Noname manuscript No.
(will be inserted by the editor)

A System for High Quality Crowdsourced Indigenous Language Transcription

Ngoni Munyaradzi · Hussein Suleman

Received: date / Accepted: date

Abstract In this article, a crowdsourcing method is proposed to transcribe manuscripts from the Bleek and Lloyd Collection, where non-expert volunteers transcribe pages of the handwritten text using an online tool. The digital Bleek and Lloyd Collection is a rare collection that contains artwork, notebooks and dictionaries of the indigenous people of Southern Africa. The notebooks, in particular, contain stories that encode the language, culture and beliefs of these people, handwritten in now-extinct languages with a specialised notation system. Previous attempts have been made to convert the approximately 20000 pages of text to a machine-readable form using machine learning algorithms but, due to the complexity of the text, the recognition accuracy was low. This article presents details of the system used to enable transcription by volunteers as well as results from experiments that were conducted to determine the quality and consistency of transcriptions. The results show that volunteeers 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 a gold standard, the volunteers achieve an average accuracy of 64.75%, which exceeded that in previous work. Finally, the degree of transcription agreement correlates with the degree of transcription accuracy. This suggests that the quality of unseen data can be assessed based on the degree of agreement among transcribers.

Keywords crowdsourcing · transcription · cultural heritage

N. Munyaradzi
University of Cape Town
Tel.: +27216502663
Fax: +27216503551
E-mail: ngoni.munyaradzi@uct.ac.za

H. Suleman
University of Cape Town
Tel.: +27216505106
Fax: +27216503551
E-mail: hussein@cs.uct.ac.za

1 Introduction

The digital Bleek and Lloyd Collection [13] is a collection of scanned notebooks, dictionaries and artwork that document the culture and beliefs of the indigenous people of Southern Africa. The notebooks, specifically, contain 20000 pages of bilingual text that document the stories and languages of speakers of the now-extinct |Xam and !Kun languages. These notebooks were created by linguistics researchers in the mid-1800s and are the most authoritative source of information on the then indigenous population. Figure 1 shows a typical set of facing pages from one of the notebooks.

Transcriptions of the scanned notebooks would make the text indexable and searchable. It would also enable translation, text-to-speech and other forms of processing that are currently not possible. Manual transcription is a possibility but this is an expensive solution and not one that can easily be adapted to similar problems for other digital collections and other forms of document processing, especially in resource-constrained environments.

An alternative is presented by the Citizen Cyberscience movement [4], where ordinary citizens are recruited to volunteer their time and/or computational resources to solve scientific problems, often with benefit to the public. Such problems include mapping of roads in rural Africa and monitoring of disease spread, (e.g.,

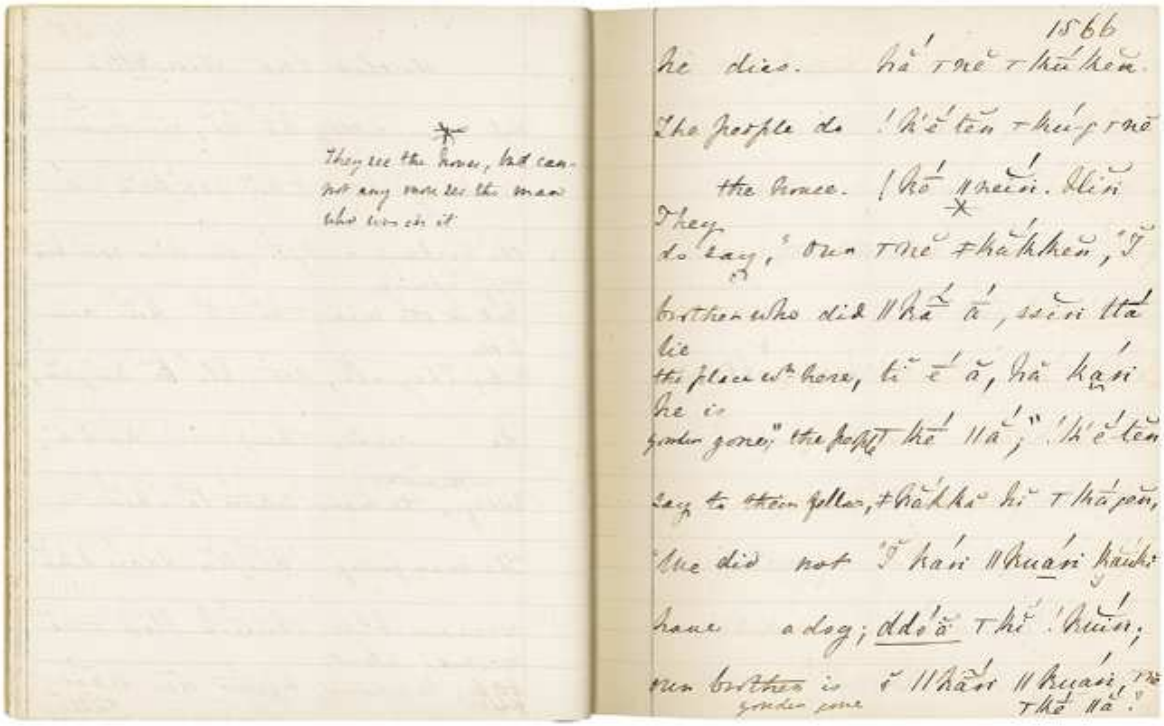

Fig. 1 Sample page from Bleek and Lloyd notebooks

FightMalaria@Home). In typical projects, each volunteer is given one or more small tasks via a Web interface and these tasks are collated to solve a larger problem.

This project is based on the premise that the preservation of cultural heritage is of importance to ordinary citizens, who could therefore be recruited as volunteers to transcribe handwritten documents. The Bossa [2] framework for distributed/volunteer thinking was used to develop a transcription application.

This article investigates the feasibility and accuracy of volunteer transcription, as one example of an intellectually intensive task in digital libraries, and how it compares to computational techniques like machine learning.

The rest of this article is structured as follows: Section 2 discusses the background and related work that serves as a foundation and motivation for the approach used in this research; Section 3 describes the Bossa volunteer framework used to harness distributed human computation power and how it was applied to this problem; Section 4 describes the transcription tool; Section 5 presents an analysis of the initial results; and Section 6 draws conclusions and discusses future work.

2 Related Work

Crowdsourcing (or volunteer thinking) has been applied to solve various problems related to information search and discovery. Volunteer thinking may be defined as crowdsourcing with volunteers, as opposed to paid workers.

Shachaf [12] investigated the quality of answers on the Wikipedia Reference Desk, and compared it with library reference services to determine whether volunteers can outperform expert reference librarians. Their results show that both systems provide reference services at the 55% accuracy level. Overall, the volunteers outperform the expert librarians – this is significant because the volunteers are amateurs and not paid for their services. The individual responses submitted by volunteers were comparable to those of librarians, but the amalgamated responses from volunteers produced answers that were similar or better than those of expert librarians.

Clickworkers [6] is an example of a citizen science project, set up by NASA, where volunteers identify and classify the age of craters on Mars images. The objectives of such citizen science projects include determining if volunteers are ready and willing to contribute to science and if this new way of conducting science produces results that are as good as earlier established methods. Ongoing work by Callison-Burch [3], Nowak

[11] and others has shown that both questions can be answered in the affirmative.

reCAPTCHA[1] is a snippet transcription tool used for security against automated programs. reCAPTCHA is used to digitize books, newspapers and old time radio shows. 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% [14]. The tasks are, however, very small and there is a strong motivation to complete them successfully as failure prevents access to whatever resource is being protected by reCAPTCHA. This is not typical of transcription projects.

The work by Causer and Wallace [5] in the Transcribe Bentham project gives an enlightening picture of the effort required to successfully create awareness about a transcription project and costs involved. Early reported results in 2012 were promising but the project included the use of professional editors and thus relied on project funding to ensure quality. In contrast, this article investigates what level of quality can be achieved solely by volunteers and automated post-processing techniques.

Williams [15] attemped to transcribe the Bleek and Lloyd notebooks solely using machine learning techniques, by performing a detailed comparison of the best known techniques. Using a highly-tuned algorithm, a transcription accuracy of 62.58% was obtained at word level and 45.10% at line level. As part of that work, Williams created a gold standard corpus of |Xam transcriptions [16], which was used in the work reported on in this article.

In summary, there have been numerous attempts at transcription, with a focus on the mechanics of the process. This article, additionally, focuses on the assessment of transcription accuracy, which is further in the context of a language that is unfamiliar to volunteers. The mechanics were greatly simplified by use of the Bossa toolkit, as discussed in the next sections.

3 The Bossa Framework

3.1 Bossa Architectural Overview

The Berkeley Open System for Skill Aggregation (Bossa) [2] is an open source software framework for distributed thinking - where volunteers complete tasks online that require human intelligence. Bossa was developed by David Anderson[2], and is part of the larger Berkeley Open Infrastructure for Network Computing (BOINC) framework - BOINC is the basis for volunteer computing projects such as SETI@Home [1]. 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 based on the concept of volunteer work with no monetary incentives.

The framework simplifies the task of creating distributed thinking projects by providing a suite of common tools and an administrative interface to manage user accounts and tasks/jobs. A well-defined machine interface in the form of a set of PHP call-back functions allows for interconnection with different custom applications.

For each application, a core database with important application details is pre-populated and can be expanded with application-specific data. The programmer can then define the actual task to be performed as a Web application, and link this to the call-back 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.

Figure 2 provides a cross-sectional view of the whole Bossa-based transcription tool, and shows how Bossa and Boinc are integrated, including the Bolt training module for Bossa. 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.

3.2 Job Distribution Policy

A task in Bossa is defined by a job and each job may have multiple instances. Each of the multiple instances is performed by a different user, thus yielding multiple results for each 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; and what threshold values have been set for each job or which jobs have higher priority. Applications have different job distribution policies, which are user-defined.

There are two standard models for job distribution: either a limited set of jobs and thousands of volunteers; or an unbounded set of jobs but a limited number of volunteers.

[1] http://www.google.com/recaptcha
[2] http://boinc.berkeley.edu/anderson/

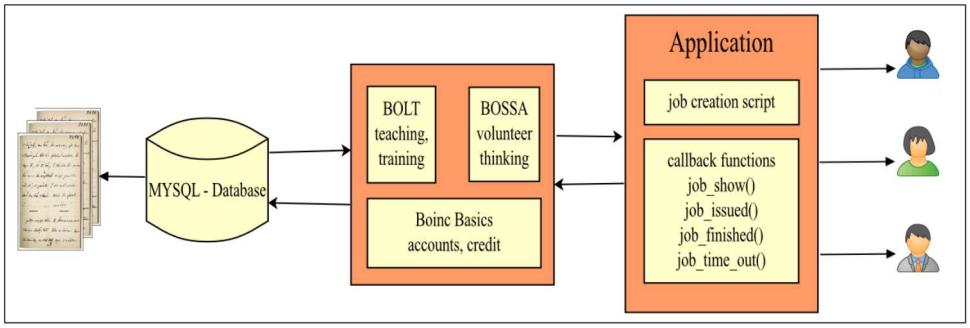

Fig. 2 Bossa-based interface for transcription of pages

In the first case, 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.

In the second case, where there is a targeted threshold of accuracy, 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 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 initially instantiated once. A volunteer can transcribe as many pages as they like. A job replication policy (discussed in Section 3.3) was then implemented to improve accuracy of results.

3.3 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 [8] for music mood classification, three relevance judgements were collected from participants. Lee [7] again collected three judgements for music similarity. Marge et al [10] 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 transcription, 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.

3.4 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:

- 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. 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 required by Bossa were implemented to display the jobs on the transcription tool interface and store result representations within the database.

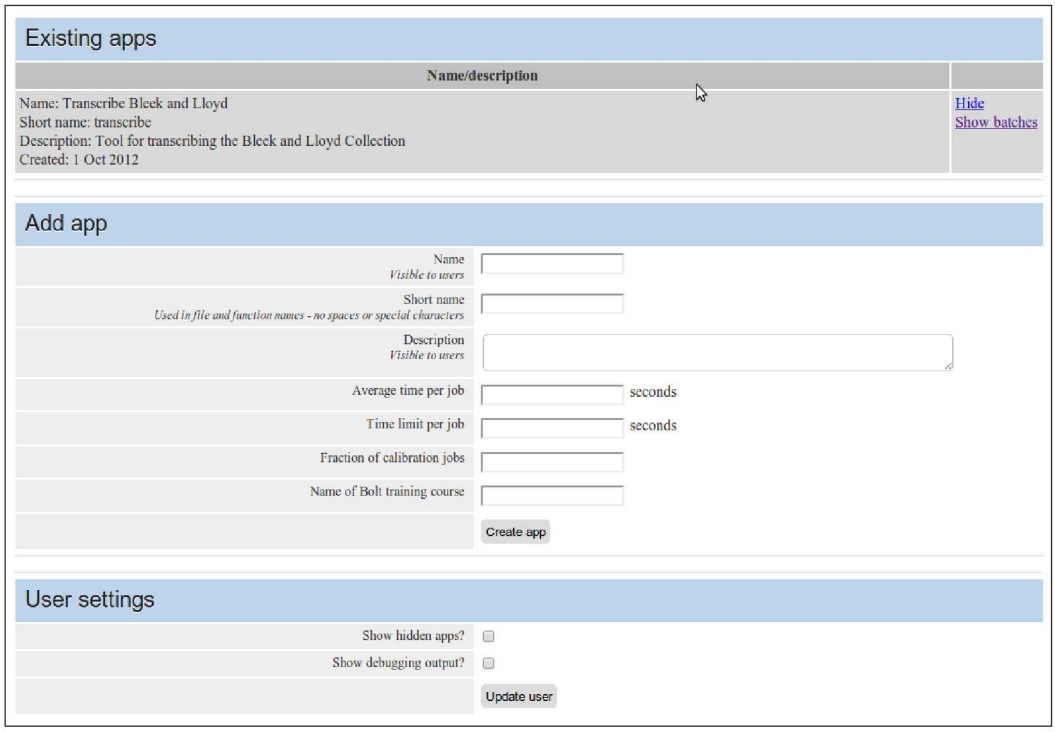

Fig. 3 Administrative interface

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.

4 Transcription Tool

4.1 Login, Registration and Qualification

In order 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 first watch a short transcription tutorial video. After the transcription tutorial, the user can begin transcribing. Other crowdsourcing projects require users to complete an assessment exercise to determine volunteer skill. This was not done for this project.

4.2 Characters and Diacritics Panel

More than 300 diacritics of the |Xam language are used in the transcription tool. Still more diacritics are being discovered in the notebooks. This language is not supported in standard Unicode representation. A specialized encoding tool was developed by Williams [15] to represent this complex script. The custom encoding tool was developed using LATEX and 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 4).

The visual representation of the encoding is a near approximation of the text in the notebooks. Future work as suggested by Williams [15] would be to develop a custom font for the languages.

4.3 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. 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 cannot be transcribed for some reason, volunteers are told to click on the Cannot Transcribe Page button (see Figure 5) The |Xam and English text are supposed to be typed into the left and right textareas respectively.

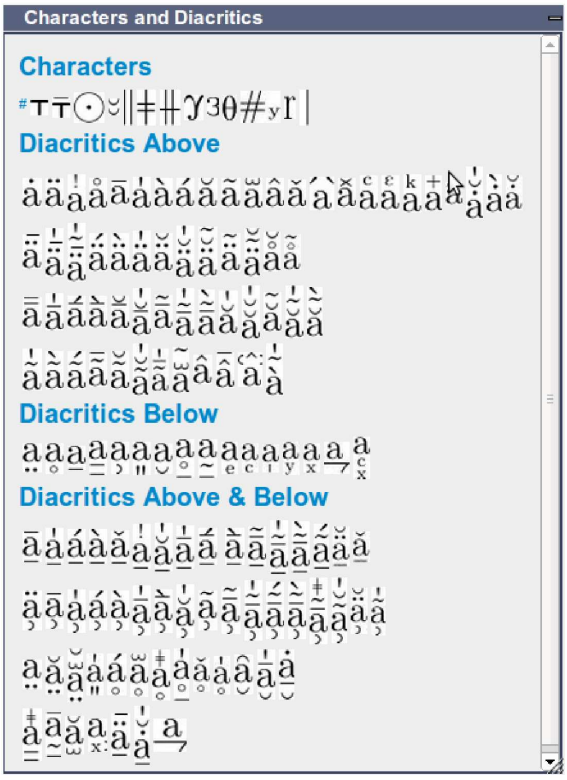

Fig. 4 Characters and diacritics panel

Once a volunteer completes transcribing a page, they would then click on the Finish and Exit button.

Further instructions on how to use the transcription tool were embedded above the transcription tool interface. Figure 5 shows the instructions, which are simple and short and emphasis is placed on the important points.

4.4 Transcription Interface

A simplistic design was used for the transcription tool interface, to cater for the varying volunteer skill levels. 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 5 below.

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, a zoom feature was included.

5 Evaluation

An evaluation of transcription accuracy was conducted by: checking the consistency of multiple transcriptions; comparing transcriptions to a known gold standard; and correlating consistency with accuracy.

5.1 Transcription Similarity Metric

The Levenshtein distance [9] 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 [17] note that, because of this, there is need for a normalized version of the method.

Notation-wise, Σ represents the alphabet, Σ^{Λ} is the set of strings in Σ and $\lambda \notin \Sigma$ denotes the null string. A string X $\in \Sigma^{\Lambda}$ 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:

$$LD(X,Y) = min\{\gamma(S_{x,y})\} \tag{1}$$

Yujian and Bo [17] define the normalized Levenshtein distance as a number within the range 0 and 1, where 0 means that the strings are different and 1 means that they are similar.

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

where $\alpha = \max\{\gamma(\text{a}\rightarrow \lambda), \gamma(\lambda \rightarrow b)\}$

5.2 Inter-transcriber Agreement

The normalized Levenshtein distance metric was used to measure transcription similarity or inter-transcriber agreement among users who have transcribed the same text. The inter-transcriber agreement can be used to

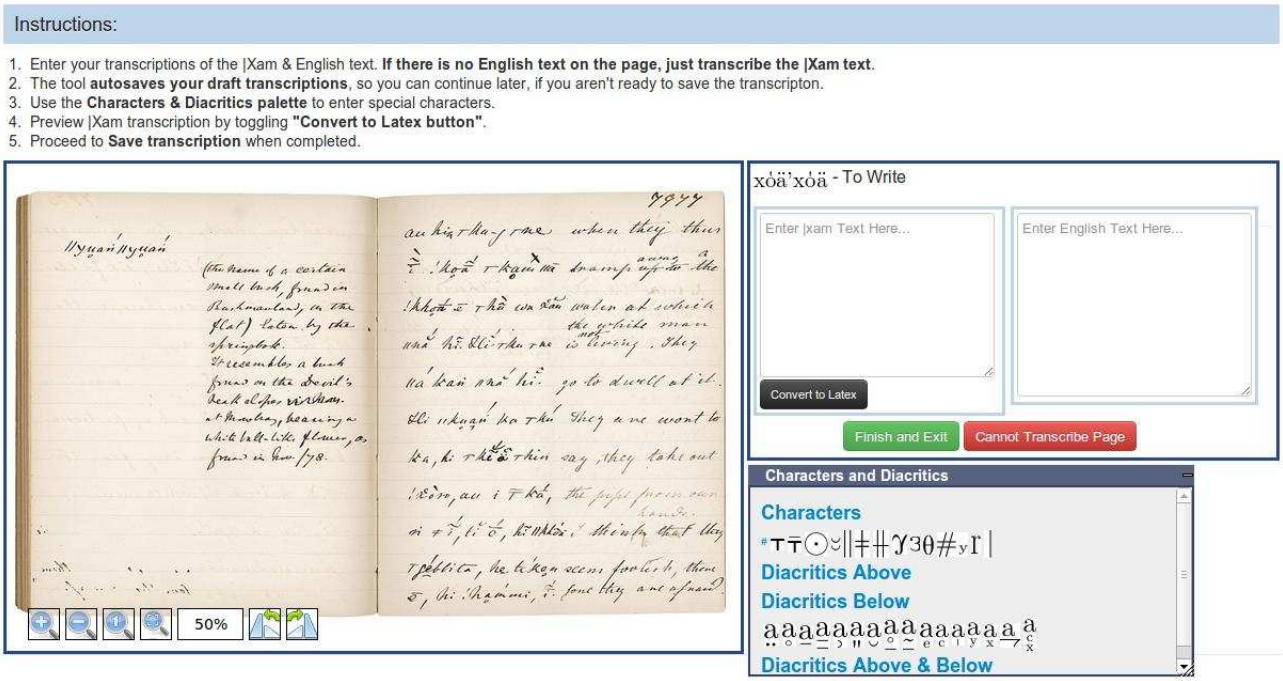

Fig. 5 Transcription system interface

assess reliability of the data from volunteers or consistency in the transcriptions.

Transcription similarity or inter-transcriber agreement is calculated at line level. The overall similarity among 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 similarity.

The minimum, average and maximum similarity values were calculated independently for the English and |Xam text.

5.2.1 English Text

Figure 6 is a plot of the minimum, average and maximum similarity for each transcription of English text. The blue, red and green data points represent the maximum, average and minimum values respectively. The transcriptions have been sorted on average similarity to clearly show clusters of similar values.

A total of 371 transcriptions were plotted in Figure 6. Single transcriptions or perfect correspondences are indicated by the convergence at an agreement value of 1. Approximately one third of the transcriptions (225-371) result in perfect agreement, while another one third (100-224) have at least 80% agreement. For higher levels of agreement, the variance in values is also low. For the lowest one third of the transcriptions (1-99), there is a higher variance but the appearance of many high maximum values suggest that 2 transcriptions have high agreement while the third is an outlier.

The results show that volunteers (non-experts) are able to produce English transcriptions that are reliable and consistent, with an overall similarity measure of $\mu = 0.95$ for all the transcriptions.

5.2.2 |Xam Text

Figure 7 is a plot of the minimum, average and maximum for each transcription of |Xam text. The blue, red and green data points represent the maximum, average and minimum values respectively. The transcriptions have been sorted on average similarity to clearly show clusters of similar values.

A total of 412 transcriptions were plotted in Figure 7. Single transcriptions or perfect correspondences are indicated by the convergence at an agreement value of 1, and only account for approximately 10% of the transcriptions. However, about 80% of transcriptions (80-412) have an agreement value of at least 75%. The variance is also relatively low and there are few transcriptions with small agreement values.

As before, the results show that volunteers (non-experts) are able to produce |Xam transcriptions that are reliable and consistent, with an overall similarity measure of $\mu = 0.80$ for all the transcriptions.

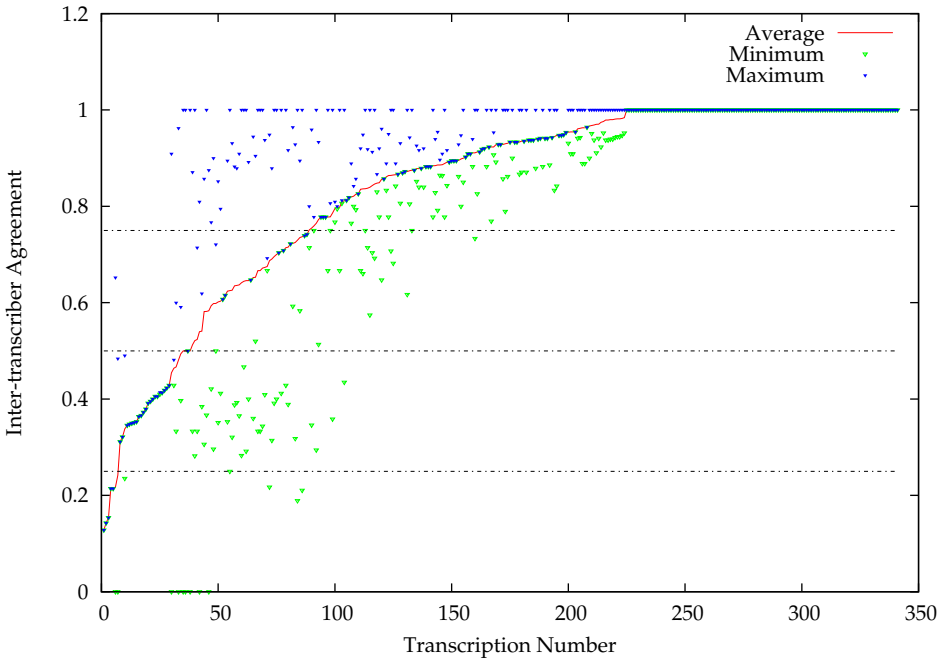

Fig. 6 Inter-transcriber similarity for English text

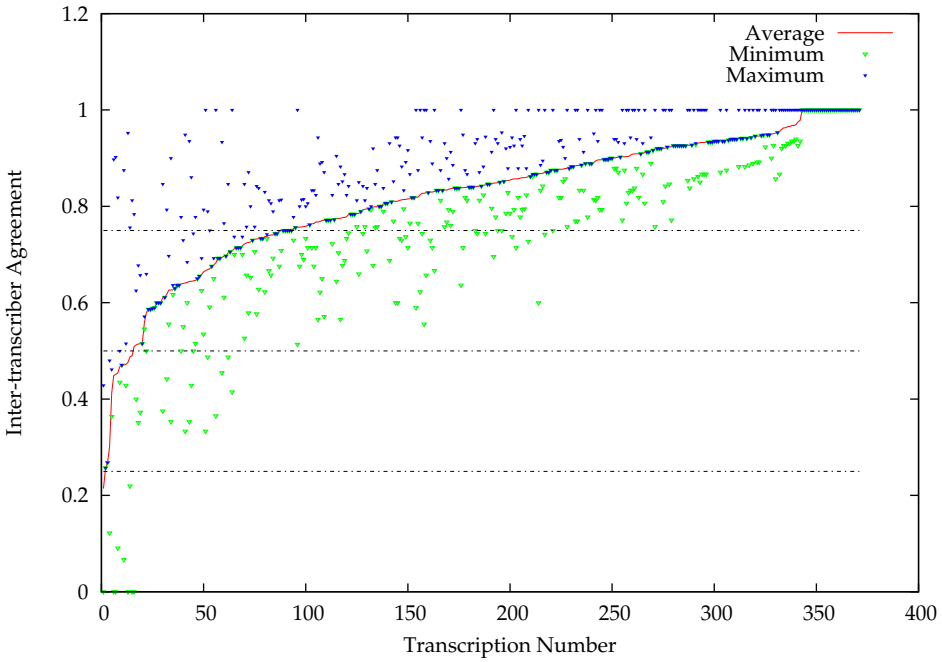

Fig. 7 Inter-transcriber similarity for |Xam text

5.3 Transcription Accuracy

In this experiment, the Bleek and Lloyd transcription gold standard (Corpus-G) [16] was used as a comparison for the transcriptions produced by the crowdsourced volunteers (Corpus-V). Transcription accuracy was measured by calculating the normalized Levenshtein distance between two strings. A total of 186 transcriptions were used.

Table 1 depicts the transcription accuracy distribution. 34.41% of the transcriptions have an average accuracy higher than 70%, while 40.86% have an accuracy between 51% and 69%. 14.51% of the transcriptions have an accuracy between 36% and 50%, and the re-

Table 1 Accuracy Distribution for Corpus-V with Corpus-G

Accuracy	Data Points	Percentage
0.70 - 1.00	64	34.41%
0.51 - 0.69	76	40.86%
0.36 - 0.50	27	14.51%
0.00 - 0.35	16	8.60%

maining 8.60% have an accuracy lower than 35%. The global average accuracy is 64.75%.

The average accuracy is therefore substantially higher than previous studies at line level and marginally higher than previous studies at word level. In addition, this accuracy was obtained on the basis of the "wisdom of the crowd" rather than highly tuned machine learning algorithms.

5.4 Correlation of Inter-transcriber Agreement and Accuracy

The final experiment considered whether intertranscriber agreement correlates with accuracy. Inter-transcriber agreement can be calculated mechanically during processing of tasks while accuracy can only be computed based on an existing gold standard. Thus, if there is a correlation, it suggests that inter-transcriber agreement could be used as an alternative metric to accuracy for non-training data.

Figure 8 is a box-and-whisker plot of the correlation, with agreement levels separated into 10 discrete bands. The graph shows clearly that there is a linear relationship between average inter-transcriber agreement and transcription accuracy. Thus, greater agreement among transcriptions of a line of text may translate to a higher level of accuracy and this could be exploited in the crowdsourcing application by, for example, injecting additional jobs into the queue if inter-transcriber agreement is low.

6 Conclusions

This article considered the feasibility of volunteer thinking for the transcription of historical manuscripts, with a focus on quality of transcriptions.

The experiments have demonstrated that: (a) transcriptions produced by volunteeers have a high degree of similarity, suggesting that the transcriptions are reliable and consistent; (b) the acccuracy of transcriptions produced by volunteeers is higher than that obtained in previous research; and (c) a high degree of consistency correlates with a high degree of accuracy.

Thus, it may be argued that is possible to produce high quality transcriptions of indigenous languages using volunteer thinking. Furthermore, this technique should be considered to complement or as an alternative approach for other heritage preservation tasks where the "wisdom of the crowd" may produce comparable or better results.

Future work related to transcription includes the use of language models for suggestion, correction and merging of transcriptions; and result merging to produce synthetically-derived transcriptions with potentially higher levels of accuracy.

Acknowledgements This research was partially funded by the National Research Foundation of South Africa (Grant numbers: 85470 and 83998), the Citizen Cyberscience Centre and University of Cape Town. The authors acknowledge that opinions, findings and conclusions or recommendations expressed in this publication are that of the authors, and that the NRF accepts no liability whatsoever in this regard.

References

1. Anderson, David P., Jeff Cobb, Eric Korpela, Matt Lebofsky and Dan Werthimer. SETI@home: An Experiment in Public-Resource Computing. Communications of the ACM, Vol. 45 No. 11, November 2002, pp. 56-61.
2. Bossa. http://boinc.berkeley.edu/trac/wiki/bossaintro.
3. Callison-Burch, Chris. Fast, cheap, and creative: evaluating translation quality using amazons mechanical turk. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, EMNLP '09, pages 286-295, Stroudsburg, PA, USA (2009) Association for Computational Linguistics.
4. Catlin-Groves, Christina L. The Citizen Science Landscape: From Volunteers to Citizen Sensors and Beyond, International Journal of Zoology, Vol. 2012, Article ID 349630, 14 pages (2012) doi:10.1155/2012/349630
5. Causer, Tim, and Valerie Wallace. Building a volunteer community: results and findings from Transcribe Bentham, Digital Humanities Quarterly, Vol. 6, No. 2 (2012)
6. Kanefsky, B., 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)
7. Lee, J. H. Crowdsourcing music similarity judgments using mechanical turk. Proc. of ISMIR 2010, pages 183188 (2010)
8. Lee, Jin Ha, 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 129138, New York, NY, USA, ACM (2012)
9. Levenshtein, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8):707-710 (1966)
10. Marge, Matthew, 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 Mechan-

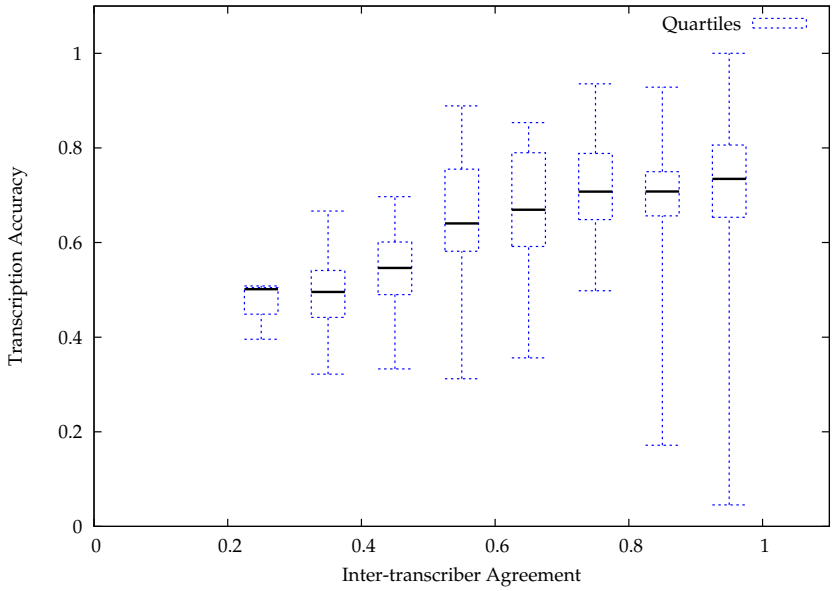

Fig. 8 Correlation between inter-transcriber similarity and accuracy

ical Turk, CSLDAMT '10, pages 99107, Stroudsburg, PA, USA, Association for Computational Linguistics (2010)
finding of usability problems. In Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, CHI '93, pages 206213, New York, NY, USA, ACM (1993)
11. Nowak, Stefanie 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, ACM (2010)
12. Shachaf, P. The paradox of expertise: Is the wikipedia reference desk as good as your library? Journal of Documentation, 65(6):977-996 (2009)
13. Suleman, H. Digital libraries without databases: The Bleek and Lloyd collection. Research and Advanced Technology for Digital Libraries, pages 392-403 (2007)
14. Von Ahn, L., Benjamin Maurer, Colin McMillen, David Abraham and Manuel Blum. RECAPTCHA: Human-based character recognition via web security measures. Science, 321:1465-1468 (2008)
15. Williams, Kyle. Learning to Read Bushman: Automatic Handwriting Recognition for Bushman Languages. MSc, Department of Computer Science, University of Cape Town (2012)
16. Williams, Kyle 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, Springer-Verlag (2011)
17. Yujian, Li and Liu Bo. A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6):1091-1095, June (2007)