions, but it was assumed that they would only transcribe relevant text. Cases like this would affect the quality of results.



want to give us
drink.
quickly to Therefore.
shall
we loose our:

Figure 4.7: Additionally Inserted Text

Alignment of Diacritics

In multiple cases, volunteers were confused on how to transcribe certain words with regard to alignment of diacritics and characters. Figure 4.8 shows two words that are likely similar, but can be recognized as being different due to the alignment of diacritics. A transcriber could try and correct this or transcribe the text as is. This is likely to affect accuracy of results.

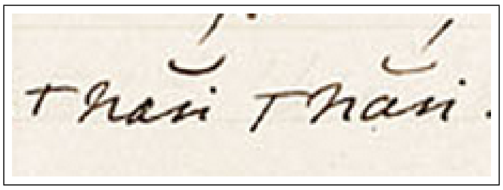


Figure 4.8: Confusion of diacritic placement

Summary

The discussion in this section has outlined the noted sources of confusion for transcribers during the transcription phase. Some of the issues raised in this section were easily addressable, while others were not. The challenges highlight the difficulty of transcribing complex |Xam text. Measures were put in place to address some of the potential challenges for volunteers, for example the zooming in & out feature to scroll the image. Some issues mentioned have a high likelihood of affecting the results obtained when calculating the inter-transcriber agreement and transcription accuracy. In the next section, measures taken to rectify these sources of errors are discussed.

4.4 Corpus Analysis

This section discusses the pre-processing and transformations applied to the corpus. The experimental data composed of transcriptions obtained from volunteers is termed Corpus-V.

4.4.1 Corpus-V Pre-processing

Punctuation and Space Alignment

Figure 4.9 shows a case where the placement of the semi-colon is not clearly defined due to author mistakes. This was a source of confusion for some volunteers. Based on the norms used in English grammar, it is expected that a semi-colon appears immediately after a word, but in this case it appears closer to the second word. Not all the volunteers corrected this inconsistency in the text, and thus further analysis had to be done. Other characters that had to be corrected were: the period, colon and exclamation mark.

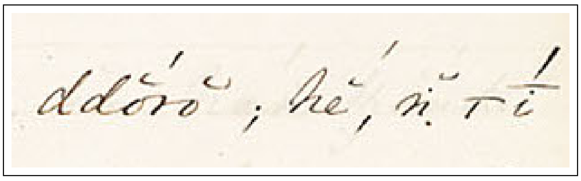


Figure 4.9: Placement of punctuations

Case standardization

The normalized Levenshtein algorithm used in the calculation of the inter-transcriber agreement and transcription accuracy is sensitive to minor differences in the text. The algorithm measures the similarity of two strings based on the alignment of the text. In the case that one of the strings is typed out in upper case *FATHER* and the other in lower case *father*, these would be recognized as different texts. Hence to reduce possible inaccuracies in measurement of transcription similarity or accuracy, all the characters were converted to lower case.

Deletion of volunteer comments

A volunteer who transcribed Figure 4.7 could not read the text in the image and put placeholder characters in the transcription: *Therefore/.....?/we loose our/....?/.* In another case a volunteer transcribed the image as *see thee at [......]Bushman rice, and thou must often[....], tomorrow* showing more instances where volunteers were unsure about the true representation of the text. The confusion is attributed to faint text lines and inconsistent handwriting used by the author. Another volunteer included a comment in the transcription stating that he/she could not transcribe a particular section of the text: *up, he again hove??? (can't transcribe) flapping his wings.* For the same text a different transcriber just used placeholder text indicating uncertainty of the word: *he again hovered (?) flapping his wings.* The placeholder text and comments were filtered out as it would be noise within our test data, and would affect overall accuracy. No standard placeholder was used as this would increase the amount of irrelevant text to be filtered out for experiments.

Deletion of irrelevant transcriptions

Figure 4.4 has text transcribed on both sides of the page. One volunteer consistently transcribed the text on both pages in a number of transcriptions they produced; it had been clearly mentioned that only text on the right of the page was required. This additional text was deleted, and only the relevant text kept for experimentation. Each page had an associated page number, and in some cases the volunteers would also transcribe the page number into the text. The page numbers were discovered using a manual process. This information was removed as this would affect calculations of the inter-transcriber agreement and transcription accuracy of the |Xam or English text. The next two sections describe experiments conducted to evaluate the quality of transcriptions produced by volunteers.

Further transformation

During analysis of the corpus of volunteer transcriptions, it was observed that some volunteers consistently mis-transcribed specific characters that appear in the language. For instance, volunteers would use the character *T* instead of the special character defined in the characters and diacritics palette. To also have consistency with the gold standard data, that was going to be used in later experiments, the exclamation mark *!* had to be transformed to the one defined in the gold standard \excl{}. No other transformations were perfomed for the diacritics or base symbols that appear in the language.

4.5 Transcription Similarity

The Levenshtein distance [Lcvenshtcin, 1966] or edit distance is a measure of the similarity between strings. It can be defined as the minimum cost of transforming string X into Y through basic insertion, deletion and substitution operations. This method is popularly used in domains of pattern recognition and error correction. This method is not suitable to solve certain problems as the method is sensitive to string alignment; noisy data would significantly affect its performance. The method is also sensitive to string lengths; shorter strings tend to be more inaccurate if there are minor errors than longer strings. [Yujian and Bo, 2007] note that because of this there is need for a normalized version of the method.

Notation-wise, Σ represents the alphabet, Σ^* is the set of strings in Σ and $\lambda \notin \Sigma$ denotes the null string. A string X $\in \Sigma^*$ is represented by X $=x_1x_2...x_n$, where x_i is the ith symbol of X and n is the length of the string calculated by taking the magnitude of X across $x_1x_2...x_n$ or $\mid X \mid$. A substitution operation is represented by $a \rightarrow b$, insertion by $\lambda \rightarrow a$ and deletion by $b \rightarrow \lambda$. $S_{x,y}$ $= S_1S_2...S_u$ are the operations needed to transform X $\rightarrow$ Y. γ is the weight function equivalent to a single edit transformation that is non-negative, hence the total cost of transformation is $\gamma(S_{x,y}) = \Sigma_{j=1}^{u}\gamma(S_j)$

The Levenshtein distance is defined as:

$$LD(X,Y) = min\{\gamma(S_{x,,y})\} \tag{4.1}$$

[Yujian and Bo, 2007] define the normalized Levenshtein distance as:

$$NLD(X,Y) = \frac{2 \cdot LD(X,Y)}{\alpha(\mid X \mid + \mid Y \mid) + LD(X,Y)} \tag{4.2}$$

where the result is a number within the range 0 and 1, where 0 means the strings are different and 1 means they are similar. where α = max$\{\gamma($a$\rightarrow \lambda)$, $\gamma(\lambda \rightarrow b)\}$

Methodology

The normalized Levenshtein distance metric was used to measure the inter-transcriber agreement. The inter-transcriber agreement metric can be used to assess how reliable the data from volunteers is. In this experiment the assumption is that similar transcriptions have lower Levenshtein values.

The inter-transcriber agreement is calculated at line level. The overall agreement amongst documents can be trivially calculated using the compound sum of each individual line in a document. During the data collection phase, each individual page was transcribed by up to three unique volunteers. From the individual transcriptions, each line is compared with the other two for agreement. Figure 4.10 illustrates the methodology used to calculate the inter-transcriber agreement. Transcription 1 & 2 have a agreement value of 0.65; Transcription 1 & 3 score 0.80; and Transcription 2 & 3 have a score of 0.50.

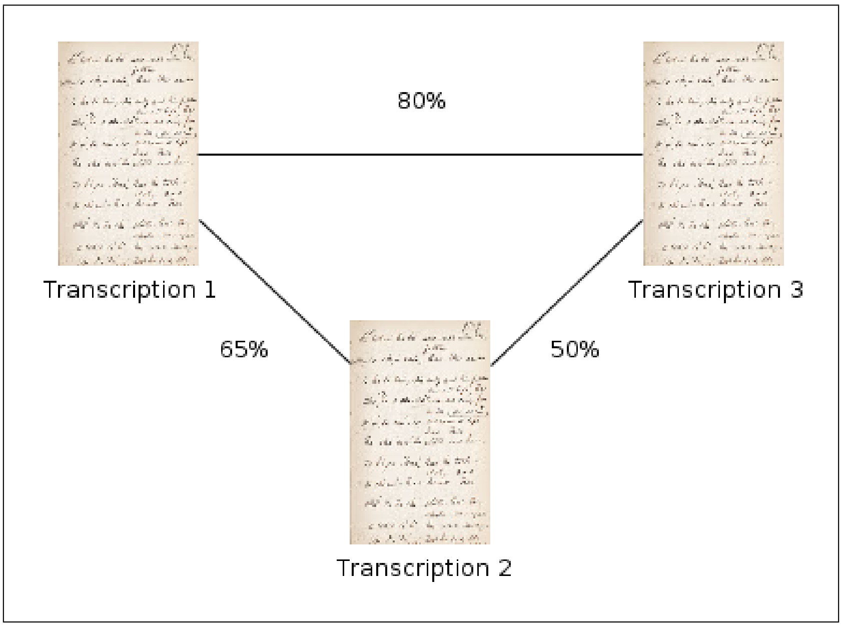


Figure 4.10: Measurement of Transcription Similarity

For each of the experiments performed in this section, the minimum, average and maximum transcriber-agreement was calculated. A custom normalized Levenshtein metric was used for calculations of the inter-transcriber agreement. The custom algorithm took into account the characters and diacritics. Characters that contained diacritics were recognized as a single character. Two experiments were performed to evaluate the reliability of transcriptions and agreement of transcribers.

- Experiment 1: Inter-transcriber agreement for English text.
- Experiment 2: Inter-transcriber agreement for |Xam text.

Experiment 1

In this experiment the level of inter-transcriber agreement for English transcriptions is calculated.

Observations

Figure 4.11 is a plot of the minimum, average and maximum inter-transcriber agreement values for up to three transcriptions. The green, red and blue data points represent the minimum, average and maximum values respectively. The data has been sorted on the increasing inter-transcriber agreement average to clearly show clusters. Table 4.1 represents the distribution of data points for English transcriptions. The data points in Figure 4.11 and Table 4.1 are split into equal proportions of the total number of points; in this case quartiles.

Analysis is done for each quartile range, the dotted-dashed black lines drawn in Figure 4.11 to show split in data.

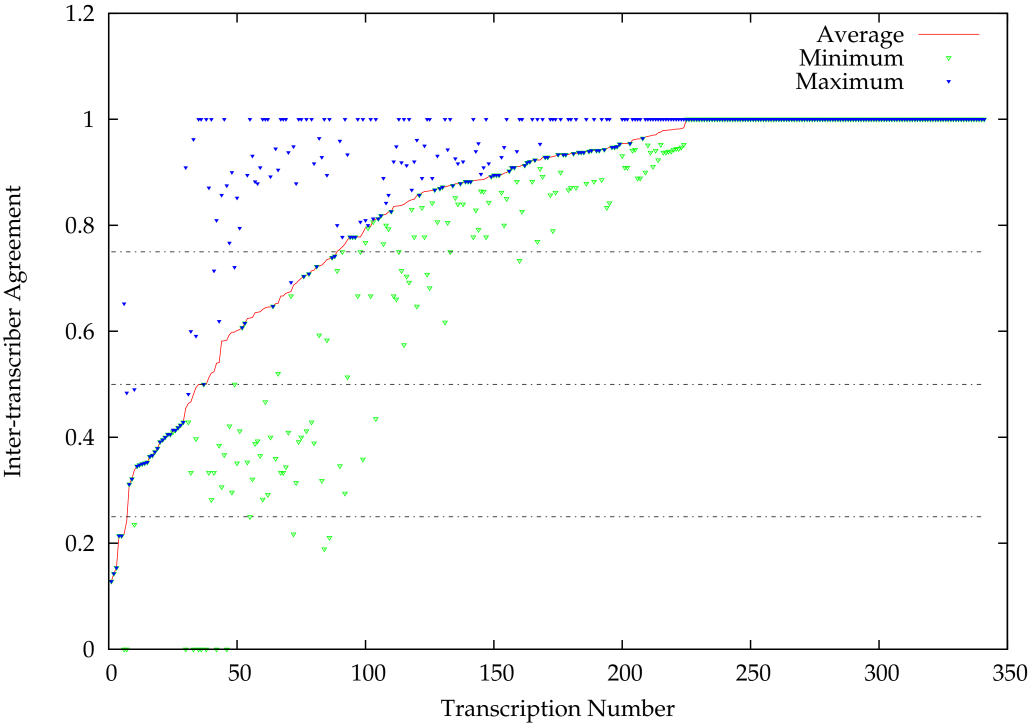


Figure 4.11: Level of inter-transcriber agreement for English transcriptions

Table 4.1: English Text Data Distribution

$Transcriber Agreement$	$Data Points$	$Percentage$
0.75 - 1.00	253	74.19%
0.50 - 0.75	50	14.66%
0.25 - 0.50	31	9.09%
0.00 - 0.25	7	2.05%

Discussion

A total of 371 transcriptions were plotted in Figure 4.11. Table 4.1 shows the distribution and percentage of data points within a particular inter-transcriber agreement range. The majority of transcriptions (74%) submitted by volunteers have an inter-transcriber agreement higher than 0.75. 15% of the transcriptions have an inter-transcriber agreement between 0.50 and 0.75, and 11% of the transcriptions have an inter-transcriber agreement below 0.50.

Analysis of Figure 4.11 shows that from transcription number 218 to 371 the transcriptions have a average agreement of $\mu = 1.00$, variance of $\sigma^2 = 0.000$ and standard deviation of $\sigma_x = 0.000$. This shows that 41.23% of the total transcriptions are similar, and reveals strong agreement amongst the transcribers. For transcription number 100 to 217, the average agreement is $\mu = 0.92$, variance is $\sigma^2 = 0.0014$ and standard deviation is $\sigma_x = 0.037$. This shows that 31.54% of the transcriptions on average have a agreement of 92%, and the agreement rate amongst volunteers was high. The spread of points from the average is small and evenly distributed, showing consistency amongst the volunteers for that portion of transcriptions.

For transcription number 1 to 99, the average agreement is $\mu = 0.89$, variance is $\sigma^2 = 4.5828$ and standard deviation is $\sigma_x = 2.1401$. 26.68% of the data points lies within this region. As can be observed from Figure 4.11, the spread of points within this region is wide, also revealed by the variance of 4.5828. Most of the transcriptions in this region reveal that there was low agreement amongst transcribers regarding the representations of the text. This observation can be attributed to volunteers submitting incomplete transcriptions or the fact that specific texts were generally difficult to transcribe, possibly due to noise within the text. Though some of the transcriptions had a low inter-transcriber agreement (26.68%), the maximum values suggest that some lines of text were reliably transcribed. The results show that volunteers are able to produce English transcriptions that are reliable with an overall inter-transcriber agreement of $\mu = 0.95$ for all the transcriptions.

Experiment 2

In this experiment the level of inter-transcriber agreement for |Xam transcriptions is calculated.

Observations

Figure 4.12 is a plot of the minimum, average and maximum inter-transcriber agreement values for up to three transcriptions. The blue, red and green data points represent the minimum, average and maximum values respectively. The data has been sorted on the increasing inter-transcriber agreement average to clearly show clusters. Table 4.2 represents the distribution of data points for |Xam transcriptions. The data points in Figure 4.12 and Table 4.2 are split into equal proportions of the total number of points; in this case quartiles. Analysis is done for each quartile range, the dotted-dashed black lines drawn in Figure 4.12 to show split in data.

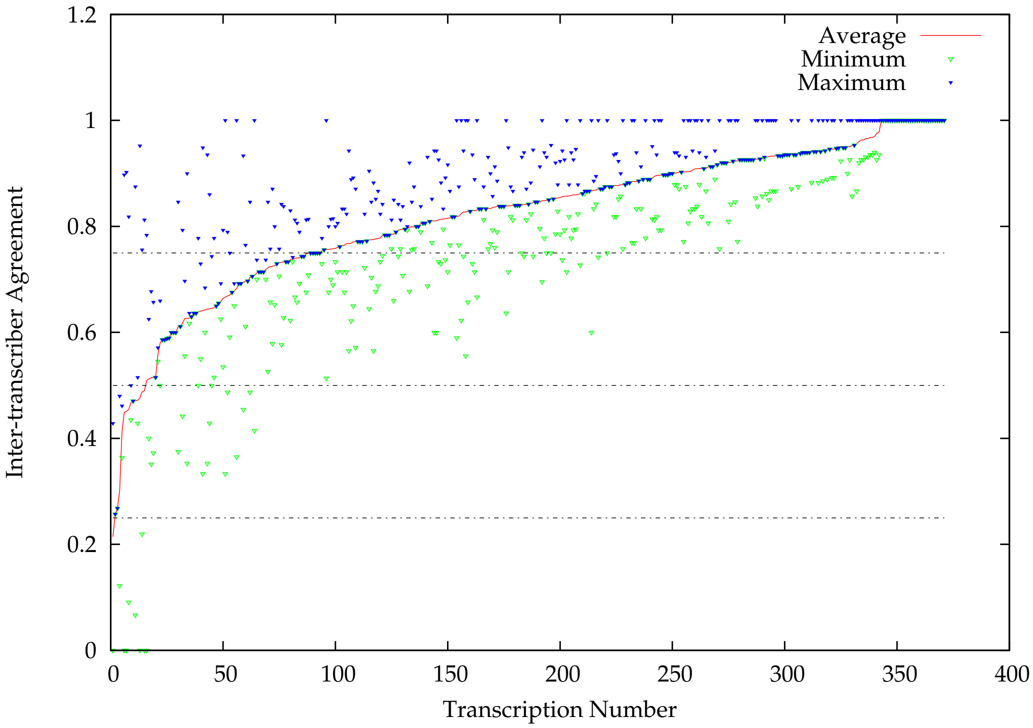


Figure 4.12: Level of inter-transcriber agreement for |Xam transcriptions

Table 4.2: |Xam Text Data Distribution

TranscriberAgreement	*DataPoints*	*Percentage*
0.75 - 1.00	283	76.28%
0.50 - 0.75	73	19.68%
0.25 - 0.50	14	3.77%
0.00 - 0.25	1	0.27%

Discussion

A total of 412 transcriptions were plotted in Figure 4.12. Table 4.2 shows the distribution and percentage of the data points within a particular inter-transcriber agreement range. The majority of transcriptions (76%) submitted by volunteers have an inter-transcriber agreement higher than 0.75. 20% of the transcriptions inter-transcriber agreement rate is between 0.50 and 0.75, and 4% of the transcriptions inter-transcriber agreement rate is below 0.50.

Analysis of Figure 4.12 shows that transcription number 375 to 412 have a average agreement of $\mu = 1.00$, variance of $\sigma^2 = 0.000$ and standard deviation of $\sigma_x = 0.000$. This shows that 8.98% of the total transcriptions are similar and there is a strong agreement rate amongst the volunteers on the text representation. For transcription number 200 to 374, 42.23% of the transcriptions have an average transcription agreement of $\mu = 0.89$, variance of $\sigma^2 = 0.0016$ and the standard deviation is $\sigma_x = 0.04$. More than half of the data points lie within this region and the former, evidence that most of the transcriptions have an inter-transcriber agreement value of 0.89 or better. The data points are evenly distributed across the average line, which can be confirmed by the low variation value of 0.0016. This shows a high agreement rate amongst the volunteers but a decrease in transcription similarity compared with the former region analysed.

For transcription number 100 to 199, the average agreement is $\mu = 0.79$, the variance is $\sigma^2 = 0.0004$ and standard deviation is $\sigma_x = 0.2000$. This region contains 24.03% of the total transcriptions. The maximum and minimum values are evenly distributed from the average line, with a variation of 0.0004. For transcription number 1 to 99, the average agreement is $\mu = 0.59$, variance is $\sigma^2 = 0.0269$ and standard deviation is $\sigma_x = 0.1640$. 24.03% of the transcriptions lie within this region. The data points are evenly distributed across the average line with a few outliers being observed in the bottom left of Figure 4.12. The outlier points can be attributed to poorly transcribed text lines; the minimum value for this line would be lower than other data points.

A number of data points lie on the average line. This could represent lines that were transcribed by only one transcriber due to limited participants or similar transcriptions. The data points that lie either above or below the average line were transcribed by at least two transcribers. The average agreement for the 412 transcriptions is $\mu = 0.80$, the variance is $\sigma^2 = 0.0249$ and standard deviation is $\sigma_x = 0.1579$. Based on the results from this experiment, it is plausible to state that the |Xam transcriptions collected from volunteers are reliable.

Summary

An analysis of Figure 4.11 and 4.12 shows that the English text has higher average agreement than the |Xam text. The average inter-transcriber agreement of both are $\mu = 0.95$ and $\mu = 0.80$ respectively. The results suggest that on average the volunteers transcribed the English text better than the |Xam text. This finding is expected as the |Xam text contains complex characters and diacritics within the script hence would pose a challenge to transcribe. Though issues were raised (see section 4.3) with regard to poor legibility of the English text by transcribers; this did not greatly affect the final results. The figures in Table 4.1 and Table 4.2 are plots of the distribution of submitted transcriptions. Though the figures in Table 4.1 for english are less than Table 4.2 the average inter-transcriber for english results is still higher than for |Xam text.

4.6 Volunteer Accuracy

This section aims to assess how volunteers in this project perform compared with previous efforts at recognizing |Xam text.

Methodology

The gold standard data used for evaluating the accuracy of volunteers at the task of transcription was obtained from the corpus created by [Williams and Suleman, 2011a]. For these experiments, two corpora were used for experimentation. Corpus 1 shall be termed Corpus-V to denote transcriptions produced by volunteers for this research project. Corpus 2 shall be termed Corpus-G to represent the gold standard data from Williams' research. Corpus-G is a dataset containing transcriptions produced from the Bleek and Lloyd notebooks. Although Corpus-G is termed *Gold Standard*, this corpus is known to have inaccuracies, but is the best approximate representation available. Two experiments were conducted in this section:

- Experiment 1: Evaluation of accuracy of the individual volunteer transcriptions with the gold standard data.
- Experiment 2: Evaluation of volunteer accuracy versus inter-transcriber agreement.

Experiment 1

This experiment aims to: (1) measure the accuracy of the individual transcriptions produced by volunteers with the gold standard data; (2) assess how this compares with previous efforts.

Observations

Figure 4.13 is a plot of the minimum, average and maximum inter-transcriber agreement values for up to three transcriptions. The green, red and blue data points represent the minimum, average and maximum values respectively. The data has been sorted on the increasing inter-transcriber agreement average to clearly show clusters. Table 4.3 represents the distribution of data points for |Xam transcriptions. The data points in Figure 4.13 and Table 4.3 are split into equal proportions of the total number of points; in this case quartiles. Analysis is done for each quartile range, the dotted-dashed black lines drawn in Figure 4.13 to show split in data.

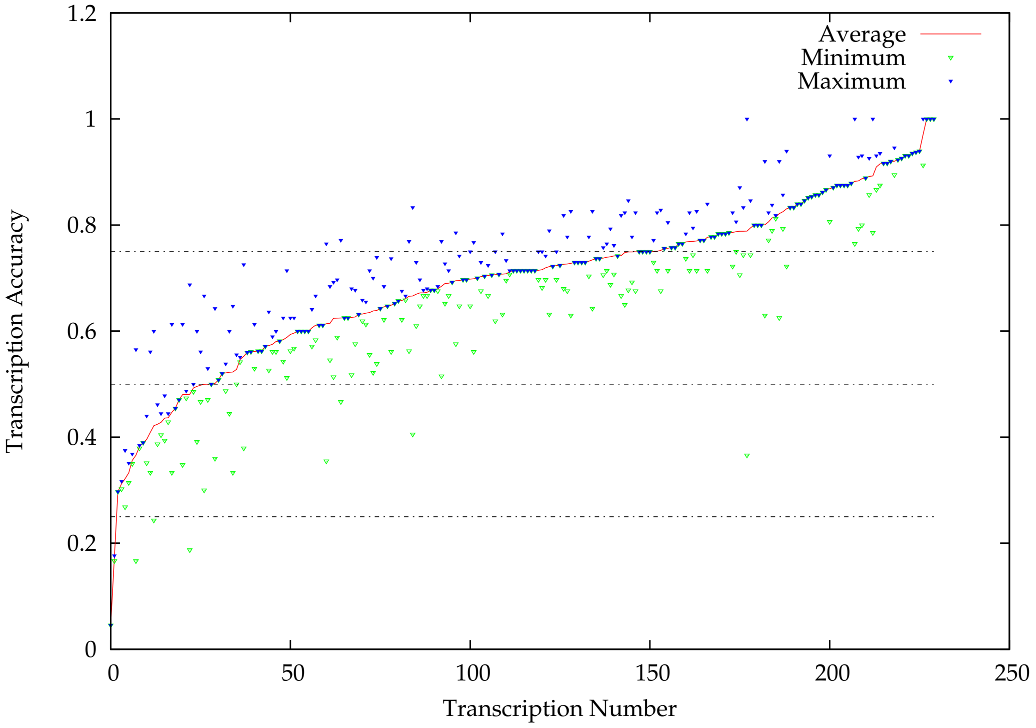


Figure 4.13: Accuracy of volunteers

Table 4.3: Accuracy Distribution for Corpus-V with Corpus-G

Accuracy	*DataPoints*	*Percentage*
0.75 - 1.00	67	32.84%
0.50 - 0.75	117	57.35%
0.25 - 0.50	19	9.31%
0.00 - 0.25	1	0.49%

Discussion

A total of 204 transcriptions were plotted in Figure 4.13. The majority of the transcriptions have an increasing match rate with the gold standard data. The trend in the average line is fairly linear, increasing positively. The variation of data points from the average line is generally consistent and small, showing consistency in the transcriptions.

Figure 4.13 shows that transcriptions 154 to 204 have an average agreement of $\mu = 0.8621$, variance of $\sigma^2 = 0.0029$ and standard deviation of $\sigma_x = 0.0542$. 32.84% of the transcriptions from Corpus-V have an 86% match to those of Corpus-G. Only one transcription has a perfect match. For transcriptions

103 to 153, the average transcription accuracy is μ =0.7372, variance is $\sigma^2 = 0.0006$ and the standard deviation is $\sigma_x = 0.0248$. This region contains 57.35% of the total transcriptions with a match rate of 74%. Transcriptions 52 to 102, the average accuracy is $\mu = 0.6588$, the variance is $\sigma^2 = 0.0007$ and standard deviation is $\sigma_x = 0.0262$. This region contains 9.31% of the total transcriptions with a match rate of 66%. Transcriptions 1 to 51, the average accuracy is $\mu = 0.4832$, the variance is $\sigma^2 = 0.0124$ and standard deviation is $\sigma_x = 0.1115$. This region contains 0.49% of the total transcriptions with a match rate of 48%.

The overall average accuracy of volunteer transcriptions compared with the gold standard is 69.69%. In light of the inconsistencies known to exist between both corpora, the accuracy achieved is good. It was observed that some volunteers represented their transcriptions in a different way to that of the gold standard, but recognizing similar text. The diacritics can be stacked around characters in varying order but still have the same meaning. If the inconsistencies between the representations of the diacritics were fully-resolvable, then a higher accuracy is expected.

The research by [Williams and Suleman, 2011b] that used machine learning techniques to recognize the |Xam text achieved an accuracy 45.10% at line level compared with the same gold standard. For this research, when the volunteer transcriptions were compared with the gold standard, an accuracy of 69.69% was achieved at line level. As mentioned earlier, the accuracy value is expected to increase if the inconsistencies between the two corpora are resolved. This result shows that volunteer thinking outperforms machine learning techniques at recognizing complex |Xam texts.

Experiment 2

The aim of this experiment is to establish whether a correlation exists between the level of inter-transcriber agreement and the accuracy for |Xam transcriptions.

Observations

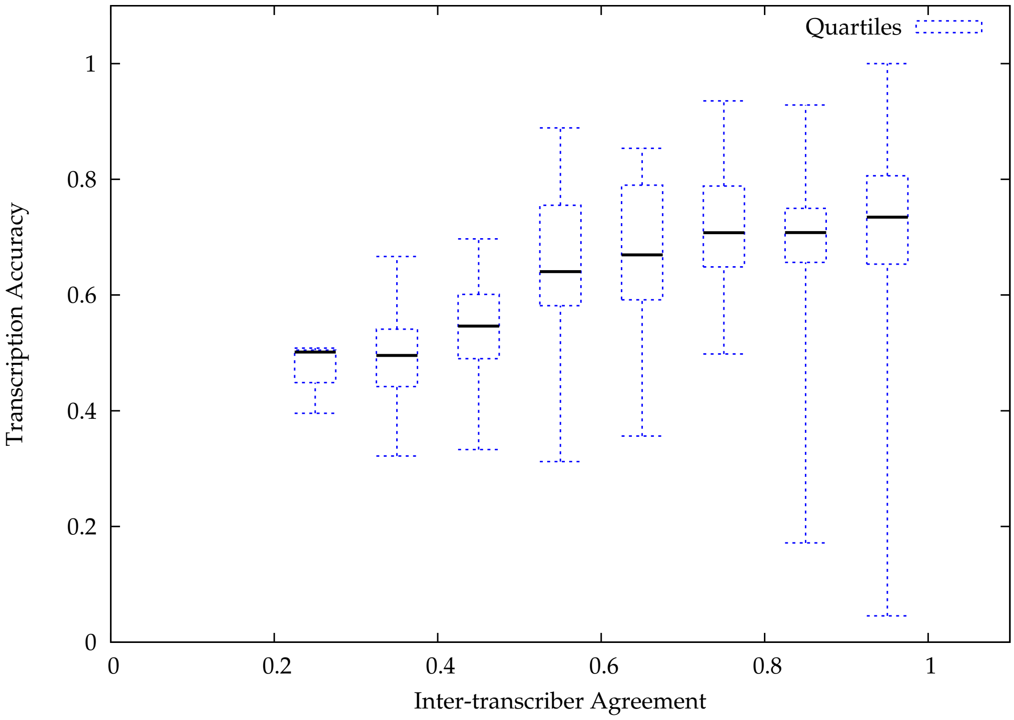


Figure 4.14: Correlation of Inter-transcriber Agreement and Volunteer Accuracy

Table 4.4: Distribution of Correlation data

Range	*Minimum*	*1stQuartile*	*Median*	*3rdQuartile*	*Maximum*
0.20 - 0.30	0.3957	0.4486	0.5014	0.5048	0.5082
0.30 - 0.40	0.3216	0.4419	0.4957	0.5409	0.6667
0.40 - 0.50	0.3328	0.4900	0.5463	0.6011	0.6969
0.50 - 0.60	0.3122	0.5817	0.6403	0.7553	0.8889
0.60 - 0.70	0.3561	0.5920	0.6693	0.7900	0.8537
0.70 - 0.80	0.4981	0.6486	0.7077	0.7885	0.9355
0.80 - 0.90	0.1716	0.6563	0.7079	0.7499	0.9286
0.90 - 1.00	0.0455	0.6534	0.7346	0.8063	1.0000

Discussion

Table 4.4 tabulates the data plotted for the box and whisker plots in Figure 4.14. The tails of the box and whisker plots reveal the minimum and maximum values of accuracy within a certain bin range. The data points have been sorted into bins and the average accuracy of points in a particular bin is calculated, equivalent to the median values in the table above.

If a line was to be plotted through the median points in Figure 4.14 this would reveal a positive linear increasing relationship between the inter-transcriber agreement and volunteer accuracy, showing a correlation between the two. Generally the graph shows that as the inter-transcriber agreement increases so does the transcription accuracy.

The box and whisker plot with 0.95 inter-transcriber agreement represents single transcriptions or two or three transcriptions submitted by volunteers that are similar. The plot has long tails because the accuracy of these transcriptions lies between 0 and 1; they can be totally accurate or a mismatch. The average accuracy of the transcriptions lie between 50% and 73% for all the bins. An implication of this finding is that only the inter-transcriber agreement values are needed to estimate the general accuracy of transcriptions in the case that no gold standard data exists for comparison. If the inter-transcriber agreement is 0.55 it is expected that the average transcription accuracy would be above 60%. The next section discusses the survey conducted for this research.

4.7 User Survey

The first aim of the user survey was to better understand the user experience of volunteers using the transcription tool. Secondly, the survey was intended to obtain feedback to pick out important user issues. Previous user studies [Kittur et al., 2008] have shown that early user input can solicit important feedback, and potentially improve the user interaction with a system.

4.7.1 Design and Tools

The online study methodology was used for the design of this survey, as this project is Web-based. Lime survey[1] - an online Web survey tool - was used to design the survey questions. The software is freely available for download online under the open source license. See appendix A.2 for the survey questionnaire. Reference was made to the All About User Experience[2] website and some of the survey questions were adapted from [Lund, 2001] USE questionnaire. USE stands for usefulness, satisfaction and ease of use. User experience is evaluated based on these three factors.

The five-point Likert rating scale was [von Ahn et al., 2008] used for the first set of questions:

[1] http://www.limesurvey.org/

[2] www.walkabout.org

1. Strongly Agree
2. Agree
3. Neutral
4. Disagree
5. Strongly Disagree

An initial draft of the questionnaire was designed; this was then evaluated by an academic expert. Corrections to the first draft were made, and the final result later used on Limesurvey. Two sets of questions were posed to users. In the first set, users were required to state how strongly they agreed or disagreed with a list of statements. The second set of questions were open-ended; users optionally gave general comments regarding the transcription tool. The survey was hosted on a server in the Computer Science department at the University of Cape Town. The link to the survey was then embedded on a user's account page. In the video tutorial, users are asked to fill in the online survey.

4.7.2 Questionnaire Results

For analysis purposes, the categories, *Strongly Agree & Agree* and *Strongly Disagree & Disagree* were combined, hence three categories were used including the *Neutral* response. The responses to questions are calculated as percentages. The labels for each question have been shortened. The full questions can be viewed in Appendix A.2. Next the observations are given and discussed in the context of this research, and lastly potential implications for crowdsourced transcription projects are mentioned.

Participant Background Information

A total of fourteen volunteers responded to the online survey, and of these two incomplete responses were discarded. Two participants have contributed to other crowdsourcing projects, namely GalaxyZoo[3] and Distributed Proof-Readers[4].

Evaluation of transcription tool usefulness

Four questions were posed to the volunteers to determine their perception of the usefulness of such a tool for the purpose of transcription.

[3] www.galaxyzoo.org

[4] www.pgdp.net/

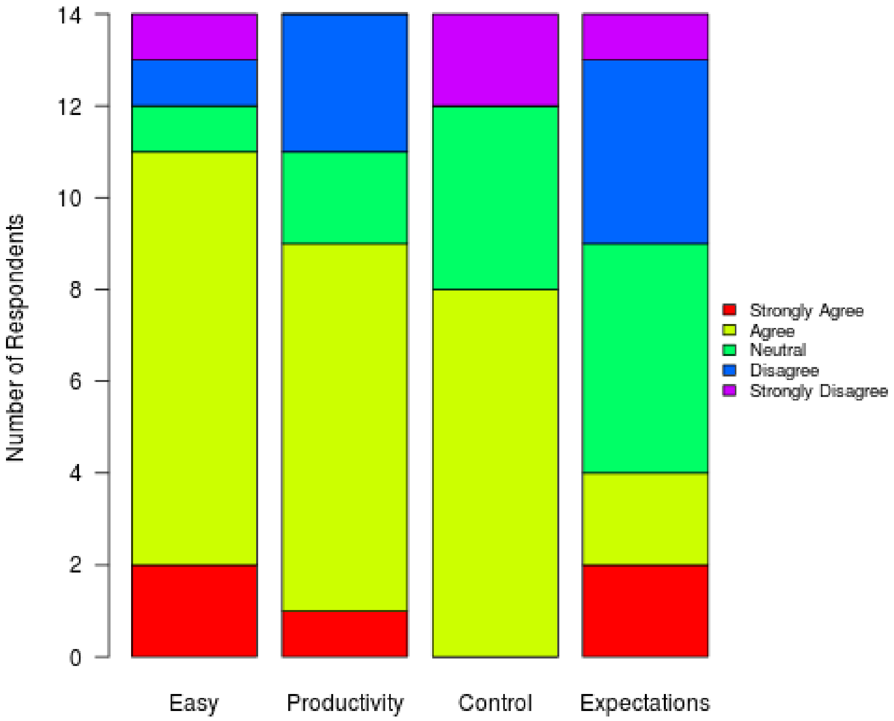


Figure 4.15: Usefulness of Transcription Tool

Figure 4.15 shows that 79% of the participants found the learning curve of the transcription tool easy, while 21% of the participants were unsure or thought it was difficult to use. In response to the question of being productive using the tool, 64% of participants said they became very productive after a few times using the tool. 14% of the participants were neutral in their response, and the other 21% disagreed to becoming productive after a few attempts. Inference can be made from the question on the ease of learning the tool and volunteer productivity that overall the tool was easy to learn and volunteers became efficient with the tool after a few practices; 21% of the participants who were either unsure or disagreed that the tool was easy to use could likely account for the 21% that disagreed to being productive.

57% of participants felt they were in control while using the tool, 29% were unsure with the remaining 14% disagreed. For the question regarding volunteer expectations of the tool, 14% of participants thought it worked in a manner they expected; this could be possibly be accounted for by the two participants who had previously contributed to other crowdsourcing projects. 36% were neutral and another 36% disagreed. Such a large percentage of uncertainty and disagreement might be associated with unfamiliarity with online transcription projects.

In summary, approximately 67% of the first three questions regarding the tool usefulness were answered in the affirmative, and 16% of the responses were negative. Responses from questions on expectations of the volunteers could hint that further improvement of the tool might be needed, but the majority found the tool useful and it could potentially be adapted for other purposes.

Evaluation of transcription tool satisfaction

Eight questions were posed to better understand volunteers user experience with regard to satisfaction using the tool. The majority of questions in this section have been posed in a positive and negative manner to determine if there is consistency in the responses obtained. Figure 4.16 and 4.17 have been used to fully capture the responses for this section.

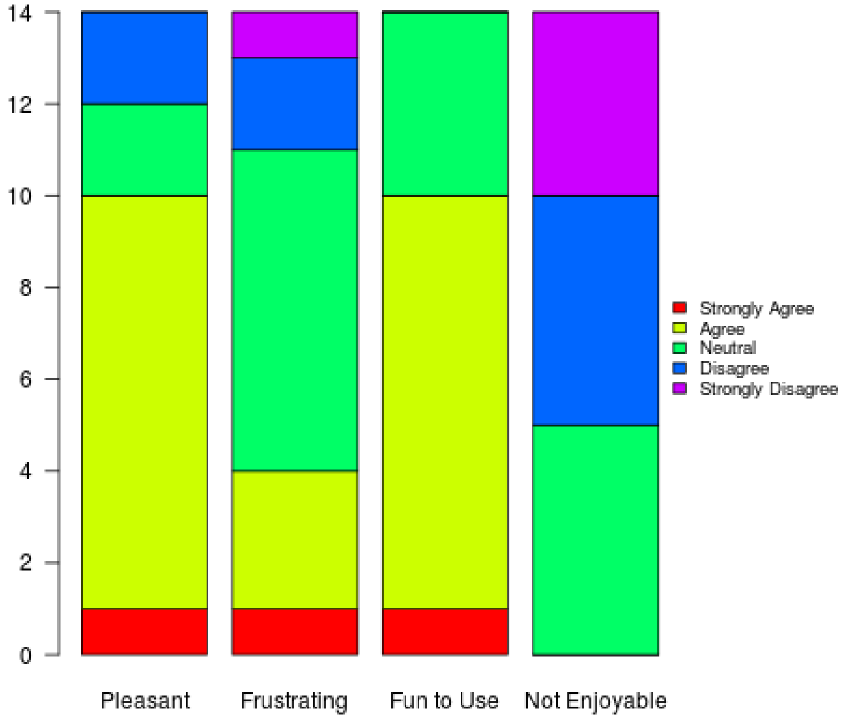


Figure 4.16: User Satisfaction using Transcription Tool

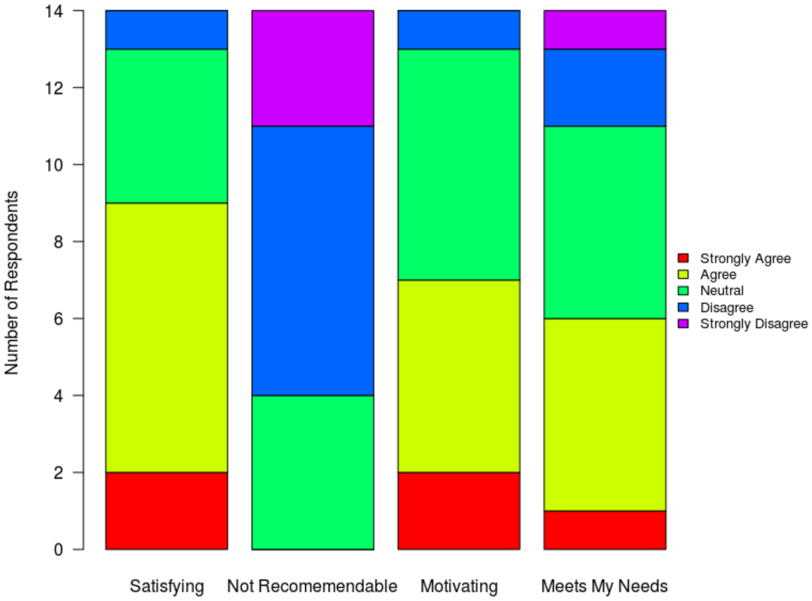


Figure 4.17: User Satisfaction using Transcription Tool

Figures 4.16 and 4.17 are plots of participant responses to their user experience with the transcription tool, and these are discussed in their respective orders. Figure 4.16, shows that 71% of respondents found the tool pleasant to use, while 29% were either unsure or disagreed. 29% participants felt frustrated using the tool, 50% were unsure and 21% disagreed. 71% of participants found the tool fun to use, with 29% unsure. Since only 21% felt frustrated using the tool compared with the 71% who found the tool pleasant, it would be reasonable to conclude that the majority found the tool pleasant to use. 71% of participants found it fun to use the tool, while 29% were neutral. When asked whether participants felt the tool was not enjoyable to use, 36% were neutral and 64% disagreed with this statement. Analysis of questions on whether the tool was fun to use or it was not enjoyable confirm that most of the participants had fun using tool.

Figure 4.17, shows that 64% of participants felt a sense of satisfaction using the tool, while 29% were neutral and the remaining 7% disagreed. 71% of respondents said that they would recommend the tool to friends, with the other 29% being neutral. Responses from the question on recommending the transcription tool confirm the validity of responses about user satisfaction with the tool. A majority of the participants felt satisfied using the tool and would tell others about it. From some of the replies received on the invitation (see section 3.7), a few people said they would tell their colleges about the project. 50% were motivated to continue using the tool, while 43% were neutral and 7% disagreed. The 50% observation could account for the high number of contributions made by particular volunteers to the project (see Figure 4.1).

43% of respondents found the tool adequate for the task at hand, with 36% being unsure and 21% disagreeing. Of note, most participants who contributed to the project had backgrounds in linguistics, and this could account for the 21% of participants who answered in the negative regarding the tool being sufficient for the transcription task. In conclusion, most participants answered with mostly positive responses regarding recommending the tool, feeling motivated and it being sufficient to user needs. Based on this it can be concluded that the tool was satisfying to use.

Evaluation of transcription tool ease of use

Six questions were posed to the volunteers to determine the ease of use of the transcription tool, as it meant for non-experts.

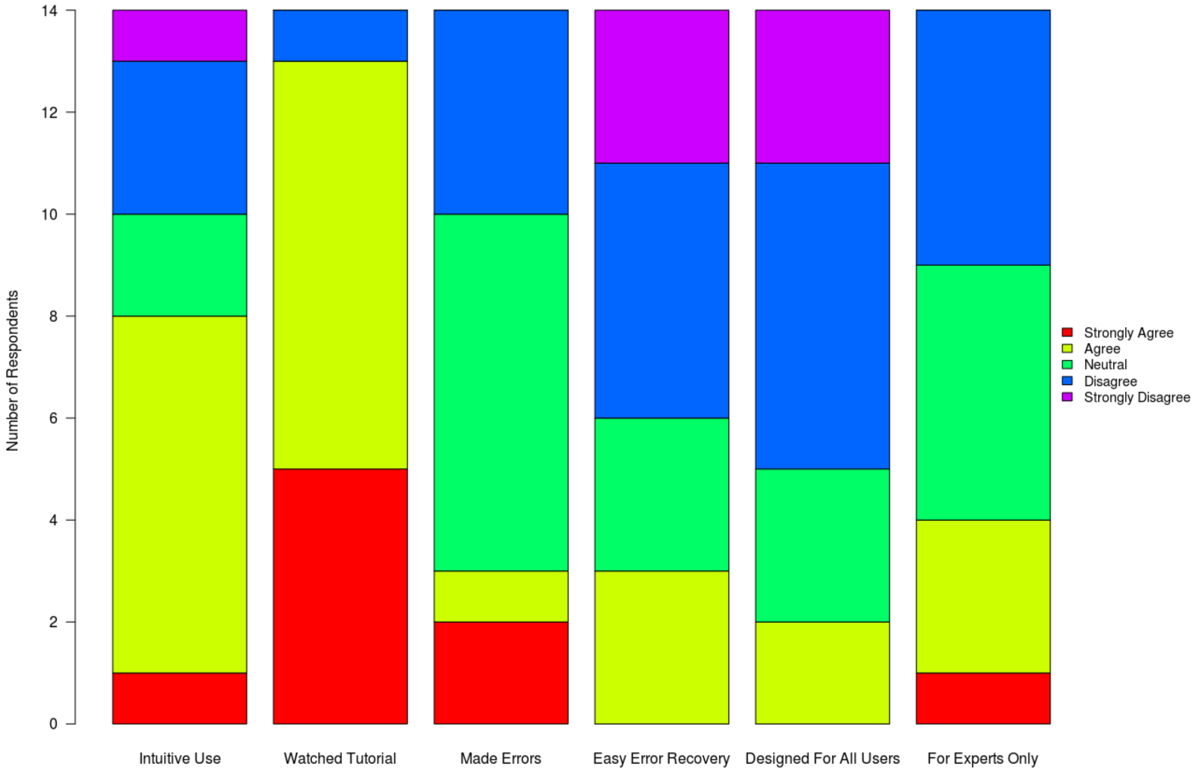


Figure 4.18: Transcription Tool Ease of Use

Figure 4.18 shows that 57% of respondents found the transcription tool intuitive to use. 14% were neutral, while 29% disagreed. 93% of participants acknowledged that it was essential to watch the transcription tutorial first before they could use the tool. 21% of participants confirmed that they made many mistakes using the tool, 50% were neutral and 29% disagreed. 21% of respondents found it easy to recover to correct their mistakes, 21% were neutral and 57% disagreed. Such a high percentage of users struggling to correct their errors can be attributed to the non-standard encoding of the complex |Xam text. As mentioned earlier, this language is not supported in standard Unicode, hence it would be the first time for participants to use it.

14% of participants felt the tool was designed for use by all users, 21% were unsure and 64% disagreed. 29% of participants thought that the tool could only be used by experts, while 71% were either unsure or disagreed. Responses to questions on who the tool was designed for show that though the tool at first glance seems difficult to use for non-experts, with sufficient training all volunteers are capable of using the tool.

Positive and Negative Aspects of Transcription Tool

In this section volunteers were asked to state positive and negative aspects about the transcription tool, based on their experience. The positive aspects shall be discussed first, followed by negative aspects. The detailed answers from the volunteers are given in Appendices A.3.1 and A.3.2 respectively.

Generally, a number of users found the transcription tool easy to understand and use after a few practices. Some volunteers report that they enjoyed the process. One volunteer noted the difficulty in transcribing such scripts and found the Characters & Diacritics palette sufficient in meeting this need. The same user found the tool to be well thought out to meet the challenges, and suggests that the tool can still be improved. Deciphering some of the characters proved to be a challenge to some volunteers, but one volunteer acknowledges the benefit of the zooming in and out feature to easily view the image. Lastly, one user found the application interesting, and helped them in understanding some of the languages spoken by hunter-gatherer people.

Though a number of positive aspects were stated, negative aspects about the tool were also raised. A number of volunteers found that not all the diacritics were included in the Characters & Diacritics palette, and some were weary of submitting incorrect transcriptions. As mentioned earlier, more diacritics are still be found, and the ones used in the tool is a best approximate set of the popular diacritics found in the script. Concerns were also raised with regard to the arrangement of features on the display area. For instance, users found the the text preview obstructed their view of the text input area. This can be attributed to design challenges experienced, where a compromise was made between the available text input areas, and display area for the image. The image display area was made bigger to make it easy to view the transcription image, while the text input boxes were made small. So on small display screens this would pose a challenge. One volunteer suggested that there should be a way of marking places in the manuscript to indicate crossed out words or possibly transcribing all the text to get a complete image of the original manuscript.

Overall, the volunteers raised relevant issues, some of which were easily addressable and others to be considered as part of future work, as more time is required to handle these. A number of negative aspects mentioned can be attributed to design limitations of the transcription interface. The interface itself was designed to only capture text on the right side of the transcription image.

General Comments

The last section of the user survey required participants to give any general comments they had regarding the transcription tool. Based on the general comments given, the volunteers had a good experience and found the tool useful and thought it to be a good way to transcribe this text. Further suggestions were given on how to improve the functionality of the transcription interface to suit volunteers with small screens. The comments given in this section further confirm both positive and negative aspects mentioned earlier, e.g. how to deal with crossed out text and marking words that are difficult to decipher. One user thought this to be a great idea, and suggested that having an automatic transcription tool would be helpful. One user commented that they expected the English text to have already been transcribed. This highlighted that they did not fully understand the goal of the project, which was to transcribe both the |Xam and English text. A volunteer states that they were confused on whether a particular character was a ”u“ or ”w“. This was observed during corpus pre-processing mentioned in section 4.4.1, where some volunteers were consistently mistaking particular characters. One volunteer requested to view links to additional descriptive material about the |Xam language.

4.8 Summary

This chapter begins by stating the research questions of the project, then describes the data collection phase, followed by a discussion highlighting some of the challenges experienced by volunteer during the data collection phase. Some of these issues were resolved in the corpus analysis section. Four experiments are then described under the transcription similarity and volunteer accuracy sections. These experiments are aimed at answering the research questions posed. Lastly, the discussion analyses results from the survey on the user experience using the transcription tool.

Findings

The need for post-editing shows that within the context of this project, quality control is necessary to ensure the transcriptions are consistent before further inference can be made. This finding is attributed to the difficulty in the handwriting, complexity of the text and noise within some of the pages. It is important to highlight that it took over 48 hours for a single graduate student to complete some of the post-editing tasks with the transcripts obtained in this project. Similar efforts are noted in other projects [Causer et al., 2012].

Research question 1 in section 4.1 can be answered in the affirmative after evaluation of the findings. Volunteer thinking can be used to crowdsource intellectually-intensive tasks in digital libraries (like transcribing handwritten manuscripts).The high transcription similarities for both |Xam and English text of 80% and 95% respectively suggest that transcriptions from volunteers are of good quality. This was proven true in section 4.6 by the positive linear correlation between the inter-transcriber agreement and transcription obtained.

After analysing all responses based on the usefulness, user satisfaction and ease of use of the tool, it is possible to conclude that most users had a good experience using the tool. Emotions of satisfaction, fun and enjoyment were experienced, but also frustration. Based on the findings, it is possible to say that once the negative issues raised by transcribers have been dealt with, this is a tool a number of volunteers would find fun to use. It is very interesting that volunteers would find the task of transcription fun, as this is often described in literature [Clocksin, 2003] as being laborious. A number of volunteers would personally send emails enquiring how long the transcription project would still be available, showing enthusiasm in contributing to the research.

Research question 2 aimed to assess how volunteer thinking compares to machine learning techniques when applied to the problem of transcription. The research by [Williams and Suleman, 2011b] that used machine learning techniques to recognize the gold standard achieved an accuracy 45.10% at line level. For this research, when the volunteer transcriptions were compared with the gold standard, an accuracy of 69.69% was achieved at line level. The volunteer thinking approach outperforms machine learning techniques at recognizing the |Xam text. Using volunteers to transcribe the Bleek and Lloyd collection yields more accurate results than previous methods. This approach can possibly be used to transcribe other historical handwritten manuscripts. In light of the inconsistencies known to exist between both corpora, the accuracy achieved is good. If the inconsistencies between the representations of the diacritics were fully-resolvable, then a higher accuracy is expected. Based on the findings research question 2 has been successfully answered. Another potentially interesting experiment would be to determine how the volunteer accuracy varies when poorly performing transcriptions are discarded.

Chapter 5

Conclusion

The digital Bleek and Lloyd notebooks form part of a rare collection detailing the history and culture of the early inhabitants of Southern Africa, the hunter-gather people. Over the past few years, research has been conducted to make this easily accessible online. Though the notebooks are available online, the stories described in the books are only available as images. This research set out to explore whether volunteer thinking can be used to expose accurate transcriptions of these texts.

A transcription tool was successfully created for the transcription of the Bleek and Lloyd Collection. Two deployment phases were employed in announcing the project to volunteers. The first phase was aimed at announcing the project to local African researchers participating in similar work. The first phase was also used as a testing period for any possible bugs. Within this period, useful feedback was obtained, from our colleagues involved with Citizen Cyberscience. In terms of participation from local African researchers, the response was generally poor. This can be attributed to a couple of reasons:

- There are a few communities in Africa involved with work related to the earliest inhabitants of Southern Africa and their visibility online is low.
- Not many researchers in Africa are involved in crowdsourcing research.

Ben Brumfield, a recognized blogger on crowdsourced transcription projects, also noted that this was the first crowdsourcing transcription project in the Southern hemisphere. From a perspective of promoting future research within this area, there has to be a raised awareness amongst researchers in Africa of potential benefits they could reap through crowdsourcing their work. In the second phase of the deployment, the project was announced internationally to any interested participants. Most of the contributors to this project originated from North-America, Europe, New Zealand and Australia.

As with many other crowdsourcing projects, there is need to enforce control measures to block spam. For this project one user submitted spam via the transcription tool. Apart from that incident, the whole data collection process went well. Due to the complexity of the problem, it was noted that a few volunteers would personally email to clarify whether their transcriptions were correct before they submitted the results. Though the video tutorial used to train volunteers was adequate to get volunteers transcribing, more assessment exercises would be beneficial to improve the learning rate of volunteers at the task. The transcription of the Bleek and Lloyd notebooks proved to be difficult.

The project set out to answer two research questions; four experiments were conducted to answer the research questions. Research question 1 aimed to investigate: *If volunteer thinking can be used to crowdsource intellectually-intensive tasks in digital libraries (like transcribing handwritten manuscripts).* The results from the experiment on calculating the similarity of the transcriptions produced by volunteers showed that volunteers had a high inter-transcriber agreement. The inter-transcriber agreement for the |Xam and English texts was 80% and 95% respectively. This means that the transcriptions produced by volunteers were very similar.

An experiment was conducted to determine if there exists a correlation between the accuracy of transcriptions submitted by volunteers and their corresponding inter-transcriber agreement. The results detailed in section 4.6 showed that a positive increasing linear relationship exists between the inter-transcriber agreement and transcription accuracy. It shows that a high inter-transcriber agreement would lead to more accurate transcriptions. Based on these three experiments it can be concluded that volunteer thinking can be applied to complex tasks like transcribing complex handwritten manuscripts. The implication of this is that it supports the idea of shifting away from using trained lab experts, for intellectually intensive tasks. If researchers were to adopt this, strict control measures need to be implemented first. Another implication of the finding is that it has been shown that scholarly data can be transcribed by volunteers at no cost to the project, as no payments were made in this research.

Research question 2 aimed to assess *how volunteer thinking compares to machine learning techniques when applied to the problem of transcription.* One notable difference between the two methods is that volunteer thinking does not require a training dataset to recognize text. [Callison-Burch, 2009] states that the performance of machine translation depends on the size of the training data. This is also applicable to transcription of the |Xam language using machine learning techniques. The accuracy of machine learning techniques in recognizing the complex scripts depends on the accuracy of the training data. [Williams and Suleman, 2011b] states that diacritics in the |Xam language affect the recognition accuracy of the texts. Volunteers in this project found it challenging to decipher some of the diacritics, and this is what made the transcription process tedious. The outcomes of applying crowdsourcing to the task of transcription for heritage collections in digital libraries will result in enhancing the quality of content archived in digital library repositories, thereby improving the ability to easily expose and disseminate cultural heritage content over the World Wide Web

The research by [Williams and Suleman, 2011b] that used machine learning techniques to recognize the gold standard achieved an accuracy of 45.10% at line level. For this research, when the volunteer accuracy was compared with the gold standard an accuracy of 69.69% was achieved at line level. In light of the inconsistencies known to exist in both corpora the accuracy achieved is good. The volunteer thinking approach outperforms machine learning techniques at recognizing the |Xam text. Based on the findings research question 2 has been successfully answered.

Hopefully many other African researchers will be inspired to explore using crowdsourcing in their different domains, from the findings of this research. Africa possesses a number of unique heritage collections, most of which have not been digitized yet. The Timbuktu manuscripts mentioned earlier in section 2.4.1 would be a really interesting project to potentially transcribe using crowdsourcing. Most citizen science projects originate from overseas. Crowdsourcing presents a new unexplored avenue for future research within Africa. This project should serve as motivation to explore the potential application within the Africa context.

5.1 Future Work

Research regarding the transcription of historical manuscripts has many possible areas that can be explored as future work. This section attempts to address and highlight a few of these areas.

Statistical language modelling is widely used in the domain of Natural Language Processing. This provides important knowledge about the language,for instance the distribution of words in the language or likelihood of word occurrence. N-Gram language models have been used to solve problems related to spell checking [Gupta and Mathur, 2012], word segmentation [Yoshimura et al., 2012], text reuse [Adeel Nawab et al., 2012] and similarity [Islam and Inkpen, 2008]. N-gram language models have been shown to improve the recognition of recognisers [Williams, 2012] significantly. Future work can attempt to analyse whether N-gram language models can be applied to the |Xam language, specifically exploring (1) whether unigrams, bigrams or trigrams used to improve the accuracy of volunteer transcriptions by merging the best possible results; (2) calculate the likelihood probability that sentences of the language are likely correct; (3) determine if a correlation exists between the inter-transcriber agreement and sentence likelihood probability.

For this project, the Bossa job distribution policy implemented was set to issue three instances of a job. Once the three transcriptions had been collected the job would be considered as COMPLETED; if more than five instances are issued without a consensus then the job would be considered INCONCLUSIVE. An alternative job distribution policy that can be used would be to use submission accuracy thresholds. Based on the findings of the research that there exists a positive linear correlation between inter-transcriber agreement and volunteer transcription accuracy, this would be feasible. A submission accuracy threshold is a desired accuracy level for a transcription; that can be calculated using the inter-transcriber agreement. In this case the similarity of the transcriptions are calculated using either the normalized Levenshtein algorithm described in section 4.5 or any other suitable algorithm. Job instances are issued until the threshold level has been achieved, thereby ensuring accurate volunteer transcriptions.

[Kittur, 2010] suggests that volunteer collaboration could achieve improved results and performance amongst volunteers. Incorporating features that would allow communication amongst volunteers in the system could greatly improve the task of transcription by making it a fun and enjoyable experience as volunteers share ideas. A suggestion that was raised by volunteers in the user study was the need for a feature that would allow one to save a draft transcription and return to it later for completion. Due to limited development time, the solution implemented for this was an auto-save plugin. Ideally future work on the project could explore the use of the MediaWiki tool, that was used for the Transcribe Bentham project [Causer et al., 2012]. Another potential implementation feature is implementing real time rendering of volunteer transcriptions and possibly including English voice translations of the rendered transcription.

Lastly, some limitations of the project are mentioned to be considered when designing a production system of the tool. A limitation in the current design of the transcription interface is the ability to capture side notes that appear in the transcription pages. The side notes would be useful in understanding the context of the stories and shed light into authors thoughts. This would be a functionality potentially fascinating to volunteers in helping give more contextual knowledge. Another limitation to be addressed is comprehensively capturing all the diacritics that appear in the |Xam and !Kun languages. The encoding tool captures the most frequent diacritics appearing in the texts. As mentioned earlier, 300 characters and diacritics have been found. In the user survey one user noted that certain characters were not captured in the characters and diacritics palette (see section 3.6.2); this was then resolved.

Bibliography

Tomasz Adamek, Noel O'Connor, and Alan Smeaton. Word matching using single closed contours for indexing handwritten historical documents. *International Journal on Document Analysis and Recognition*, 9:153–165, 2007. ISSN 1433-2833. URL http://dx.doi.org/10.1007/s10032-006-0024-y. 10.1007/s10032-006-0024-y.

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 665–674, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367587. URL http://doi.acm.org/10.1145/1367497.1367587.

Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Clough. Detecting text reuse with modified and weighted n-grams. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 54–58, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2387636.2387646.

L. Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. RECAPTCHA: Human-based character recognition via web security measures. *Science*, 321:1465–1468, 2008. doi: 10.1126/science.1160379.

Vicent Alabau and Luis Leiva. Transcribing handwritten text images with a word soup game. In *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts*, CHI EA '12, pages 2273–2278, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1016-1. doi: 10.1145/2223656.2223788. URL http://doi.acm.org/10.1145/2223656.2223788.

O. Alonso and S. Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

Omar Alonso and Ricardo Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In Paul Clough, Colum Foley, Cathal Gurrin, Gareth Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch, editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 153–164. Springer Berlin / Heidelberg, 2011. ISBN 978-3-642-20160-8. URL `http://dx.doi.org/10.1007/978-3-642-20161-5_16`. 10.1007/978-3-642-20161-5_16.

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, November 2008. ISSN 0163-5840. doi: 10.1145/1480506.1480508. URL `http://doi.acm.org/10.1145/1480506.1480508`.

D.P. Anderson. Boinc. `http://boinc.berkeley.edu/wiki/How_BOINC_works`, a. Accessed: 25/03/2013.

D.P. Anderson. Bolt. `http://boinc.berkeley.edu/trac/wiki/BoltIntro`, b. Accessed: 25/03/2013.

D.P. Anderson. Bossa. `http://boinc.berkeley.edu/trac/wiki/BossaIntro`, c. Accessed: 25/03/2013.

D.P. Anderson. Boinc: a system for public-resource computing and storage. In *Grid Computing, 2004. Proceedings. Fifth IEEE/ACM International Workshop on*, pages 4 – 10, nov. 2004. doi: 10.1109/GRID.2004.14.

D.P. Anderson and G. Fedak. The computational and storage potential of volunteer computing. In *Cluster Computing and the Grid, 2006. CCGRID 06. Sixth IEEE International Symposium on*, volume 1, pages 73 –80, may 2006. doi: 10.1109/CCGRID.2006.101.

John CL Andreassen. Archives in the library of congress. *American Archivist*, 12(1):20–26, 1949.

Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

Jeremy Bentham. *The Collected Works of Jeremy Bentham: Correspondence, Volume 11: January 1822 to June 1824*, volume 1. Clarendon Press, 2000.

Louis Brenner and David Robinson. Project for the conservation of malian arabic manuscripts. *History in Africa*, 7:pp. 329–332, 1980. ISSN 03615413. URL `http://www.jstor.org/stable/3171669`.

Michael S Brown and W Brent Seales. Document restoration using 3d shape: a general deskewing algorithm for arbitrarily warped documents. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 367–374. IEEE, 2001.

Michael S Brown and W Brent Seales. Image restoration of arbitrarily warped documents. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1295–1306, 2004.

B Brumfield. The year of crowdsourcing transcription in the collaborative manuscript transcription blog (2011). URL `http://manuscripttranscription.blogspot.com/2011/02/2010-year-of-crowdsourcing.html`.

N.R. Budhathoki. Participants ‘motivations to contribute geographic information in an online community. 2010.

M.F. Bulut, Y.S. Yilmaz, and M. Demirbas. Crowdsourcing location-based queries. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 513 –518, march 2011. doi: 10.1109/PERCOMW.2011.5766944.

Stefan Büttcher, Charles Clarke, and Gordon V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines.* The MIT Press, 2010. ISBN 0262026511, 9780262026512.

Chris Callison-Burch. Fast, cheap, and creative: evaluating translation quality using amazon’s mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP ’09, pages 286–295, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL `http://dl.acm.org/citation.cfm?id=1699510.1699548`.

T Causer and V Wallace. Building a volunteer community: results and findings from transcribe bentham. *Digital Humanities Quarterly*, 6, 2012. URL `http://discovery.ucl.ac.uk/1362050/`.

Tim Causer, Justin Tonra, and Valerie Wallace. Transcription maximized; expense minimized? crowdsourcing and editing the collected works of jeremy bentham. *Literary and Linguistic Computing*, 27(2):119–137, 2012. doi: 10.1093/llc/fqs004. URL `http://llc.oxfordjournals.org/content/27/2/119.abstract`.

G Sayeed Choudhury, Robert Ferguson, Michael Droettboom, Ichiro Fujinaga, and Tim DiLauro. Document recognition for a million books. *D-Lib Magazine*, 12(3):4, 2006.

AltonY.K. Chua and RadhikaShenoy Balkunje. Comparative evaluation of community question answering websites. In Hsin-Hsi Chen and Gobinda Chowdhury, editors, *The Outreach of Digital Libraries: A Globalized Resource Network*, volume 7634 of *Lecture Notes in Computer Science*, pages 209–218. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34751-1.

doi: 10.1007/978-3-642-34752-8_27. URL http://dx.doi.org/10.1007/978-3-642-34752-8_27.

Cyril Cleverdon. Readings in information retrieval. chapter The Cranfield tests on index language devices, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1-55860-454-5. URL http://dl.acm.org/citation.cfm?id=275537.275544.

William F Clocksin. Towards automatic transcription of syriac handwriting. In *Image Analysis and Processing, 2003. Proceedings. 12th International Conference on*, pages 664–669. IEEE, 2003.

J.P. Cohn. Citizen science: Can volunteers do real research? *BioScience*, 58 (3):192–197, 2008.

Patrick W. Conner. The beowulf workstation: One model of computer-assisted literary pedagogy. *Literary and Linguistic Computing*, 6(1):50–58, 1991. doi: 10.1093/llc/6.1.50. URL http://llc.oxfordjournals.org/content/6/1/50.abstract.

G. Cook. How crowdsourcing is changing science. *Boston Globe, November*, 11, 2011.

Khalil Dahab and Anja Belz. A game-based approach to transcribing images of text. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Reim Doumat, Elöd Egyed-Zsigmond, Jean-Marie Pinon, and Emese Csiszar. Online ancient documents: Armarius. In *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, pages 127–130, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-081-4. doi: 10.1145/1410140.1410167. URL http://doi.acm.org/10.1145/1410140.1410167.

C. Eickhoff. Introduction to crowdsourcing. *Delft University of Technology*, 2011.

A. Farouk-Alli and M.S. Mathee. The tombouctou manuscript project: social history approaches.

Shaolei Feng, N Howe, and R Manmatha. A hidden markov model for alphabet-soup word recognition. In *Proc. IEEE Int. Conf. on Frontiers in Handwriting Recognition (ICFHR 2008)*, pages 210–215, 2008.

P. Fichman. A comparative assessment of answer quality on four question answering sites. *Journal of Information Science*, 37(5):476–486, 2011. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-80054066889&partnerID=40&md5=ede071964a9560c9c102849ad4123b59`. cited By (since 1996) 2.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1866696.1866709`.

Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. Crowddb: answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, SIGMOD '11, pages 61–72, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0661-4. doi: 10.1145/1989323.1989331. URL `http://doi.acm.org/10.1145/1989323.1989331`.

Qin Gao and Stephan Vogel. Consensus versus expertise: a case study of word alignment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 30–34, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1866696.1866700`.

Peter S Graham. New roles for special collections on the network. *College & Research Libraries*, 59(3):232–239, 1998.

L. Guichard, J. Chazalon, and B. Couasnon. Exploiting collection level for improving assisted handwritten word transcription of historical documents. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 875 –879, sept. 2011. doi: 10.1109/ICDAR.2011.179.

Neha Gupta and Pratistha Mathur. Spell checking techniques in nlp: A survey. *International Journal*, 2(12), 2012.

Aria Haghighi, Percy Liang, Taylor B. Kirkpatrick, and Dan Klein. Learning Bilingual Lexicons from Monolingual Corpora. In *Proc. ACL-HLT*, pages 771–779, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology-new/P/P08/P08-1088.bib`.

Thomas A Hale. Manuscripts of timbuktu (review). *African Studies Review*, 55(2):198–199, 2012.

S. Heaney. Introduction'to beowulf, 1999.

G. Heng. An experiment in collaborative humanities: Imagining the world, 500–1500 ce. 2007.

Vaughn Hester, Aaron Shaw, and Lukas Biewald. Scalable crisis relief: Crowdsourced sms translation and categorization with mission 4636. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 15:1–15:7, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0473-3. doi: 10.1145/1926180.1926199. URL `http://doi.acm.org/10.1145/1926180.1926199`.

Sanjika Hewavitharana and Stephan Vogel. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 61–68, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-015. URL `http://dl.acm.org/citation.cfm?id=2024236.2024248`.

M. Hossain. Users' motivation to participate in online crowdsourcing platforms. In *Innovation Management and Technology Research (ICIMTR), 2012 International Conference on*, pages 310 –315, may 2012. doi: 10.1109/ICIMTR.2012.6236409.

John Hutchins. Two precursors of machine translation: Artsrouni and trojanskij. *International Journal of Translation*, 16:1–11, 2004.

Panagiotis G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17(2):16–21, December 2010. ISSN 1528-4972. doi: 10.1145/1869086.1869094. URL `http://doi.acm.org/10.1145/1869086.1869094`.

Aminul Islam and Diana Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, 2 (2):10:1–10:25, July 2008. ISSN 1556-4681. doi: 10.1145/1376815.1376819. URL `http://doi.acm.org/10.1145/1376815.1376819`.

Shaun Kane, Andrew Lehman, and Elizabeth Partridge. Indexing george washington's handwritten manuscripts. *Center for Intelligent Information Retrieval, Computer Science Department, University of Massachusetts, Amherst, MA*, 1003, 2001.

B. Kanefsky, N. G. Barlow, and V. C. Gulick. Can Distributed Volunteers Accomplish Massive Data Analysis Tasks? In *Lunar and Planetary Institute Science Conference Abstracts*, volume 32 of *Lunar and Planetary Inst. Technical Report*, page 1272, March 2001.

N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. In *Proceedings of* *the Seventeenth Americas Conference on Information Systems, Detroit, MI*, 2011.

Gabriella Kazai, Natasa Milic-Frayling, and Jamie Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 452–459, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572019. URL `http://doi.acm.org/10.1145/1571941.1572019`.

Kevin S. Kiernan. Digital image processing and the beowulf manuscript. *Literary and Linguistic Computing*, 6(1):20–27, 1991. doi: 10.1093/llc/6.1.20. URL `http://llc.oxfordjournals.org/content/6/1/20.abstract`.

Soojung Kim and Sanghee Oh. Users' relevance criteria for evaluating answers in a social q&a site. *Journal of the American Society for Information Science and Technology*, 60(4):716–727, 2009. ISSN 1532-2890. doi: 10.1002/asi.21026. URL `http://dx.doi.org/10.1002/asi.21026`.

Aniket Kittur. Crowdsourcing, collaboration and creativity. *XRDS*, 17(2):22–26, December 2010. ISSN 1528-4972. doi: 10.1145/1869086.1869096. URL `http://doi.acm.org/10.1145/1869086.1869096`.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357127. URL `http://doi.acm.org/10.1145/1357054.1357127`.

Steve Klass. Monsters. artsedge curricula, lessons and activities. 2002.

V. Lavrenko, T.M. Rath, and R. Manmatha. Holistic word recognition for handwritten historical documents. In *Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on*, pages 278 – 287, 2004. doi: 10.1109/DIAL.2004.1263256.

VI Lcvenshtcin. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, volume 10, 1966.

John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 21–26, 2010.

Matthew Lease and Emine Yilmaz. Crowdsourcing for information retrieval. *SIGIR Forum*, 45(2):66–75, January 2012. ISSN 0163-5840. doi: 10.1145/2093346.2093356. URL `http://doi.acm.org/10.1145/2093346.2093356`.

J.H. Lee. Crowdsourcing music similarity judgments using mechanical turk. *Proc. of ISMIR 2010*, pages 183–188, 2010.

Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, pages 129–138, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1154-0. doi: 10.1145/2232817.2232842. URL `http://doi.acm.org/10.1145/2232817.2232842`.

Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey∗. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2008.13689.x. URL `http://dx.doi.org/10.1111/j.1365-2966.2008.13689.x`.

P.J. Lucas. The place of judith in the beowulf-manuscript. *Review of English Studies*, pages 463–478, 1990.

A.M. Lund. Measuring usability with the use questionnaire. *Usability and User Experience*, 8(2):8, 2001.

R. Manmatha and J.L. Rothfeder. A scale space approach for automatically segmenting words from historical handwritten documents. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1212 –1225, aug. 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.150.

Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky. Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 99–107, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1866696.1866712`.

Jim Maurer. A conversation with david anderson. *Queue*, 3(6):18–25, July 2005. ISSN 1542-7730. doi: 10.1145/1080862.1080872. URL `http://doi.acm.org/10.1145/1080862.1080872`.

David Meyer. Icelanders approve their crowdsourced constitution @ONLINE, October 2012a. URL `http://gigaom.com/europe/icelanders-approve-their-crowdsourced-constitution/`.

David Meyer. Finland is about to start using crowdsourcing to create new laws @ONLINE, September 2012b. URL `http://gigaom.com/europe/online-crowdsourcing-can-now-help-build-new-laws-in-finland/`.

Stefano Mizzaro. Relevance: the whole history. *J. Am. Soc. Inf. Sci.*, 48(9):810–832, September 1997. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(199709)48:9⟨810::AID-ASI6⟩3.0.CO;2-U. URL `http://dx.doi.org/10.1002/(SICI)1097-4571(199709)48:9<810::AID-ASI6>3.0.CO;2-U`.

A.B. M'kadem and P. Nieuwenhuysen. Digital access to cultural heritage material: Case of the moroccan manuscripts. *Collection Building*, 29(4):137–141, 2010. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-78049490434&partnerID=40&md5=f05fcf0e9becf39ca6226fc09916c80e`. cited By (since 1996) 1.

Matteo Negri and Yashar Mehdad. Creating a bi-lingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 212–216, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1866696.1866730`.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 670–679, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4. URL `http://dl.acm.org/citation.cfm?id=2145432.2145510`.

G.B. Newby and C. Franks. Distributed proofreading. In *Digital Libraries, 2003. Proceedings. 2003 Joint Conference on*, pages 361 – 363, may 2003. doi: 10.1109/JCDL.2003.1204888.

Jakob Nielsen and Thomas K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pages 206–213, New York, NY, USA, 1993. ACM. ISBN 0-89791-575-5. doi: 10.1145/169059.169166. URL `http://doi.acm.org/10.1145/169059.169166`.

Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-815-5. doi: 10.1145/1743384.1743478. URL `http://doi.acm.org/10.1145/1743384.1743478`.

C. Oppenheim and D. Smithson. What is the hybrid library? *Journal of information science*, 25(2):97–112, 1999.

P. Oxy. The oxyrhynchus papyri. *Published by the Egypt Exploration Society in Graeco-Roman Memoirs. London*, 2007, 1898.

R. Plamondon and S.N. Srihari. Online and off-line handwriting recognition: a comprehensive survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):63 –84, jan 2000. ISSN 0162-8828. doi: 10.1109/34.824821.

Dorothy Carr Porter. *Medievalists' Use Of Electronic Resources: The Results of a National survery of Faculty Members in Medieval Studies.* PhD thesis, University of North Carolina, 2002.

A. Prescott. The electronic beowulf and digital restoration. *Literary and Linguistic Computing*, 12(3):185–196, 1997.

Andrew Prescott. Constructing electronic beowulf. *Towards the Digital Library: The British Library's 'Initiatives for Access' Programme*, pages 30–49, 1998.

Alexander J. Quinn and Benjamin B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979148. URL `http://doi.acm.org/10.1145/1978942.1979148`.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1866696.1866717`.

R Manmatha Toni M Rath. Indexing of handwritten historical documents-recent progress. In *Proceedings 2003 Symposium on Document Image Understanding Technology*, page 77. Umd, 2003.

Toni M. Rath, R. Manmatha, and Victor Lavrenko. A search engine for historical manuscript images. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 369–376, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. doi: 10.1145/1008992.1009056. URL `http://doi.acm.org/10.1145/1008992.1009056`.

Tony M Rath and Rudrapatna Manmatha. Word spotting for historical documents. *International Journal on Document Analysis and Recognition*, 9(2):139–152, 2007.

D. Ryan. Aluka: Digitization from maputo to timbuktu. *OCLC Systems and Services*, 26(1):29–38, 2010. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-77049089713&partnerID=40&md5=e557419ae9a6cda5d0738c2a22e8508a`. cited By (since 1996) 1.

Matthew L. Saxton and John V. Richardson. *Understanding Reference Transactions: Transforming an Art into a Science*. Academic Press, 2002.

Nicolás Serrano, Adrià Giménez, Albert Sanchis, and Alfons Juan. Active learning strategies for handwritten text transcription. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, pages 48:1–48:4, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0414-6. doi: 10.1145/1891903.1891962. URL `http://doi.acm.org/10.1145/1891903.1891962`.

P. Shachaf. The paradox of expertise: Is the wikipedia reference desk as good as your library? *Journal of Documentation*, 65(6):977–996, 2009. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-71849106183&partnerID=40&md5=dd19d32c364e358e4da0bfbfaa7e6b27`. cited By (since 1996) 13.

J. Silvertown. A new dawn for citizen science. *Trends in Ecology and Evolution*, 24(9):467–471, 2009. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-68949137010&partnerID=40&md5=d6f6aaca0d6d28894ef99a89db16c360`. cited By (since 1996) 71.

M.D. Smucker and C.P. Jethani. The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613715.1613751`.

A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1 –8, june 2008. doi: 10.1109/CVPRW.2008.4562953.

H. Suleman. An african perspective on digital preservation.

H. Suleman. Digital libraries without databases: The bleek and lloyd collection. *Research and Advanced Technology for Digital Libraries*, pages 392–403, 2007.

James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

A.H. Toselli, V. Romero, E. Vidal, and L. Rodriguez. Computer assisted transcription of handwritten text images. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 944 –948, sept. 2007. doi: 10.1109/ICDAR.2007.4377054.

Alejandro H. Toselli, VerÃ³nica Romero, MoisÃ©s Pastor, and Enrique Vidal. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814 – 1825, 2010. ISSN 0031-3203. doi: 10.1016/j.patcog.2009.11.019. URL `http://www.sciencedirect.com/science/article/pii/S0031320309004385`.

Deborah J. Trumbull, Rick Bonney, Derek Bascom, and Anna Cabral. Thinking scientifically during participation in a citizen-science project. *Science Education*, 84(2):265–275, 2000. ISSN 1098-237X. doi: 10.1002/(SICI)1098-237X(200003)84:2⟨265::AID-SCE7⟩3.0.CO;2-5. URL `http://dx.doi.org/10.1002/(SICI)1098-237X(200003)84:2<265::AID-SCE7>3.0.CO;2-5`.

Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008. doi: 10.1126/science.1160379. URL `http://www.sciencemag.org/content/321/5895/1465.abstract`.

C. Webb. *Guidelines for the preservation of digital heritage.*

A. J. Westphal, J. von Korff, D. P. Anderson, A. Alexander, B. Betts, D. E. Brownlee, A. L. Butterworth, N. Craig, Z. Gainsforth, B. Mendez, T. See, C. J. Snead, R. Srama, S. Tsitrin, J. Warren, and M. Zolensky. Stardust@home: Virtual Microscope Validation and First Results. In S. Mackwell and E. Stansbery, editors, *37th Annual Lunar and Planetary Science Conference*, volume 37 of *Lunar and Planetary Inst. Technical Report*, page 2225, March 2006.

Kyle Williams. Feasibility of automatic transcription of neatly rewritten bushman texts. Technical report, 2010.

Kyle Williams. Learning to read bushman: Automatic handwriting recognition for bushman languages. 2012.

Kyle Williams and Hussein Suleman. Creating a handwriting recognition corpus for bushman languages. In *Proceedings of the 13th international conference on Asia-pacific digital libraries: for cultural heritage, knowledge dissemination, and future creation*, ICADL'11, pages 222–231, Berlin, Heidelberg, 2011a. Springer-Verlag. ISBN 978-3-642-24825-2. URL `http://dl.acm.org/citation.cfm?id=2075271.2075305`.

Kyle Williams and Hussein Suleman. Using a hidden markov model to transcribe handwritten bushman texts. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 445–446, New York, NY, USA, 2011b. ACM. ISBN 978-1-4503-0744-4. doi: 10.1145/1998076.1998177. URL `http://doi.acm.org/10.1145/1998076.1998177`.

Kyle Williams, Sanvir Manilal, Lebogang Molwantoa, and Hussein Suleman. A visual dictionary for an extinct language. In Gobinda Chowdhury, Chris Koo, and Jane Hunter, editors, *The Role of Digital Libraries in a Time of Global Change*, volume 6102 of *Lecture Notes in Computer Science*, pages 1–4. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-13653-5. doi: 10.1007/978-3-642-13654-2_1. URL `http://dx.doi.org/10.1007/978-3-642-13654-2_1`.

Mamoru Yoshimura, Fuminori Kimura, and Akira Maeda. Word segmentation for text in japanese ancient writings based on probability of character n-grams. In Hsin-Hsi Chen and Gobinda Chowdhury, editors, *The Outreach of Digital Libraries: A Globalized Resource Network*, volume 7634 of *Lecture Notes in Computer Science*, pages 313–316. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34751-1. doi: 10.1007/978-3-642-34752-8_38. URL `http://dx.doi.org/10.1007/978-3-642-34752-8_38`.

Li Yujian and Liu Bo. A normalized levenshtein distance metric. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1091 –1095, june 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1078.

Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1220–1229, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL `http://dl.acm.org/citation.cfm?id=2002472.2002626`.

Appendix A

Appendices

A.1 Bossa Job Creation Script

```
 <?php
//Given a directory of images, make a batch of jobs
//Usage:
//bossa_transcribe_make_jobs.php options
//--dir dir

$app_name = "transcribe";
$cli_only = true;

require_once("../inc/bossa.inc");
require_once("../inc/util_ops.inc");

function make_job($path, $batchid, $appid) {
$info = null;
$info->path = $path;

if (!bossa_job_create($appid, $batchid, $info, false)) {
    exit("bossa_create_job() failed\n");
}
echo "created job for $path\n";
}

function make_jobs($dir, $appid) {
$batchid = bossa_batch_create($appid, date(DATE_RFC822), false);
if (!$batchid) {
    exit("bossa_create_batch() failed\n");
}

$d = "../xoaxoa/$dir";
$iterator = new RecursiveIteratorIterator(new RecursiveDirectoryIterator($d),
        RecursiveIteratorIterator::CHILD_FIRST);

        foreach ($iterator as $path) {
          if ($path->isDir()) {
//do nothing
          } else {
         if (!strstr($path, ".JPG")) continue;
         make_job("$path", $batchid, $appid);
        }
     }
}

function usage() {
exit("Usage: bossa_transcribe_make_jobs.php --dir d\n");
}

for($i=1; $i<$argc; $i++) {
if ($argv[$i] == '--dir') $dir = $argv[++$i];
else usage();
}

if (!$dir) usage();

if (!is_dir("../xoaxoa/$dir")) {
exit("../xoaxoa/$dir is not a directory\n");
}

$appid = bossa_app_lookup($app_name);
if (!$appid) exit("No application $app_name\n");

make_jobs($dir, $appid);

?>
```

A.2 Survey Questions

Survey Goal

Thank your for participating in the Transcribe Bleek and Lloyd Project, your input is greatly appreciated. As part of my Master's research , I am evaluating the transcription tool to better understand your experience using the tool.

Survey Instructions

You will be required to answer multiple choice questions and a few short free text questions, this should approximately take you 5-10minutes to complete.

Section A: Participant Background Information

A1. What is your age range?

21 and Under □
22 to 34 □
35 to 44 □
45 to 54 □
55 to 64 □
65 and Over □

A2. Gender

Female □
Male □

A3. **Have you ever worked on any other transcription/crowdsourcing projects?**

Yes □

No □

A4. **List the projects you have worked on**

Section B: System Design

B1. **Please rate your agreement with these statements.**

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The flow in the layout of the website was easy to follow/understand.	□	□	□	□	□
It is difficult to move around this web site.	□	□	□	□	□
The organization of information on the interface is clear.	□	□	□	□	□
I found the interface too cluttered.	□	□	□	□	□
I could easily find my profile details and change them.	□	□	□	□	□
The leader-board motivated me to transcribe more images.	□	□	□	□	□

Section C: Transcription Tool: Satisfaction

C1. Please rate your agreement with these statements.

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The tool is pleasant to use.	□	□	□	□	□
I felt frustrated using the tool.	□	□	□	□	□
The tool is fun to use.	□	□	□	□	□
I would not recommend the tool to a friend.	□	□	□	□	□
I felt satisfied when I used the tool.	□	□	□	□	□
I didn't enjoy using the tool.	□	□	□	□	□
It does everything I would expect it to do.	□	□	□	□	□
The tool motivated me to continue using it.	□	□	□	□	□
I felt in control when I was using the tool.	□	□	□	□	□
The tool meets my needs for this task.	□	□	□	□	□

Section D: Transcription Tool: Ease of Use

D1. Please rate your agreement with these statements.

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The transcription tool was intuitive to use.	□	□	□	□	□
I had to watch the transcription tutorial to learn how to use the tool.	□	□	□	□	□
I made many errors while transcribing.	□	□	□	□	□
Correcting my mistakes was easy.	□	□	□	□	□
The tool is designed for all levels of users.	□	□	□	□	□
Only experts can use this tool.	□	□	□	□	□

Section E: Transcription Tool: Ease of Learning and Tool Features

E1. Please rate your agreement with these statements.

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Learning how to use the tool was easy.	□	□	□	□	□
I became productive quickly using the transcription tool.	□	□	□	□	□
I could find most of the characters in the “Bushman images” on the Characters	□	□	□	□	□
I understand what the Characters	□	□	□	□	□
I understood that the left text area was for Bushman text and that the right text area was for entering English text.	□	□	□	□	□
The navigation features (e.g. zoom in and out) were helpful to properly visualize the text on the images.	□	□	□	□	□
I had no need to use the navigation features (e.g. zoom in and out).	□	□	□	□	□
I could easily preview the Bushman text I transcribed.	□	□	□	□	□

Section F: Transcriptiion Tool: Short Answers

F1. Name some positive aspects about the transcription tool.

F2. Name some negative aspects of the transcription tool.

F3. **Any general comments?**

Thank you again, for helping complete this survey and your participation in the Transcribe Bleek and Lloyd Project. Feel free to continue transcribing more pages, and invite your friends to participate.

A.3 User Feedback Comments

A.3.1 Positive Aspects of Transcription Tool

User 1: Easy to understand and use.

User 2: The tool in itself is simple to use and tutorial video is very helpful.

User 3: I think the palette handles the complexity of the character set very well. This material is inherently difficult to transcribe. The tool has, on the whole, been well thought out to meet this challenge. I think it needs to be improved in some ways, but considering the difficulties it is remarkably well done.

User 4: Very intuitive, after a few practice transcriptions. I actually enjoyed using the tool after a page was done.

User 5: It is great in helping to learn and perhaps understand the Bushman languages.

User 6: Image of the source material was of good quality.

User 6: I like that you can zoom into and move around the page and that you can write the Bushman characters using either the pallet or by writing the latex command if you have become familiar with it.

User 7: It covered almost all of the symbols/letters. Made it easier to understand the languages.

User 8: It was very intuitive, simple and not cluttered by any means! It was fun to use.

User 9: Interesting and unique.

A.3.2 Negative Aspects of Transcription Tool

User 1: Sometimes I felt like I couldn't find the diacritics or special characters I needed and had to choose an option I felt was incorrect.

User 2: It takes a lot of time trying to figure out the characters and some characters were hard to read on the images.

User 3: Some of the diacritics were not in the palette or if put into the text were displayed a bit differently than they were supposed to. Also, maybe there should be a possibility of marking places in the manuscript that were crossed out or marked by *. I think we should be able to transcribe anything, also crossed out words so that you have a complete image of the original manuscript and nothing is left out.

User 4: In the original English is on the left-hand side but in the tool we have to type it into the right column.

User 5: The Latex markup makes it rather difficult to find where to edit a mistake.

User 6: A lot of diacritics are missing in the palette, also (and I know this is not your fault) the handwriting is pretty often impossible to decipher.

User 7: The "save" button should say something like "finish and exit" to show that you cannot continue after pressing it. In other software, "save" means keep your work but keep on working. I pressed "save" before I finished a page, thinking I would be able to keep on working, but it took me out of the page and I could not find a way to get back in. The result is that I left a page unfinished. It is not clear to me whether this comment that I am writing now will be associated with the specific page that I left unfinished. I hope it will be. If it is not, I hope that a way can be provided to report specific problems or questions for an individual page.

User 8: The 'Save Page' button was too close to the 'Latex' button - I accidentally submitted a page as complete when I was not done with it.

User 9: Page in Google Chrome scrolls back to 0 when adding a diacritic.

User 10: I like the Latex preview and think it is important but it is annoying since it obscures the working area and the way to close it is not immediately evident. If working on a smaller screen using the tool can be frustrating since there isn't enough space for the image and the work area and the characters pallet on the screen. Sometimes when you add characters from the pallet focus does not return automatically to the text area and sometimes the text area scrolls to the beginning of the text, this is very frustrating.

User 11: Some frustrating moments where I couldn't find the symbols. When you would preview your translating, it only showed about the first line.

User 12: I had some problems with the Convert to Latex function, so it was difficult to review the text I had worked on about halfway through.

User 13: Some of the diacritics were not in the palette or if put into the text were displayed a bit differently than they were supposed to. Also, maybe there should be a possibility of marking places in the manuscript that were crossed out or marked by *. I think we should be able to transcribe anything, also crossed out words so that you have a complete image of the original manuscript and nothing is left out.

User 14: Could not understand the audio on video AT ALL; watched 2x; did not understand CONVERT TO LATEX–it did not always correspond to what I wrote so I could not evaluate my transcription; difficult to line up Bushman/English transcriptions; scrolling down to increase transcription space was not evident to me on 1st page so did not finish it; extraneous marks on English side, so could not always read letters.

A.3.3 General Comments

User 1: What I really need as a user is the ability to save my progress on a job and continue working on it later before submitting it. The individual jobs are too long without this feature. Being unfamiliar with the language, I would also appreciate specimen samples of the handwriting. Many times I felt unsure whether I was looking at an r, n, s, k, h, u, etc. Specimen samples would also be useful in cases of transcribing the English because it seemed that some of the words were unfamiliar cultural loanwords or place names.

User 2: It would be great if the tutorial gave instructions on 1) how to transcribe text that has been crossed out by the author and 2)how to transcribe/mark a character that you cannot read. It would also be great if you could go back to pages you've already transcribed if you realize you've made a mistake on one of them as you gain more experience using the tools.

User 3: I think there should be a way to report a problem or make notes when working on a page. For example, when looking at one of the hand-written letters, I could not tell if it was a ”u” or a ”w”. I wish I could have attached a note about that to the page. I wish I had a way to return to a page after pressing the ”save” button. I know now that ”save” really means ”finish and exit”, so I will be careful not to press it until I am sure I am done. But even then, I might later think of something to improve, and want to go back.

User 4: It is sometimes difficult reading the original text; the writer's handwriting was NOT intuitive. I think this is a great tool; I would like to see links to descriptive material on how the language works.

User 5: This is a great idea. If it's possible to have an automatic transcription tool, it would be very helpful.

User 6: I would like to see some real time rendering of the latex, possibly adjacent the source image in some way.

User 7: The tool is very useful and a very good way to transcribe this text.

User 8: The English should have been already translated and it was frustrating trying to figure some of the words for the English side. Also, sometimes it was difficult knowing what had to be transcribed and what not.

User 9: Very good experience!

User 10: Original tutorial and initial interface needs improvement–role of Convert Latex, how to evaluate transcription.

User 11: On the page I just worked on, there were notes on the right page explaining words or pointing different spellings that I think were relevant, but the video indicates they were not the focus so i did not include them, and I couldn't even if i wanted to. Maybe someone should check that.