Quality Assessment in Crowdsourced Indigenous Language Transcription

Ngoni Munyaradzi and Hussein Suleman

Department of Computer Science, University of Cape Town,
Cape Town, South Africa
{ngoni.munyaradzi@uct.ac.za,hussein@cs.uct.ac.za}

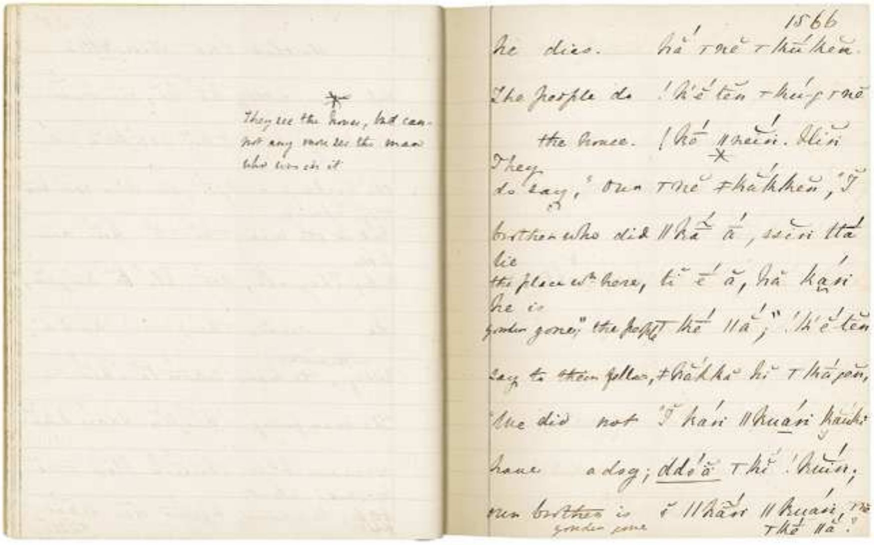
Abstract. The digital Bleek and Lloyd Collection is a rare collection that contains artwork, notebooks and dictionaries of the indigenous people of Southern Africa. The notebooks, in particular, contain stories that encode the language, culture and beliefs of these people, handwritten in now-extinct 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. In this paper, a crowdsourcing method is proposed to transcribe the manuscripts, where non-expert volunteers transcribe pages of the notebooks using an online tool. Experiments 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

1 Introduction

The digital Bleek and Lloyd Collection [10] 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 page 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



*
They see the house, but cannot any more see the man who lives in it

1866
he dies. há ⊤ne ⊤kúken.
The people do !k'é tẽn ⊤kúi ⊤ne
the house. !kó ǁnein. ǁin
They do say, "our ⊤ne ǂkakken, "I
brother who did ǁká ā, sein ǁā
lie the place wh. here, ti ē ā, há kań
he is yonder gone," the people kē ǁā;" !k'é tẽn
say to their fellow, ǂkákka ki ⊤kábbu,
"We did not "I kań ǁkuań kaúki
have a dog; dde ā ⊤kē !kuin,
our brother is yonder gone ĩ ǁkań ǁkuań, "ne ⊤kē ǁā."

Fig. 1. Sample page from Bleek and Lloyd notebooks

processing that are currently not possible. Manual translation 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., 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 paper investigates the feasibility and accuracy of volunteer transcription, as one example of an intellectually-intensive tasks in digital libraries, and how it compares to computational techniques like machine learning.

The rest of this paper 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; Section 4 focuses on the analysis of the initial results; and Section 5 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 [9] 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 [8] 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% [11]. 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 paper investigates what level of quality can be achieved solely by volunteers and automated post-processing techniques.

Williams [12] 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 [13], which was used in the work reported on in this paper.

In summary, there have been numerous attempts at transcription, with a focus on the mechanics of the process. This paper, instead, focuses on the assessment of transcription accuracy, which is further in the context of a language

[1] http://www.google.com/recaptcha

that is unfamiliar to volunteers. The mechanics were greatly simplified by use of the Bossa toolkit, as discussed in the next section.

3 Bossa Framework

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 the 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.

The Transcribe Bleek and Lloyd project used a Web application that defined each page of text to be transcribed as a single task. Volunteers were presented with a Web interface where the original text was displayed and they were asked to enter their transcriptions, with special characters and diacritics entered using a visual palette. This palette-oriented editing interface was adapted from earlier work by Williams [12]. Figure 2 shows the transcription interface.

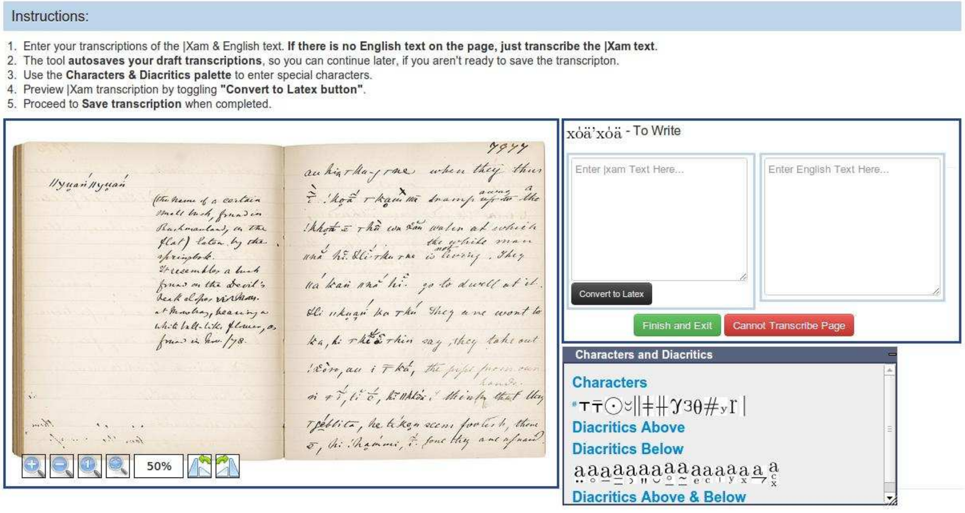
4 Evaluation

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.

4.1 Transcription Similarity Metric

The Levenshtein distance [7] 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

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



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.

Fig. 2. Bossa-based interface for transcription of pages

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 [14] 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 $= x_1 x_2 ... x_n$, 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 $x_1 x_2 ... x_n$ or $\mid X \mid$. A substitution operation is represented by $a \rightarrow b$, insertion by $\lambda \rightarrow a$ and deletion by $b \rightarrow \lambda$. $S_{x,y} = S_1 S_2 ... S_u$ 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:

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

Yujian and Bo [14] 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 $\alpha = \max\{\gamma(\text{a} \rightarrow \lambda),\, \gamma(\lambda \rightarrow b)\}$

4.2 Inter-transcriber Agreement

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

English Text Figure 3 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.

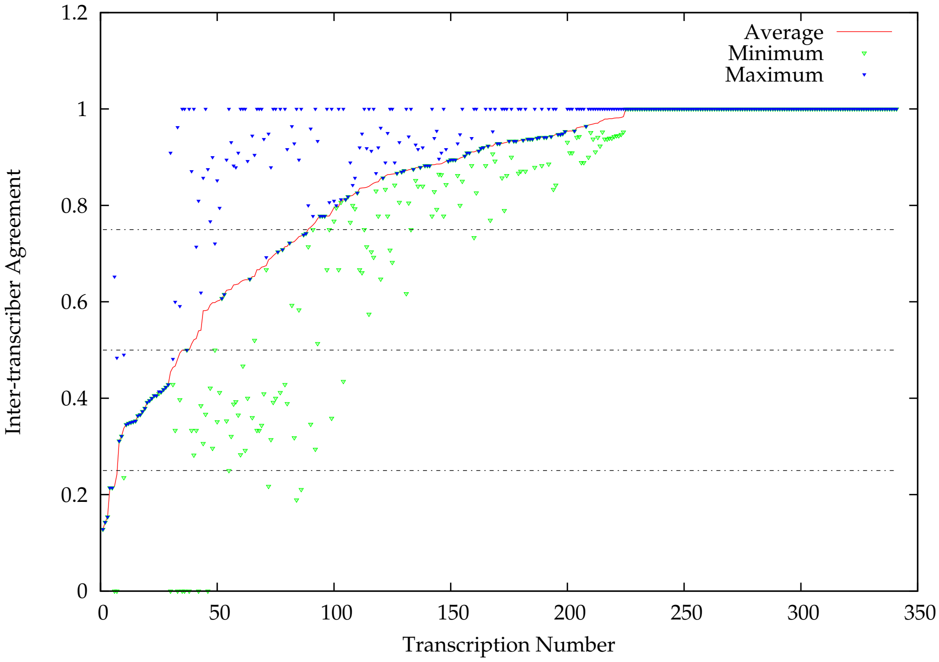


Fig. 3. Inter-transcriber similarity for English text

A total of 371 transcriptions were plotted in Figure 3. 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.

|Xam Text Figure 4 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.

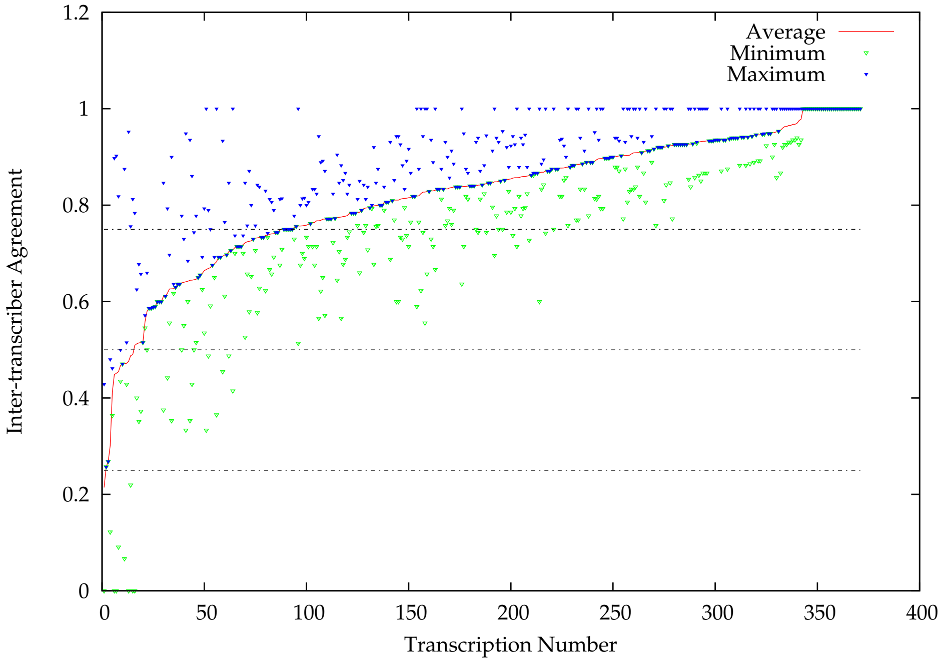


Fig. 4. Inter-transcriber similarity for |Xam text

A total of 412 transcriptions were plotted in Figure 4. 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.

4.3 Transcription Accuracy

In this experiment, the Bleek and Lloyd transcription gold standard (Corpus-G) [13] was used as a comparison for the transcriptions produced by the crowd-sourced 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 remaining 8.60% have an accuracy lower than 35%. The global average accuracy is 64.75%.

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

Accuracy	*DataPoints*	*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%

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 optimized algorithms.

Correlation of Inter-transcriber Agreement and Accuracy The final experiment considered whether inter-transcriber 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 5 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

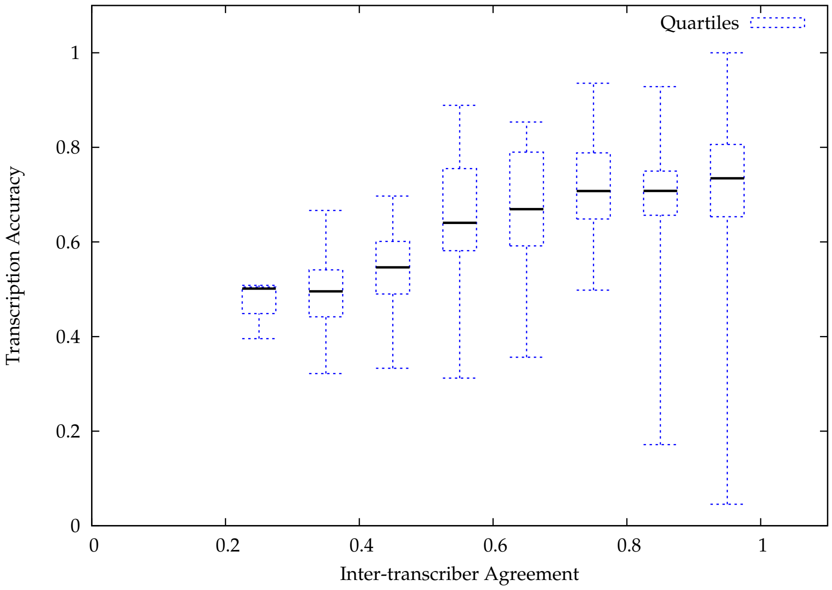


Fig. 5. Correlation between inter-transcriber similarity and accuracy

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.

5 Conclusions

This paper 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.

6 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. Levenshtein, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8):707-710, 1966.
8. 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, 2010. ACM.
9. Shachaf, P. The paradox of expertise: Is the wikipedia reference desk as good as your library? Journal of Documentation, 65(6):977-996, 2009.
10. Suleman, H. Digital libraries without databases: The Bleek and Lloyd collection. Research and Advanced Technology for Digital Libraries, pages 392-403, 2007.
11. 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.
12. Williams, Kyle. Learning to Read Bushman: Automatic Handwriting Recognition for Bushman Languages. MSc, Department of Computer Science, University of Cape Town, 2012.
13. 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, 2011. Springer-Verlag.

14. Yujian, Li and Liu Bo. A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6):1091-1095, June 2007.