Crowdsourcing a Text Corpus is not a Game

Sean Packham and Hussein Suleman

Centre for ICT4D, Department of Computer Science,
University of Cape Town, South Africa
{pcksea001@uct.ac.za, hussein@cs.uct.ac.za}

Abstract. Building language corpora for low resource languages such as South Africa's isiXhosa is challenging because of limited digitized texts. Language corpora are needed for building information retrieval services such as search and translation and to support further online content creation. A novel solution was proposed to source original and relevant multilingual content by crowdsourcing translations via an online competitive game where participants would be paid for their contributions. Four experiments were conducted and the results support the idea that gamification by itself does not yield the widely expected benefits of increased motivation and engagement. We found that people do not volunteer without financial incentives, the form of payment does not matter, they would not continue contributing if the money is taken away and people preferred direct incentives and the possibility of incentives was not as strong a motivator.

Keywords: Crowdsourcing · Gamification · Translation · Language corpora· Information retrieval

1 Introduction

isiXhosa (Xhosa) is the second most spoken first language in South Africa, spoken by more than 8 million people - 16% of the country's population [1]. isiXhosa is categorised as a low resource language with a scarcity of digital content and well defined linguistic models and tools [2]. isiXhosa is a morphologically rich and a highly agglutinative language, forming words by gluing together part to a word's base form [3]. For example the base form of the isiXhosa word for month is "inyanga", gluing "i" in front produces the plural form "iinyanga". Developing automatic translation systems for agglutinative languages with few morphological models is particularly challenging because the base form of words are often incorrectly categorised [4]. Low resource languages are further challenged by the difficulty of assembling sufficient content to build language corpora, a problem worsened when trying to assemble multilingual language corpora. Attempts to assemble monolingual and multilingual isiXhosa corpora from South African governmental websites [2], [4] or by crawling isiXhosa specific websites [5] found that the quantity and quality of the content was not sufficient to produce a working machine translation system [4,5].

A gamified crowdsourcing system was proposed as a novel approach to affordably gather original multilingual content for building language corpora for low resource languages. A custom crowdsourcing system was created and evolved over four experiments. The aim of the experiments was to investigate if intrinsic motivation or gamified motivation could influence users to perform a clearly important social task, with monetary payments being only secondary. Thus, 2 of the experiments appealed to the users based on the intrinsic value of the task. The other 2 experiments offered payments, but these were gamified to test whether the game element appealed to users more than the financial reward. Additional motivation factors, such as physical rewards and user feedback, were not considered for this study. Furthermore, motivation factors for sustained participation weren't explored [6] because of the short duration of the experiments.

The rest of this paper describes these experiments and their results, preceded by a discussion of related literature.

2 Crowdsourcing

Crowdsourcing is the process of outsourcing tasks normally done by an employee or contractor to an anonymous crowd [7]. Crowdsourcing can be successful on projects that can be subdivided into small repeatable Human Intelligence Tasks (HITs), which are challenging for computers to perform but can be performed by a human in a reasonable amount of time [8]. Zaidan and Burch used crowdsourcing to produce Urdu to English translations where the quality was near professional levels by using redundant translations, translation edits and translator screening to automatically select the best translations [9]. Crowdsourcing has also been used for emergency response after the Haiti earthquake in 2010 to translate more than 40,000 emergency messages over six days from Haitan Kreyol to various languages [10]. A systematic classification of 46 crowdsourcing projects identified motivation via remuneration and quality control via a pre-qualification assessment to be prominent and important characteristics of successful crowdsourcing projects [11].

Table 1 shows the results of a literature survey that was conducted to sample the reward amounts for translation HITs on crowdsourcing platforms such as Mechanical Turk and CrowdFlower. The survey uncovered payment points for both translating and ranking tasks. A few studies specified payment points per task rather than per word and where possible a translation word cost was calculated. The survey shows that it was normal to find translation jobs between 2009 and 2014 that offered rewards between $0.01 and $0.25 to translate/edit a sentence.

Crowdsourcing marketplaces such as Mechanical Turk (MTurk) or CrowdFlower offer a crowdsourcing platform and access to a large number of users. A sampling of MTurk users revealed that 85% were from the United States and India and the remaining 15% were scattered across the rest of the world [8], therefore alternative means of specifically gathering bilingual English-isiXhosa speakers were investigated.

Table 1. Rewards offered by various crowdsourcing translation studies

Source	Task Detail	Reward	/Word
[9]	Translate Urdu to English	$0.10	$0.005
	Edit 10 sentence	$0.25	
	Rank 4 translation groups	$0.06	
[12]	Translate English to Spanish		$0.01
	Validate translation		$0.002
[13]	Rank 5 German to English machine translations	$0.01	
	Translate German to English	$0.10	
	Detect if a machine translation	$0.006	
[14]	Translate Spanish to English	$0.01	
	Translate Teluga to English	$0.02	
	Translate English to Creole	$0.06	
	Translate Urdu to English	$0.03	
	Translate Hindi to English	$0.03	
	Translate Chinese to English	$0.02	

2.1 Gamification

Gamification is the process of using gaming elements in a non-gaming context to improve user experience and motivation [15]. Rewarding a person with virtual points [16,17,18,19] and badges or achievements [19,20,21,22,23] for completing tasks are all examples of gamification. Like many games, gamification can be implemented as a competitive system where users compete against others for placement on a leaderboard [16], [18], [22], [24].

3 Methodology

The four experiments allowed users to translate English sentences from Wikipedia articles on South African topics on a custom created online crowdsourcing website. Each sentence needed to be translated by three separate users and the translations ranked in order of correctness by another three separate users. The ranks for each translation were totalled and the translation with the lowest score selected as the model translation. For example if all the users agreed and ranked the same translation first, that translation will be the model answer because it will have the lowest total ranking of 3 = 1 + 1 + 1. Users were rewarded with points for each contribution and their total score was reflected on a leaderboard.

Experiment 1 was conducted during the early stages of the research as a pilot project to find out if participants could be gathered from Twitter, a social network for sharing short messages called tweets to followers, and also to prototype a custom crowdsourcing system with scoring, leaderboards and support for paying participants.

The design of experiment 2 was inspired by games that offer increasing rewards from increasing effort over time. For the purpose of comparison, accompanying schemes that offered consistent and decreasing rewards from increasing effort and constant effort were designed. Using the surveyed rewards from past studies and a sampling of professional translation rates, a payment model was developed to select translation and ranking rewards for all the payment groups. The model took into account task redundancy and the national minimum wage for workers with a secondary school education.

The scoring system was designed to have one to one mapping to money earned - each point was equivalent to ZAR0.01. Users were rewarded with points for translating and ranking and the number of points awarded depended on which group they were in. Each group had its own leaderboard. All the payment schemes for the 6 groups were designed with a cap of 100 translations and 100 rankings. Setting a cap allowed predictable payment values to be calculated for each group. Task payment points were first chosen for the groups in the constant set and adjusted appropriately for the increasing set and decreasing set. The groups in the increasing set were adjusted to start at a lower rate and end at a higher rate. The groups in the decreasing set were adjusted to start at a higher rate and end at a lower rate. All the groups had the same average payment per task if the cap was reached. This design created a predictable reward system where rewards could not spiral out of control or become meaningless if no cap existed. All the payment groups were balanced so that users in either could earn the same amount if they reached the cap, with an average reward per sentence translated of $0.06 and reward per sentence ranked of $0.03, putting it in the range of the surveyed rewards in Table 1. The reward amounts for the increasing and decreasing groups differed by 100% at the start and end of the task limit. The selected articles had an average sentence length of 22 words, which resulted in total cost of translating one sentence, including ranking and duplication to be 5-30 times cheaper than sampled local and international translation services.

In South Africa the smallest bank note available is a R10 note. Therefore a participant's money earned was rounded down to the nearest R10; for example a score of 9100 would round down to R90.00. Paying users with cash was not an option because of the large number of expected users. It was decided to use mobile wallet and cardless transaction services offered by many of South Africa's large banks. To send money, the sender deposits cash or selects an account to pay from, and provides the recipient's mobile number. On payment, the recipient gets an SMS detailing the transaction and instructions on how to withdraw the money. The money can be withdrawn from one of the sending bank's branches, cash machines or from a list of authorized partners.

The third experiment replaced the multiple payment groups with a single group, and tested whether the same students from the University of Cape Town would contribute without any financial reward. Users were awarded 1 point for translating or ranking and a single leaderboard was used. Translation and ranking caps were removed, as there was no budget that could be exhausted.

The final experiment tested whether paying users based on where they placed on the contributions leaderboard rather than per contribution would be better at motivating users to contribute more and produce more affordable rates than experiment 2. An increasing reward for increased effort approach was adopted when choosing the payment points, which resulted in the rewards seen in Fig 1 for the users who contributed the most translations and rankings. Only the top 40 positions were allocated a reward. The experiment was designed to create a sense of heightened competition between users by having them focus on the marginal difference between their contributions and the next user's contributions. Experiment 4 allocated double the budget used in experiment 2 to rewards to further motivate participants.

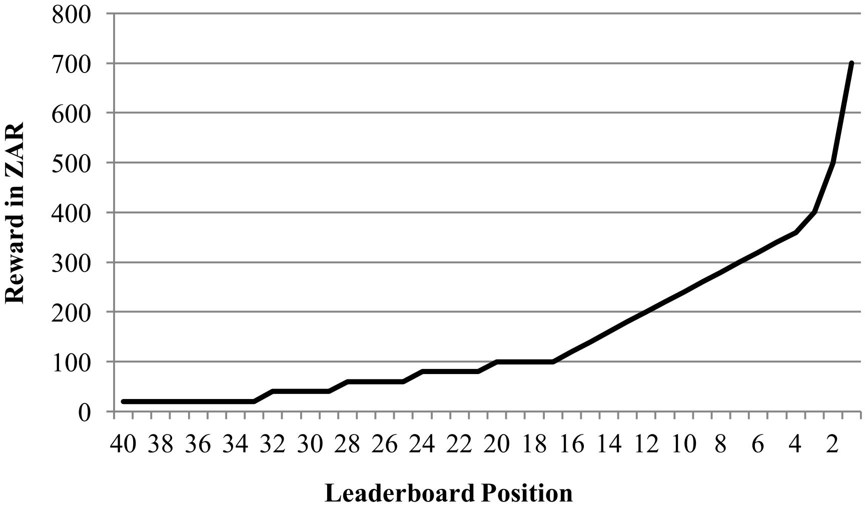

Fig. 1. Leaderboard payment scheme used in experiment 4

4 Results

Experiment 1 was run over three days from 5 August 2014. Five tweets were sent to the author's 132 followers who then shared the project four times; by the end the experiment website was visited 10 times but no one contributed any translations. There are a number of expected reasons for why people did not contribute: the author's network and extended network were not reaching isiXhosa speakers or they were not willing to contribute for free.

Experiment 2 was run for a week from 19 November 2014 (after final exams) with volunteering students from the University of Cape Town. Approximately 24,000 students were sent a "call for participants" email, 200 signed up to participate, 121 made at least one contribution and 61 users contributed enough to receive a reward. 3600 individual translations and 2589 individual rankings were contributed. 1088 sentences received 3 translations and 734 sentences received 3 rankings and could be reassembled into isiXhosa articles. The total cost of the experiment was ZAR3020.

An analysis of the translation times over the duration of the experiment showed no noticeable trends but this was due to the varying sentence length and low number of contributions in each group. Ranking times for all payment groups exhibited a similar downward trend over the duration of the experiment. This showed that the length of a sentence and its translations did not affect the time it took a user to rank it. Furthermore the different payment groups did not affect the users' motivation to rank faster.

Fig 2 shows that only 3 users reached both the translation and ranking limit and earned the maximum reward of ZAR100.00. A large percentage of users who did earn money contributed only enough to earn the first reward of ZAR10.00. 70% of the users did not contribute enough to qualify for any reward, showing that the incentive of payment was enough to motivate users to sign up but not enough to get them to contribute.

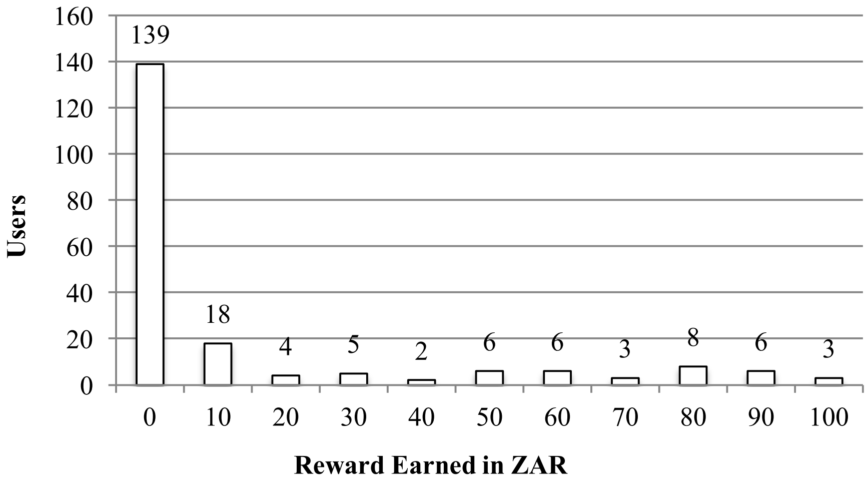

Fig 2. Number of users per reward tier

Experiment 3 was run at the start of the 2015 academic year in February. 47 users registered and only 12 made at least one contribution. The activity of the users was considerably lower than that of Experiment 2: the most active user contributed 11 translations and 2 rankings. Only 11 sentences were translated 3 times and 2 sentences were ranked 3 times. Offering a monetary reward was considerably more successful at attracting and engaging participants.

Experiment 2 showed that only 61 out of the 200 users earned money therefore experiment 4 was designed to offer rewards to only the most active users. Experiment 4 received 147 users, 57 users contributed 1865 individual translations and 1767 rankings. 39% of users in experiment 2 and 61% of users in experiment 4 made at least one contribution, a considerable difference in activity. 617 sentences received 3 translations and 584 sentences received 3 rankings. Due to the lower activity and the pre-chosen budget, experiment 4 achieved a translation cost of ZAR0.22 per word, almost double the rate of experiment 2.

Table 2. Sample of the experiment 4 leaderboard

Rank	Contributions	Reward (ZAR)	Value vs. Experiment 2
1	444	700	0,32
2	401	500	0,40
3	372	400	0,47
4	284	360	0,39
5	259	340	0,38
6	245	320	0,38
7	192	300	0,32
8	156	280	0,28
9	143	260	0,28
10	133	240	0,28
36	8	20	0,20
37	8	20	0,20
38	7	20	0,18
39	7	20	0,18
40	5	20	0,13

Table 2 shows a sample of the activity of the top 10 and bottom 5 money earners in experiment 4. The second column shows how many contributions it took to reach the respective leaderboard position and reward and the fourth column shows how the translation cost the user achieved compared to that of experiment 2 in terms of value. No user achieved a translation cost equivalent to even half that of experiment 2. The value was worse at the bottom of the paid leaderboard but steadily improved, as users were more competitive higher up the leaderboard.

It would be interesting to know if users feel more comfortable to contribute in smaller groups, like those in experiment 2, which had on average 33 users, rather than a larger group like experiment 4. Users may feel they have a greater chance at reaching the top leaderboard position when there are fewer competitors.

5 Conclusion

Employing gamification in a crowdsourcing game to translate English to isiXhosa showed that people do not volunteer without payment, the form of payment does not matter, participants would not contribute if payment is taken away and finally people wanted a guaranteed rate and the possibility of incentives is not as strong a motivator.

The guaranteed rates offered by the various payment groups in experiment 2 were considerably more effective at getting participants to contribute than the leaderboard payment scheme of experiment 4. This was an interesting result as it was expected that linking payments to leaderboard positions would create a greater competitive environment but it may have had the reverse effect and scared off users who were late to join or slow to start.

The over-arching hypothesis of this project was that gamification of a crowdsourcing system with a task with strong intrinsic motivation would make it possible to gather important data with payment being a secondary factor rather than a primary one. The various experiments have illustrated that this is indeed not true. The student users were purposefully chosen to have a higher than average level of education and to not have a desperate need for the small amounts of money paid. Ultimately, the experiments have illustrated that monetary payment is still a stronger motivation factor than intrinsic motivation or motivation because of gamification. While these results were obtained with a specific task in a specific part of the world, the fundamental lessons learnt are likely to be applicable to corpus generation projects elsewhere.

6 Future Work

Additional motivation factors such as physical rewards can be assessed and compared to financial rewards and various gamification factors. Furthermore an analysis of similar experiments conducted in low resource environments can be performed.

A deeper analysis of user behaviour will be performed by examining translation and ranking durations across groups. Expert users will be used to assess the quality of the crowdsourced data. The developed system and techniques will be improved and used to gather further data for the development of isiXhosa language processing algorithms and tools.

Acknowledgements. This research was partially funded by the National Research Foundation of South Africa (Grant numbers: 85470 and 88209) 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. Statistics South Africa: Census 2011 Census in Brief. Statistics South Africa, Pretoria (2012)
2. Eiselen, E., Puttkammer, M.: Developing Text Resources for Ten South African Languages. In: Proceedings of the LREC (2014)
3. Webb, V.N.: African Voices: An Introduction to the Languages and Linguistics of Africa. Oxford University Press (2000)

4. Johnson, K.K.: Xhosa-English Machine Translation: Working with a Low-Resource Language (2011)
5. Drummer, A.: Phrase-Based Machine Translation of Under-Resourced Languages (2013)
6. Jackson, C.B., Osterlund, C., Mugar, G., Hassman, K.D., Crowston, K.: Motivations for Sustained Participation in Crowdsourcing: Case Studies of Citizen Science on the Role of Talk. In: 2015 48th Hawaii International Conference on System Sciences (HICSS), pp. 1624–1634 (2015)
7. Howe, J. Crowdsourcing: How the Power of the Crowd is Driving the Future of Business. Random House (2008)
8. Ross, J., Irani, L., Silberman, M., Zaldivar, A., Tomlinson, B.: Who are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In: CHI'10 Extended Abstracts on Human Factors in Computing Systems, pp. 2863–2872 (2010)
9. Zaidan, O.F., Callison-Burch, C.: 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, pp. 1220-1229 (2011)
10. Munro, R.: Crowdsourced Translation for Emergency Response in Haiti: The Global Collaboration of Local Knowledge. In: AMTA Workshop on Collaborative Crowdsourcing for Translation (2010)
11. Geiger, D., Seedorf, S., Schulze, T., Nickerson, R.C., Schader, M.: Managing the crowd: towards a taxonomy of crowdsourcing processes (2011)
12. Negri, M., Mehdad, Y.: 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, pp. 212–216 (2010)
13. Callison-Burch, C.: 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, pp. 286–295 (2009)
14. Ambati, V., Vogel, S.: Can Crowds Build Parallel Corpora for Machine Translation Systems? In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 62–65 (2010)
15. Deterding, S., Sicart, M., Nacke, L., O'Hara, K., Dixon, D.: Gamification. Using Game-Design Elements in Non-gaming Contexts. In: CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 2425–2428 (2011)
16. Eickhoff, C., Harris, C.G., de Vries, A.P., Srinivasan, P.: Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 871–880 (2012)
17. Farzan, R., DiMicco, J.M., Millen, D.R., Brownholtz, B., Geyer, W., Dugan, C.: When the Experiment is Over: Deploying an Incentive System to All the Users. In: Proceedings of the Symposium on Persuasive Technology, in conjunction with the AISB (2008)
18. Farzan, R., DiMicco, J.M., Millen, D.R., Dugan, C., Geyer, W., Brownholtz, E.A.: Results from Deploying a Participation Incentive Mechanism within the Enterprise. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 563–572 (2008)

19. Montola, M., Nummenmaa, T., Lucero, A., Boberg, M., Korhonen, H.: Applying Game Achievement Systems to Enhance User Experience in a Photo Sharing Service. In: Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era, pp. 94–97 (2009)
20. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Steering User Behavior with Badges. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 95–106 (2013)
21. Denny, P.: The Effect of Virtual Achievements on Student Engagement. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 763–772 (2013)
22. Dominguez, A., Saenz-de-Navarrete, J., De-Marcos, L., Fernández-Sanz, L., Pagés, C., Martínez-Herráiz, J-J.: Gamifying Learning Experiences: Practical Implications and Outcomes. Comput. Educ. 63, 380–392 (2013)
23. Fitz-Walter, Z., Tjondronegoro, D., Wyeth, P.: Orientation Passport: Using Gamification to Engage University Students. In: Proceedings of the 23rd Australian Computer-Human Interaction Conference, pp. 122-125 (2011)
24. Halan, S., Rossen, B., Cendan, J., Lok, B.: High Score!-Motivation Strategies for User Participation in Virtual Human Development. In: Intelligent Virtual Agents, pp. 482-488 (2010)