Crowdsourcing a Text Corpus for a Low Resource Language

Sean Packham

Supervisor:

Professor Hussein Suleman

Submitted for the Degree of Master of Science

In the Department of Computer Science

University of Cape Town

February 2015

Plagiarism Declaration

I, Sean Packham, declare that I know the meaning of plagiarism and that all of the works in the dissertation save for that, which is properly acknowledged, is my own. The Chicago (Author – Date) citation style has been used.

Signed by candidate

Signature Removed

Sean Packham

License

Abstract

Low resourced languages, such as South Africa's isiXhosa, have a limited number of digitised texts, making it challenging to build language corpora and the information retrieval services, such as search and translation, that depend on them. Researchers have been unable to assemble isiXhosa corpora of sufficient size and quality to produce working machine translation systems and it has been acknowledged that there is little to know training data and sourcing translations from professionals can be a costly process. A crowdsourcing translation game which paid participants for their contributions was proposed as a solution to source original and relevant parallel corpora for low resource languages such as isiXhosa.

The objectives of this dissertation is to report on the four experiments that were conducted to assess user motivation and contribution quantity under various scenarios using the developed crowdsourcing translation game. The first experiment was a pilot study to test a custom built system and to find out if social network users would volunteer to participate in a translation game for free. The second experiment tested multiple payment schemes with users from the University of Cape Town. The schemes rewarded users with consistent, increasing or decreasing amounts for subsequent contributions. Experiment 3 tested whether the same users from Experiment 2 would continue contributing if payments were taken away. The last experiment tested a payment scheme that did not offer a direct and guaranteed reward. Users were paid based on their leaderboard placement and only a limited number of the top leaderboard spots were allocated rewards.

From experiment 1 and 3 we found that people do not volunteer without financial incentives, experiment 2 and 4 showed that people want increased rewards when putting in increased effort, experiment 3 also showed that people will not continue contributing if the financial incentives are taken away and experiment 4 also showed that the possibility of incentives is as attractive as offering guaranteed incentives.

Acknowledgements

I would like to thank my supervisor for his guidance and patience. I thoroughly enjoyed our discussions and brainstorming sessions and learnt so much from your enquiring mind.

To my wife Bianca, thank you for being by my side and supporting me through many late nights.

To my fellow lab partners, thank you for all the great discussions over our shared coffees.

Table of Contents

Plagiarism Declaration i

License ii

Acknowledgements iii

Abstract iv

Table of Contents v

List of Figures viii

List of Tables x

1. Introduction 1

1.1. Research Questions 2

1.2. Thesis Outline 3

2. Literature Review 5

2.1. isiXhosa 5

2.1.1. Language Corpora 6

2.1.1.1. Information Retrieval 8

2.1.1.2. Machine Translation 9

2.1.2. Summary 10

2.2. Crowdsourcing 11

2.2.1. Translations and Parallel Corpora 12

2.2.2. The Cost of Crowdsourcing 14

2.3. Gamification 17

2.4. Summary 20

3. System Design 21

3.1. System Requirements 21

3.2. Open Source Crowdsourcing Systems 21

3.3. Custom System 22

3.3.1. Architecture 23

3.3.2. System Configuration 25

3.3.3. Application Configuration 25

3.3.4. Data Model 26

3.3.4.1. Users 26

3.3.4.2. Content and Contributions 27

3.3.5. Experiment 1: Pilot Study 29

3.3.6. Experiments 2, 3 and 4 31

3.4. Mobile Payments 35

3.5. Summary 37

4. Experiment 1: Pilot Study 39
4.1. Methodology *39*
4.2. Dataset *40*
4.3. Results *40*
4.4. Summary *41*
5. Experiment 2: Comparing rewards 42
5.1. Methodology *42*
5.1.1. Groups 42
5.1.2. Users 44
5.1.3. Translating 44
5.1.4. Ranking 45
5.1.5. Qualifying 46
5.1.6. Payment Model 46
5.1.7. Final Rewards 50
5.1.8. Dataset 53
5.1.9. Summary 53
5.2. Results *54*
5.2.1. Pre-processing Data 56
5.2.2. Activity 60
5.2.3. Analysis 65
5.2.3.1. Comparison of All Groups 65
5.2.3.2. Comparison of Work Effort Groups 69
5.2.3.3. Comparison of Reward Groups 72
5.2.3.4. Average Active User Contributions 75
5.3. Summary *76*
6. Experiment 3: Removing rewards 78
6.1. Methodology *78*
6.1.1. Users 78
6.1.2. Dataset 79
6.2. Results *79*
6.3. Summary *79*
7. Experiment 4: Leaderboard rewards 80
7.1. Methodology *80*
7.1.1. Users 80
7.1.2. Payment Model 80
7.1.3. Dataset 81
7.1.4. Results *81*
7.1.5. Pre-processing Data 82
7.1.6. Activity 84
7.1.7. Analysis 88
7.2. Summary *91*

8. Conclusions 92
8.1. Research Questions *93*
8.1.1. Question 1 93
8.1.2. Question 2 93
8.1.3. Question 3 93
8.2. Contributions and implications. *94*
9. Limitations and future work 95
9.1. Further explore sourcing participants from social networks *95*
9.2. Further develop gamified payment schemes *95*
9.3. Further explore non-financial rewards *95*
9.4. Repeat the experiments in other countries *96*
9.5. Crowdfund crowdsourcing *96*
9.6. Further develop IR algorithms and tools *96*
Bibliography 97
Appendices 102
Appendix A Experiment 2: Call for participants email UCT *102*
Appendix B Experiment 2: Online registration form *103*
Appendix C Experiment 3: Call for participants email UCT *104*
Appendix D Experiment 4: Call for participants email UCT *105*
Appendix E Experiment dataset *106*
Appendix F Experiment 4: Leaderboard payment scheme *107*

List of Figures

Figure 1: At home isiXhosa speakers in South African (Htonl 2013; Statistics South Africa 2012)..... 5

Figure 2: The geography of Mechanical Turk workers (Pavlick et al. 2014)..... 12

Figure 3: Gamification elements used in 24 empirical peer-reviewed studies (Hamari et al. 2014).... 18

Figure 4: Gamification contexts from 24 empirical peer-reviewed studies (Hamari et al. 2014) 19

Figure 5: Custom crowdsourcing system architecture..... 24

Figure 6: Example User object 27

Figure 7: Example Content object with translations and rankings. 28

Figure 8: Experiment 1: Screenshot of prototype system..... 30

Figure 9: Experiment 1: Screenshot of final system..... 31

Figure 10: Experiment 2: Screenshot of translate page 32

Figure 11: Experiment 2,3 and 4: Screenshot of rank page..... 33

Figure 12: Experiment 2 and 3: Screenshot of leaderboard page 34

Figure 13: Experiment 4: Screenshot of leaderboard page..... 35

Figure 14: Online banking mobile wallet payment screen. 36

Figure 15: Receiving a mobile payment and withdrawing cash..... 37

Figure 16: Payment model: Input variables and payment points..... 48

Figure 17: Payment model: Word cost and users matrix..... 50

Figure 18: Experiment 2: Active users and reward earners per group 54

Figure 19: Experiment 2: Number of users per reward tier. 55

Figure 20: Experiment 2: Total translations for each group. 57

Figure 21: Experiment 2: Total ranks for each group..... 58

Figure 22: Experiment 2: Total cheat contributions for each group..... 59

Figure 23: Experiment 2: Total contributions for each group 60

Figure 24: Experiment 2: Daily contributions for all users 61

Figure 25: Experiment 2: Hourly contributions for all users..... 62

Figure 26: Experiment 2: Intervals in minutes between translations by all users 63

Figure 27: Experiment 2: Intervals in minutes between ranks by all users 63

Figure 28: Experiment 2: Intervals under 30 minutes between translations by all users 64

Figure 29: Experiment 2: Intervals under 30 minutes between ranks by all users 64

Figure 30: Experiment 2: Box and Whisker plot for translation interval groups 66

Figure 31: Experiment 2: Box and Whisker plot for rank interval groups 68

Figure 32: Experiment 2: Box and Whisker plot for translation interval groups combined on work effort 70

Figure 33: Experiment 2: Box an Whisker plot for rank interval groups combined on work effort 72

Figure 34: Experiment 2: Box an Whisker plot for translation interval groups combined on reward type 73

Figure 35: Experiment 2: Box an Whisker plot for rank interval groups combined on reward type ... 75

Figure 36: Experiment 2: Average contributions for an active user per group 76

Figure 37: Experiment 4: Leaderboard payment scheme 81

Figure 38: Experiment 4: Comparison of leaderboard contributions and rewards 82

Figure 39: Experiment 4: Total contributions and cheat contributions 83

Figure 40: Experiment 4: Daily contributions for all users 84

Figure 41: Experiment 4: Hourly contributions for all users 85

Figure 42: Experiment 4: Intervals in minutes between translations by all users 86

Figure 43: Experiment 4: Intervals in minutes between ranks by all users 86

Figure 44: Experiment 4: Intervals under 30 minutes between translations by all users 87

Figure 45: Experiment 4: Intervals under 30 minutes between ranks by all users 87

Figure 46: Experiment 2 and 4: Box and Whisker plot for translation interval groups..... 88

Figure 47: Experiment 2 and 4: Box and Whisker plot for rank interval groups 89

Figure 48: Experiment 4: Box and Whisker plot for translation and rank interval for all users 90

Figure 49: Experiment 2 and 4: Average contributions for an active user per group..... 91

List of Tables

Table 1: Financial rewards offered by crowdsourcing research 15

Table 2: Financial rewards offered by Mechanical Turk translation tasks - 18/09/2014 16

Table 3: Financial rewards offered by translation tasks on the Reddit community "HITs worth turking for" - 18/09/2014 16

Table 4: Experts evaluation of proposed tweets for Experiment 1 40

Table 5: Experiment 1 Tweets and Activity 41

Table 6: Effort and reward schemes in Experiment 2 43

Table 7: Qualifying assessment sentences and their translations 46

Table 8: Quotations for translating the English "Cape Town" Wikipedia article into isiXhosa 47

Table 9: Group 1: Consistent effort and consistent rewards. 51

Table 10: Group 2: Increasing effort and consistent rewards. 51

Table 11: Group 3: Consistent effort and increasing rewards. 51

Table 12: Group 4: Increasing effort and increasing rewards 52

Table 13: Group 5: Consistent effort and decreasing rewards. 52

Table 14: Group 6: Increasing effort and decreasing rewards. 52

Table 15: Experiment 2: Population statistics on translation intervals in seconds for all groups 66

Table 16: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups 67

Table 17: Experiment 2: Population statistics on rank intervals in seconds for all groups 67

Table 18: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on rank interval groups 69

Table 19: Experiment 2: Population statistics on translation intervals in seconds for groups combined on work effort 70

Table 20: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups combined on work effort 71

Table 21: Experiment 2: Population statistics on rank intervals in seconds for groups combined on work effort 71

Table 22: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups combined on work effort 72

Table 23: Experiment 2: Population statistics on translation intervals in seconds for groups combined on reward type 73

Table 24: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups combined on reward type 74

Table 25: Experiment 2: Population statistics on rank intervals in seconds for groups combined on reward type 74

Table 26: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on rank interval groups combined on reward type 75

Table 27: Experiment 4: Population statistics on translation and rank intervals in seconds for all users 89

Table 28: Experiment 4: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation and rank intervals 90

Table 29: Wikipedia articles used as the dataset for all experiments 106

1. Introduction

isiXhosa (Xhosa) is spoken by more than 8 million first language speakers in South Africa - 16% of the country's population (Statistics South Africa 2012). Like the majority of South Africa's eleven official languages, isiXhosa is categorised as a low resource language (Johnson 2011; Eiselen and Puttkammer 2014) with a scarcity of digital content, linguistic models and tools and Information Retrieval (IR) services such as search and translation.

People interact with IR technology on a daily basis when using Web services like Google Search and Microsoft Bing, and when searching for files on their computer, but IR technology is also used for automatic classification of content into categories and for automatically converting text from one language to another (Croft et al. 2010). Core IR algorithms include basic text processing, indexing, filtering and ranking but IR is also related to Natural Language Processing (NLP) and computational linguistics for the development of text processing algorithms and language models (Croft et al. 2010). Higher-level IR services and core algorithms can be developed though statistical analysis of text corpora assembled from large quantities of digital content or through a rule-based approach. Low resource languages such as most spoken exclusively in Africa don't have access to the same search and translation services available in European languages (Johnson 2011) because they lack the resources to assemble the necessary corpora and tools.

Attempts to assemble monolingual and multilingual isiXhosa corpora from South African governmental websites (Johnson 2011; Eiselen and Puttkammer 2014) or by crawling isiXhosa specific websites (Drummer 2013) found that the quantity and quality of the content was not sufficient to produce working machine translation systems (Johnson 2011; Drummer 2013) and that the content was highly specialised and unsuitable for general language corpora (Johnson 2011; Eiselen and Puttkammer 2014). Employing professional translation services to translate original content at the quantities required for building language corpora is expensive (Zaidan and Callison-Burch 2011) and using religious texts such as the World English Bible (WEB), which are available in multiple low resourced languages, is unsuitable for assembling practical language corpora because of specialised subject matter and unique literally style, which differs greatly from the contemporary usage of the languages (Johnson 2011).

A crowdsourcing system with gamification elements was proposed as a solution to affordably gather original multilingual content for building language corpora for low resource languages. Gamification is the process of using design elements from video games in non-gaming contexts to motivate people to engage (Deterding et al. 2011). Points, leaderboards, badges and achievements are some of the gaming elements that are often employed in software systems that span many different sectors (Hamari et al. 2014). Crowdsourcing is the process of outsourcing tasks normally done by an employee or contractor to an anonymous crowd (Howe 2008). An example of this is reCAPTCHA[1], the online image transcription service that masquerades as a human/bot detection service for websites. By transcribing words from document scans that Optical Character Recognition (OCR) algorithms have not been able to confidently identify, users are performing tasks that computers are inefficient at or are unable to do.

Various open source crowdsourcing systems were evaluated and, in the end, 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 payment as secondary motivation. Two of the experiments appealed to the users based on the intrinsic value of the task while the other two offered monetary payments to test whether the game elements appealed to users more than financial rewards. The first payment experiment tested whether paying users consistent, increasing or decreasing rewards for each subsequent contribution would affect their motivation. The second payment experiment introduced a payment scheme where users were paid based on where they placed on the leaderboard of total contributions and, when compared to the first payment experiment, gave insight into how people perceive guaranteed immediate rewards versus future potential rewards.

1.1. Research Questions

This research proposes, firstly, the use of crowdsourcing as a cost effective approach for gathering multilingual content for building language corpora for low resource

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

languages. Secondly, it explores the effects of intrinsic, gamification and monetary motivation factors - in the crowdsourcing process - on user engagement and quantity of contributions. Only the quantity and interval between contributions were studied but the quality of contributions could also be analysed to study user motivation. The specific research questions are:

1. In the context of crowdsourcing content for low resource languages, what is the effect on user engagement and contribution quantity when paying users consistent, increasing or decreasing rewards for subsequent contributions?

2. In the context of crowdsourcing content for low resource languages, will users continue to contribute if payments are taken away and only intrinsic motivators remain?

3. In the context of crowdsourcing content for low resource languages, is the possibility of future rewards more attractive to users than direct guaranteed incentives?

1.2. Thesis Outline

The literature review in Chapter 2 examines studies relating to assembling language corpora and building linguistic tools for isiXhosa, crowdsourcing and specifically crowdsourcing translations, gamification and applying gamification to crowdsourcing. Reward rates from past crowdsourcing studies were summarised where possible and were used in Experiment 2 (Chapter 5) and Experiment 4 (Chapter 7) to guide the selection of reward payment points.

Chapter 3 reviews existing open source crowdsourcing software and details the design and development of a custom crowdsourcing system to specifically address the needs of this research.

Experiment 1 was conducted as a pilot study (Chapter 4), which attempted to gather a crowd of qualified bilingual English/isiXhosa speakers, by appealing to users on social networks, to participate in a crowdsourcing game.

Chapter 5 covers Experiment 2, which explores the effects of paying users consistent, increasing or decreasing rewards for subsequent contributions and addresses the first

research question. A crowd of bilingual English-isiXhosa students was gathered from the University of Cape Town. Taking into account metrics sampled from the literature review, a comprehensive payment model was built, which allowed payment plans to be rapidly developed.

The third experiment, covered in Chapter 6, together with Experiment 1 attempts to answer the second research question. The same pool of users used in Experiment 2 were once again appealed to but all financial rewards were removed to see if users would continue contributing when the financial rewards were removed and only the intrinsic value of participating in the project remained.

Chapter 7 covers the final experiment, which tested the effects of a reward system that pays users based on where they place on a leaderboard and addresses the third research question.

The conclusion and a discussion of the experiment results and shortcomings are covered in Chapter 8. Finally, Chapter 9 proposes future work.

2. Literature Review

2.1. isiXhosa

isiXhosa (Xhosa) is an Nguni language that is part of Africa's South Eastern Bantu languages. It has the highest concentration of speakers in South Africa's Eastern Cape Province (Statistics South Africa 2012), as seen in Figure 1, and is the second most spoken first language in South Africa (more than 8 million speakers) after isiZulu.

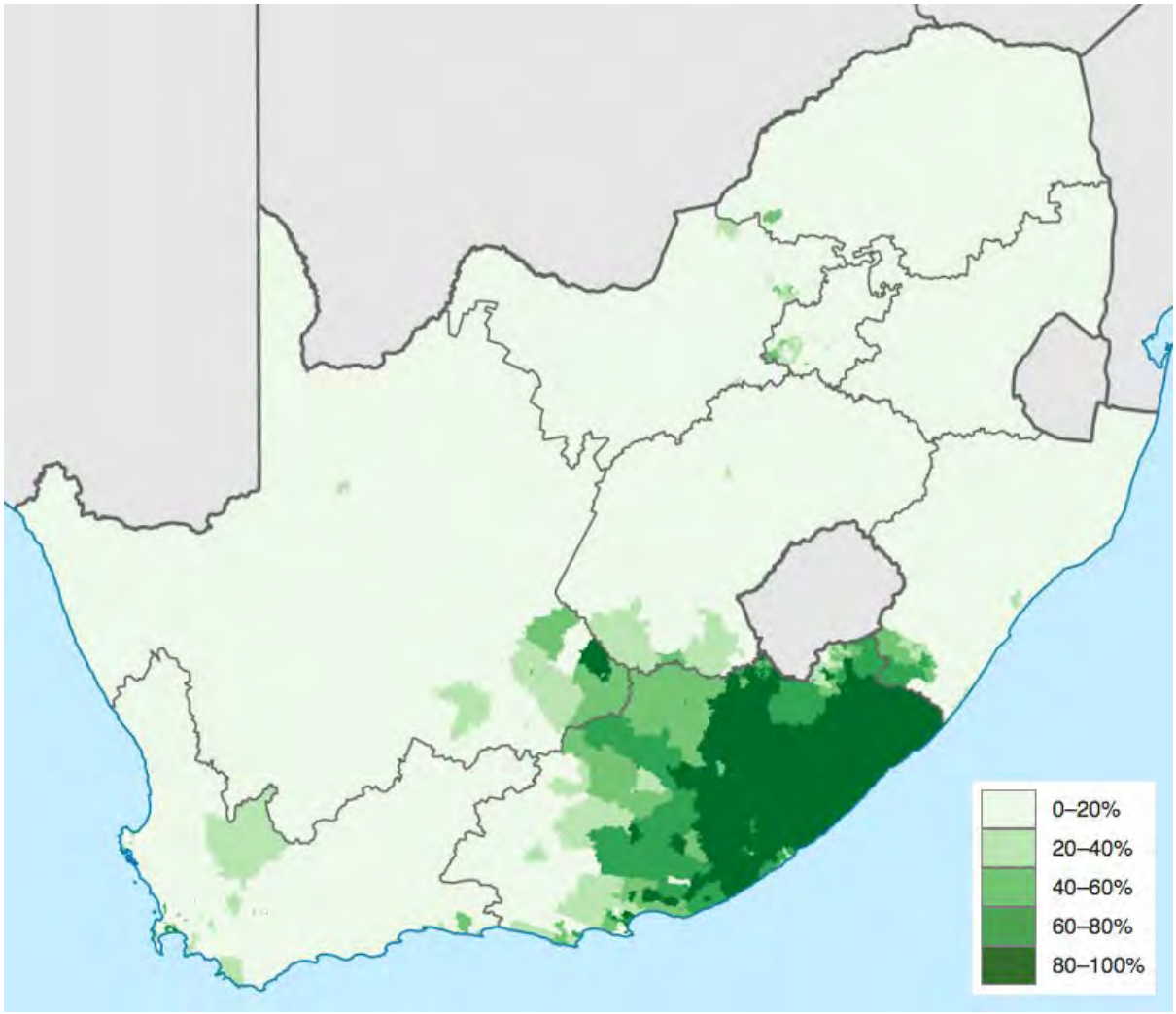

Figure 1: At home isiXhosa speakers in South African (Htonl 2013; Statistics South Africa 2012)

isiXhosa is a morphologically rich and highly agglutinative language (Webb 2000), forming words by gluing together prefixes, suffixes and other affixes to a word's base form or stem to create plural forms and for gender agreement. For example, the base

form of the isiXhosa word for month is “inyanga”; gluing “i” in front produces the plural form “iinyanga”. Developing machine translation systems for agglutinative languages with little to no language models is particularly challenging (Johnson 2011) because the base form of words and their specific usage can be incorrectly categorised or not recognised at all and marked as out of vocabulary - not appearing in the system’s known vocabulary.

The following sub-section reviews past work on language corpora creation and using language corpora to build information retrieval and machine translation systems for isiXhosa.

2.1.1. Language Corpora

A language corpus (plural corpora) is a representational collection of digital text used by linguistic research (Wynne et al. 2005) to develop language models and IR services such as search and translation and automatic speech recognition. Language corpora can be classified as monolingual or multilingual, general or specialised (McEnery, Xiao, and Tono 2006) and synchronic or diachronic in their approach (De Saussure 2011). A corpus is monolingual when it contains text in only one language and multilingual when it contains text in two or more languages. A general corpus serves the purpose of providing an overall description of the language and its variety. The British National Corpus (BNC)[2], for example, contains 100 million words of written and spoken language from a wide range of sources that represent a cross section of spoken and written British English from the late twentieth century. A specialised corpus is domain (e.g. medicine or law) or genre (e.g. newspaper or academic text) specific (McEnery, Xiao, and Tono 2006). Everyday documents and conversations are generally the best sources of content for building a general language corpus and highly specialised sources such as academic and industry journals should be avoided unless building a specialised corpus (Wynne, Arts, and Service 2005). A diachronic approach to building language corpora is concerned with the historical evolutionary development of a language while a synchronic approach looks at a specific point in

[2] http://www.natcorp.ox.ac.uk/

time and is not concerned with historical evolutionary development of the language (De Saussure 2011).

Modern languages and their uses are infinitely diverse and constantly evolving; it would be infeasible and impossible to archive every aspect (McEnery, Xiao, and Tono 2006) and this is why a corpus samples a language through various criteria in order to represent the full range of a language or its variety, whereas an archive does not. Corpus sampling criteria should be carefully chosen so that its linguistic features are not determined by the design, for example requiring a specific distribution of words or grammatical features. A corpus should be balanced in its sampling by ensuring it encompasses a full distribution of text categories (news articles, magazines, novels, conversations, etc.).

Once a corpus has been assembled it can be annotated by tagging different aspects with interpretative linguistic information (Wynne et al. 2005) to facilitate advanced linguistic services. For example part-of-speech (POS) tagging annotates words with their word class (noun, verb, adjective, adverb, etc.), which can then be used to identify words with the same spelling but different usage and meaning or to generate word frequency lists (Wynne et al. 2005). Annotations will not be required by all users (Wynne et al. 2005) and therefore should be easily separable, allowing the raw corpus to be retrieved. Many other types of annotation exist: phonetic annotation is concerned with how a word is spoken; semantic annotation describes different semantic usage of words; pragmatic annotation adds information about the different usage of the same word in speech; and lexical annotation identifies the base form of words (Garside, Leech, and McEnery 1997). Annotated corpora that can be aligned at the sentence level and compared side-by-side are called aligned or parallel corpora and are an essential component for building automatic translation systems through machine translation (MT) (Johnson 2011). Section 2.1.1.2 covers a number of different MT approaches, all of which require large corpora and some of which require linguistic models and tools.

Acknowledging the scarcity of isiXhosa digital content, Johnson (2011) attempted to automatically assemble an isiXhosa/English language corpus for the purpose of training a machine translation system by crawling South African governmental websites. After pre-processing and alignment of the parallel text, a small and specialised corpus of about 4000 parallel lines was produced. The corpus was not

large enough, and still too specialised, to produce a working machine translation system.

2.1.1.1. Information Retrieval

In 1968, Salton (Salton 1968) defined Information Retrieval (IR) as the field concerned with the structure, analysis, organisation, storage, searching, and retrieval of information. This definition is still true today (Croft, Metzler, and Strohman 2010) and it encapsulates the primary IR technologies like Web and desktop search, which we interact with on a daily basis. Search engines are primarily concerned with returning relevant results (Croft, Metzler, and Strohman 2010). Before a document can be queried, it first needs to be processed and transformed into topic indices using linguistic algorithms such as tokenising, stopping, stemming and analysed with NLP techniques to detect syntactic, grammatical and semantic features (Croft, Metzler, and Strohman 2010). Tokenising detects distinct tokens, which generally match to individual words; stopping removes common tokens that on their own don't describe the topic; stemming groups words with a common base form to increase the result and relevance (Croft, Metzler, and Strohman 2010). Without a well-defined and annotated language corpus, complete text processing cannot be done and querying takes a very basic form, simply matching and counting the occurrence of search words to find relevant documents (Croft, Metzler, and Strohman 2010).

Due to numerous social, historical and political reasons, isiXhosa has not received its due attention in the NLP research space (Johnson 2011). No prior work was found on isiXhosa IR. Most of the prior work relates to isiXhosa computational linguistics (Allwood et al. 2010). Allwood et al. performed initial tagging on a raw spoken isiXhosa corpus, paying special attention to isiXhosa's agglutinative nature, and developed graphical tools to facilitate corpus tagging (Allwood et al. 2003). de Klerk analysed selected aspects of codeswitching - the use of more than one language in the same conversation - of bilingual English/isiXhosa speakers, in a corpus of approximately 550,000 transcribed words from casual verbal discussions, to determine the level of bilingualism of users (de Klerk 2006). Pretorius et al. exploited the similarities shared by isiXhosa and isiZulu as Nguni Bantu languages and used advances in isiZulu morphological analysers to further develop isiXhosa morphological analysers, saving significant time over building them from scratch (Pretorius et al. 2009).

Eiselen and Putkamer were successful in developing multiple linguistic resources for ten of South African's official languages, including isiXhosa (Eiselen and Puttkammer 2014). Subsets of multiple corpora were annotated on token, orthographic and morphological layers to develop core IR tools such as a tokeniser, lemmatiser, part of speech tagger and morphological decomposer for each language. Eiselen and Puttkammer also acknowledged the scarcity of high quality digital content for most of South Africa's official languages. They also made use of mostly monolingual governmental texts, supplemented with smaller sets from scientific journals, news articles and journals, and experienced the difficulties of building language resources without any tools for agglutinative languages. Annotating the corpora was challenging but they were able to exploit similarities amongst many of the languages to reuse tools.

2.1.1.2. Machine Translation

Machine Translation (MT) is the translation from one language to another by a computer (Jurafsky and Martin 2000) using either a rule-based - which requires linguistic knowledge of both languages - or statistical approach. The simplest rule-based approach performs a direct word-for-word translation using a complete bilingual dictionary but does not take into account word order and sentence structure, ambiguity and semantic or cultural differences (Jurafsky and Martin 2000). More advanced approaches require extensive linguistic knowledge of both languages to analyse the structure of the source text and parse it into an intermediate syntactic tree to create the correct structure in the target language.

Statistical machine translation (SMT) systems translate by statistically analysing parallel language corpora (Sharwood 2013) to construct translation models that capture the translation probabilities. SMT produces translation systems that behave in a manner captured by the training language corpora; it is therefore important that the parallel language corpora are representative of each language and its usage and contain high quality translations (Jurafsky and Martin 2000). Some hybrid SMT approaches use linguistic tools (Sharwood 2013), such as POS taggers, as a pre-processing step to create more syntactically correct output.

The quality of SMT is strongly related to the size of the parallel corpus (Post, Callison-Burch, and Osborne 2012) and, as most low resource language pairs have

little or no available bilingual training data, they are severely underrepresented in MT research. Some language pairs are fortunate enough to have large open datasets freely available online (Koehn 2005), such as the Proceedings of the European Parliament[3], which are often used as datasets for constructing parallel corpora for SMT. Work has been done to assemble various parallel isiXhosa corpora for SMT, but in all the cases there was not enough high quality content or linguistic tools for hybrid approaches (Johnson 2011; Sharwood 2013; Eiselen and Puttkammer 2014).

Johnson's (2011) assembled parallel corpus of 4000 lines was not enough to build a working machine translation system. Johnson noted that the availability of morphological analysers and isiXhosa language recognition algorithms would make building isiXhosa machine translation systems more feasible and referred to Pretorius et al.'s (2009) usage of exploiting similarities between isiXhosa and isiZula to bootstrap the process.

Drummer and Sharwood (Sharwood 2013; Drummer 2013) were able to source additional parallel content from Nal'ibali, a website with stories in many of South Africa's official languages, and from magazines published on the Jehovah's Witness website. They also found that despite having two more sources of content than Johnson, there was still not enough high quality content to train a machine translation system without the assistance of language models and tools (Sharwood 2013; Drummer 2013). Sharwood (2013) also exploited similarities between isiXhosa and isiZulu to build a two-step machine translation system that translated from isiXhosa to English by translating isiXhosa first to isiZulu and then to English.

2.1.2. Summary

isiXhosa and isiZulu's shared Nguni origin has allowed certain tools and techniques developed for isiZulu language processing to be used for isiXhosa, but the limited availability of general digitised texts for assembling language corpora has hampered the development of information retrieval and machine translation services for isiXhosa. Section 2.2 reviews how crowdsourcing has been used for translation and

[3] http://www.statmt.org/europarl/

assembling language corpora and analyses the cost of crowdsourcing translations. Section 2.3 reviews research into using gamification, an intrinsic motivation factor, in crowdsourcing projects, as an alternative or an accompaniment to an extrinsic motivation factor like receiving financial rewards.

2.2. Crowdsourcing

Crowdsourcing overlaps with citizen science (Wiggins and Crowston 2011), a form of research collaboration that actively involves the public in scientific research projects but, unlike citizen science, it is entirely mediated by information and communication technologies (ICTs). Citizen science has been popular in ecology and environmental science projects (Silvertown 2009). For example, the Evolution MegaLab[4] project tracks snail shell evolution from photos uploaded by the public and the OPAL[5] (OPen Air Laboratories) project, which engages the public to record local wildlife and the quality of air, soil and water. 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 (Ross et al. 2010).

An important aspect of crowdsourcing is making an appeal to an anonymous crowd (Silvertown 2009). Online crowdsourcing marketplaces like Amazon's Mechanical Turk[6] (MTurk) and CrowdFlower[7] provide large communities of users across the globe, which greatly increases the ease of gathering users for most projects (Callison-Burch 2009). A sampling of user demographics for Mechanical Turk (Ross et al. 2010) revealed that 85% were from the United States and India and the remaining 15% were scattered across the rest of the world, making it challenging to gather enough volunteers if your project requires users from under-represented countries, as was the case with Ambati and Vogel (Ambati and Vogel 2010) who struggled to gather sufficient Haitian Creole speakers even after raising the reward offered. Pavlick

[4] http://www.evolutionmegalab.org/

[5] http://www.opalexplorenature.org/

[6] https://www.mturk.com/mturk/welcome

[7] http://www.crowdflower.com/

et al. (Pavlick et al. 2014) ran a more recent multi-language translation experiment to sample Mechanical Turk users in 2014, but still found that the majority (57%) of the users were from India or the United States, as seen in Figure 2. Africa was still severely under represented; for example, there were only 24 Egyptian users and 23 Swahili speakers (Pavlick et al. 2014).

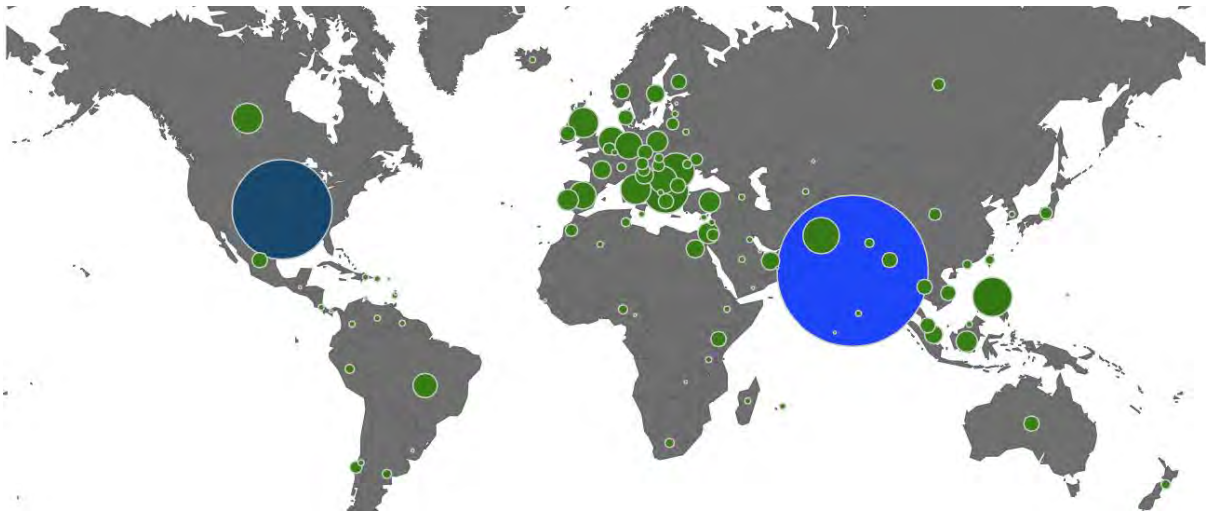

Figure 2: The geography of Mechanical Turk workers (Pavlick et al. 2014)

Crowdsourcing marketplaces implement quality control measures in the form of pre-assessments, requiring redundant contributions and blacklisting specific users but, despite these measures, it is still challenging to get high quality contributions (Callison-Burch 2009). Users will still take on tasks that they are not skilled to perform or find other methods of cheating. Ambati and Vogel (2010) discovered that more than 50% of their Spanish-English tasks were completed in India. Some translation tasks are also susceptible to having automatic translations being submitted by cheating users or bots (Callison-Burch 2009; Ambati and Vogel 2010).

2.2.1. Translations and Parallel Corpora

Only a few languages in the world receive the financial support, research interest and human effort from expert bilingual speakers for the continued development of large-scale parallel corpora for the development of automatic translation systems. For the remaining languages crowdsourcing can arguably provide a large pool of users with varying experience for free or at attractive rates (Ambati and Vogel 2010).

Callison-Burch (2009) used Mechanical Turk to: crowdsource English translations for Spanish, German, French, Chinese and Urdu; and to rank existing machine translations. In doing so, it was found that crowdsourcing was affordable enough that redundant translations and rankings could be collected. Agreement amongst users increased as redundancy was increased; at three redundant contributions, agreement amongst users was nearly the same as three expert users. Measuring agreement was an effective means of pre-assessing users, weighting user contributions and removing low quality contributions altogether. Users who contribute more than other users contributed poorer quality contributions. Overall, the crowdsourced translation quality was within the standard deviation of expert translations and was significantly better than translations produced by machine translation.

Ambati and Vogel (2010) used Mechanical Turk to crowdsource six parallel language corpora for English, Spanish, Telugu and Urdu, while comparing the effect of in-context and out-of-context phrase translations on total cost and user agreement. They found that translating phrases in the context of the original sentence led to a higher agreement amongst volunteers, as they were less likely to fabricate the context. They also found measuring agreement amongst contributed translations to be an effective means of removing low quality translations.

Negri and Mehdad (2010) used CrowdFlower to build a bilingual corpus in a short time and with a limited budget. They reviewed the work by Callison-Burch (2009) and noted that translating with up to 5 levels of redundancy was significantly more expensive than ranking with redundancy and therefore decided to use no redundant translations, 5 redundant rankings and multiple translation cycles, terminating when ranking users agree that the translation is correct. If ranking users do not agree, a new translation is sourced and ranked again. To further reduce costs and improve the overall quality of contributions, they pre-assessed users and filtered out users who did not maintain a gold standard, thereby only paying for qualified contributions.

Zaidan and Callison-Burch (2011) used Mechanical Turk to crowdsource Urdu to English translations at rates considerably more affordable than those offered by professional translations services. They were able to achieve translation quality near to professional levels by also using redundant translations, translation edits by other users, user pre-assessments and weighting of user contributions.

In 2012, recognising the established efficiency of using crowdsourcing for constructing parallel corpora, Post et al. (2012) crowdsourced English translations for six languages from the Indian sub-continent to construct parallel corpora for training machine translations systems. Their work is particularly interesting as all six languages - Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu - are low resource and highly agglutinative languages, like isiXhosa. Users were pre-assessed and filtered by checking their translations against bilingual dictionaries and tracking various metrics such as location, translation length and translation time. They discovered that there were many orthographic differences between translations, which can be attributed to either spelling mistakes, caused by the difficulties of inputting special characters used by some of the languages and/or regional language variations. Alignment of the parallel corpora was difficult because of the morphological richness of the various agglutinative languages but was further challenged by orthographic inconsistencies of spelling and sentence structure and the authors speculated that standard alignment heuristics might not be well suited under these conditions.

Crowdsourcing has also been used to source translations for emergency response scenarios (Munro 2010). The Haiti earthquake in 2010 left existing emergency response services incapable of coping with the large volume of text messages being received; majority of the volunteer emergency responders did not speak Haitan Creole. To solve the problem, a crowdsourcing system was used to allow Haitan Creole and French speaking communities to translate more than 40,000 emergency text messages over six days into various languages, with an average turn-around response time of 10 minutes. The incredible human effort resulted in hundreds of lives being saved, food and aid reaching tens of thousands of people and multiple parallel corpora being built.

2.2.2. The Cost of Crowdsourcing

Rewards offered by the various projects covered in the literature review are summarised in Table 1. They were used to guide the selection of payment points while designing the project's various experiments that offered financial rewards. Some of the studies specified payment rewards per task and others per word, all in US Dollars. It should be noted that the task reward offered by Post et al. (2012) was significantly higher than the other studies; there was no explanation given as to why the specific reward value was selected.

Table 1: Financial rewards offered by crowdsourcing research

Research	*Task Detail*	*Reward (USD)*	*Per Word*
(Callison-Burch 2009)	Rank 5 German to English machine translations	$0.01	
	Translate German to English	$0.10	
	Detect if the translation was created by machine translation	$0.006	
(Ambati and Vogel 2010)	Translate Spanish to English	$0.01	
	Translate Telugu 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	
(Negri and Mehdad 2010)	Translate English to Spanish		$0.01
	Validate translation		$0.002
(Zaidan and Callison-Burch 2011)	Translate Urdu to English	$0.10	$0.005
	Edit 10 sentence	$0.25	
	Rank 4 translation groups	$0.06	
(Post et al. 2012)	Translate Bengali, Hindi, Malayalam, Tamil, Telugu or Urdu to English	$0.70	

On 18 September 2014, Mechanical Turk was searched to find examples of more recent translation projects and the rewards they offer (see Table 2). The rewards were again reported in US Dollars. Three of the projects were for English to German translations by the same author and were created a few days after each other, with each subsequent job offering a higher rate. It is suspected that the reward offered was too low for the large amount of work required so the author reposted the jobs with higher rewards. The reward value offered by these projects were still within the range of those sampled from the literature review of crowdsourcing projects.

Table 2: Financial rewards offered by Mechanical Turk translation tasks - 18/09/2014

Task Detail	*Reward (USD)*
Grammar correction of non-first language English speakers	$0.20
Translate English letter to Spanish	$0.50
Translate 5 pages of a European Commission Document from English to German	$0.80
Translate 6 pages of a European Commission Document from English to German	$0.90
Translate 5 pages of a European Commission Document from English to German	$1.00
Translate Hindi sentences to English	$0.30

It is not possible to search past Mechanical Turk HITs but a limited source of historical Mechanical Turk projects was found on social news website Reddit[8] under the "HITs worth turking for" [9] community. The community's purpose is to share Mechanical Turk HITs that offer attractive financial rewards. The shared HITs are succinctly summarised with the reward and an expected time per task provided. On 18 September 2014, a search on the community revealed 4 translation projects, which can be seen in Table 3. Only one of the tasks offered a higher reward than what was observed in the literature review.

Table 3: Financial rewards offered by translation tasks on the Reddit community "HITs worth turking for" - 18/09/2014

Date	*Task Detail*	*Reward (USD)*	*Min*
18/01/2014	Translate Spanish letter to English	$0.30	3
04/06/2012	Translate 10 tweets from Spanish to English	$0.75	5
24/04/2012	Rank multiple Czech to English machine translations	$0.20	2
16/02/2012	Translate 10 Words from Wikipedia (Multiple Languages)	$0.15	2

The sampled rewards, gathered from past studies and projects, show that it was normal to find translation jobs between 2009 and 2014 that offered rewards between USD0.01 and USD0.75 for translation tasks consisting of a few words to entire documents, and were used to guide selection of rewards for Experiment 2 and 4.

8 http://www.reddit.com/

9 http://www.reddit.com/r/HITsWorthTurkingFor/search?q=translate&restrict_sr=on

2.3. Gamification

The Self-Determination Theory (SDT) (Ryan and Deci 2000) is a macro theory of human motivation about the inherent growth tendencies and the innate psychological needs of humans. The theory proposes that by satisfying three innate psychological needs - competence, autonomy and relatedness - intrinsic motivation can be maintained and enhanced. Intrinsic motivation is the inherent tendency to seek out novelty, challenge, exploration and learning in the absence of specific rewards (Harter 1978). Extrinsic motivation occurs when performing a task for an external reward or for compliance with an external regulation (Ryan and Deci 2000). Feeling competent by receiving positive feedback enhances intrinsic motivation while receiving negative feedback diminishes it. Competence alone will not, however, enhance intrinsic motivation unless it is accomplished autonomously, and is more likely to flourish when there is a sense of security and relatedness shared with others (Ryan and Deci 2000).

Gamification or gameful design is the process of using game design elements, rather than specialised game technology or fully-fledged games, in non-gaming contexts, regardless of intention, context and implementation, to improve user experience and motivation (Deterding et al. 2011). Games implement explicit rule systems that define the interaction between actors and achievable goals and outcomes (Deterding et al. 2011). Gamification is based around the idea that, if games can entertain and motivate people to engage, to the point that they are now ubiquitous in everyday life, applying elements found in games such as points, badges and leaderboards to non-gaming settings should make them more enjoyable and engaging too.

A literature review summarising 24 empirical peer-reviewed studies (Hamari et al. 2014) found the three most common game design elements used were virtual points (used by 38%), badges or achievements (used by 38%) and leaderboards (used by 42%) (see Figure 3). Virtual points are awarded to users for completing tasks and these points are usually added up into a total score. Badges or achievements are virtual collectables or milestones awarded for completing specific tasks and are often displayed on a user’s profile and shared with others to indicate their accomplishments. High-score leaderboards foster competition by displaying where users rank on a list of total scores. Some of the other game elements observed were: giving users clear goals

and challenges; tracking and showing progress; categorising users into levels to indicate progress and achievement; visual and audio feedback; story; and theme.

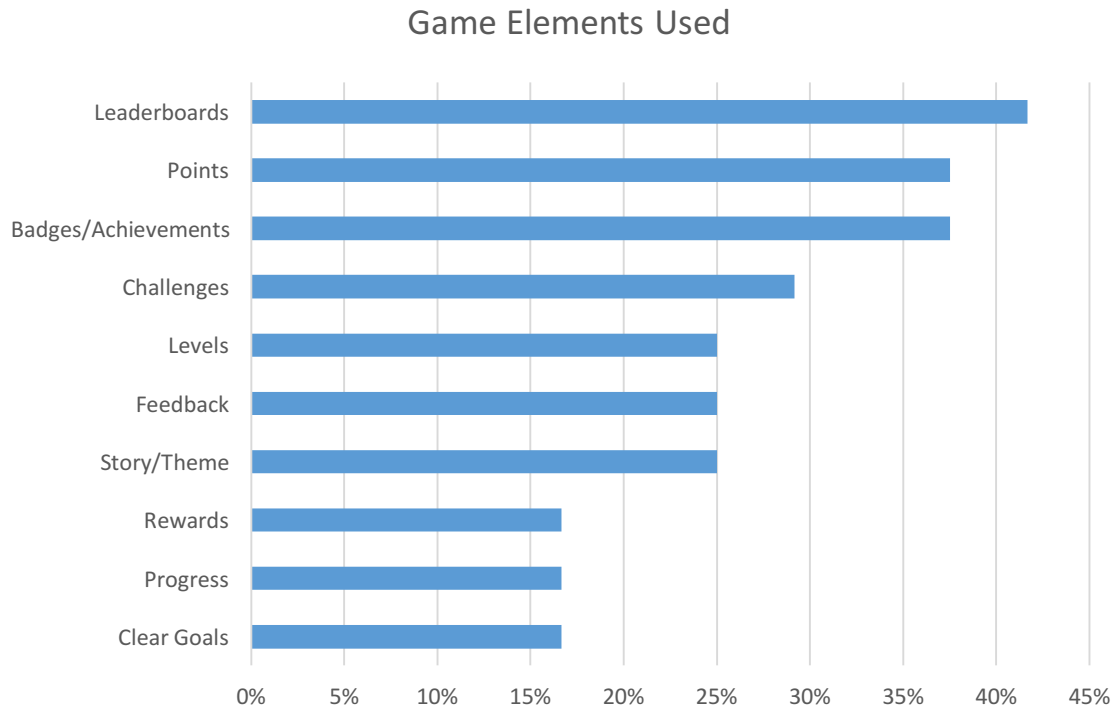

Figure 3: Gamification elements used in 24 empirical peer-reviewed studies (Hamari et al. 2014)

The majority of the studies in the review (Hamari et al. 2014) showed that gamification does have a positive effect on user motivation and engagement but is dependant on what aspect of achieving the task is being gamified and the quality of users participating. Figure 4 shows the context in which the 24 studies used gamification; education and learning was the most common context. The "work" context refers to general crowdsourcing work and was mostly conducted on crowdsourcing systems. Hamari et al. (2014) did not compare game design element uses across the studies or cover competitive systems but Havenga et al. (2012) used leaderboards and badges in a social application for building heritage archives and found that leaderboards outperformed badges at motivating users to contribute.

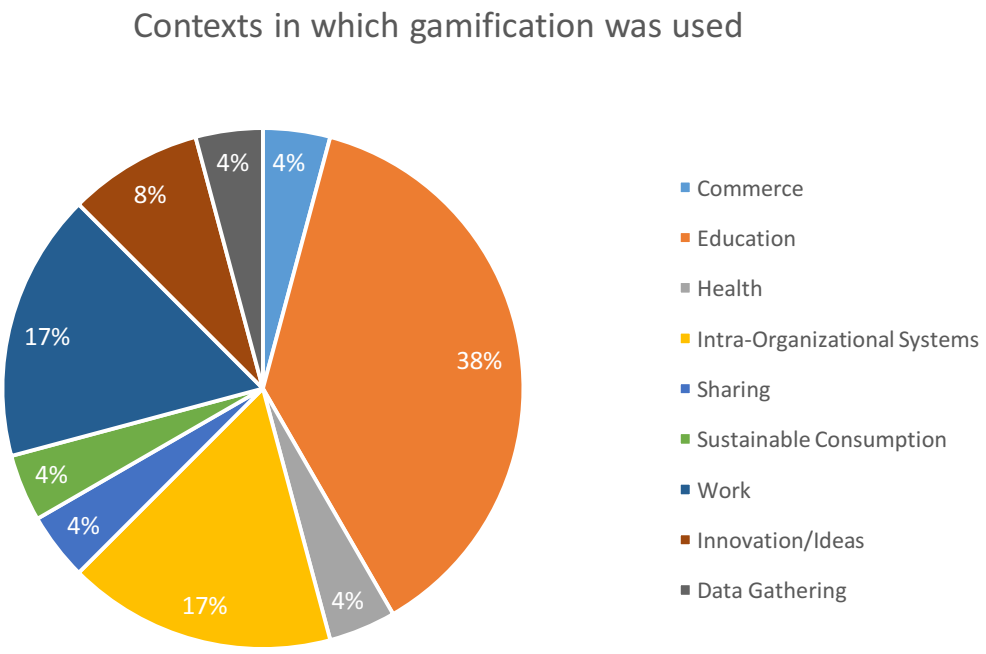

Figure 4: Gamification contexts from 24 empirical peer-reviewed studies (Hamari et al. 2014)

Eickhoff et al. (2012) acknowledge the importance of crowdsourcing for research and industry but they also warn about malicious users who are only concerned with maximising their profits through cheating and contributing low quality work. Extrinsic financial rewards motivate users to seek out the quickest and easiest means to complete tasks, undermining the purpose of crowdsourcing. They hypothesise that there are two types of crowd workers: the money driven, motivated by financial reward, and the entertainment driven, primarily driven by intrinsic needs but who will accept financial reward. Shorter task durations enhance the primary objective of money driven workers but not necessarily entertainment driven workers. Eickhoff et al. used gamification in a project to crowdsource annotations as a means of focussing and attracting entertainment driven workers who are motivated to put in more effort into producing higher quality work to earn more virtual points, achievements and higher rankings on a leaderboard. Workers were awarded points and achievements and were ranked on a high-score leaderboard. They found that through targeted task design and the implemented game elements they received fewer cheaters even from countries that authors of crowdsourcing projects would avoid.

2.4. Summary

isiXhosa is a low resource language spoken by more than 8 million first language speakers in South Africa. Past studies found there is not enough high quality non-specialised content to assemble parallel language corpora, create linguistic tools and machine translation systems. Machine translation is more challenging for agglutinative languages like isiXhosa and large high quality language corpora are essential. Paying for expert translations at the volume required for building parallel language corpora is infeasible.

Crowdsourcing has proven to be an effective and affordable means of translating content. Translation tasks on various crowdsourcing marketplaces can pay between USD0.01 and USD0.75 for translations of a few words to multi-page documents. Translation quality can be improved by using task redundancy, ranking and by pre-assessing and filtering users. Despite these measures, translations still have differences in spelling and structure. Translation tasks and their financial rewards are often time bound so there is a strong incentive to cheat and quickly submit sloppy or automated translations to maximise financial rewards. Past studies have mostly used crowdsourcing marketplaces like Mechanical Turk, which does not have a lot of users in Africa.

Crowdsourcing projects need to sufficiently motivate many people to engage over a period of time to stand a change of being successful. Gamification has proved to be an effective means of motivating and engaging users, depending on the task being gamified and the type of users. Gamification works by appealing to our intrinsic need to seek out novelty and challenges, explore and learn in the absence of specific rewards. Points, badges or achievements and leaderboards are some of the game elements used by gamification studies.

3. System Design

The literature review and background research identified that due to the goals of the research it it was not feasible to use existing crowdsourcing marketplaces to run the experiments. Firstly, higher level system and user requirements are listed and then two existing open source crowdsourcing frameworks - BOSSA and PyBOSSA - were assessed for their suitability as a foundation to build the experiments upon. Both systems were missing features to support gamification and rewards, therefore a custom crowdsourcing system with the missing features was built. The system was designed to be highly configurable, requiring no programming, and allowed all four experiments to be setup with little effort.

3.1. System Requirements

The system needs to support:

- user registration.
- grouping users into payment groups.
- user qualification through pre-assessment questions.
- gamification elements specifically points and leaderboards.
- redundant contributions.
- Limiting contributions.
- payment systems that can be linked to contributions or points.
- various frontend designs for different experiment configurations.

3.2. Open Source Crowdsourcing Systems

In 2007, David Anderson created the Berkley Open System for Skill Aggregation (BOSSA), an open source PHP framework for distributed volunteer thinking (Anderson 2004). A custom BOSSA project is implemented by defining various policies that affect how: HITs flow through the system; users are assessed; and the visual design of the site. BOSSA can also pre-assess users against a set of control

tasks. BOSSA has been used to power popular crowdsourcing projects such as Stardust@Home[10], which had 23,000 volunteers identifying interstellar dust particles, and GalaxyZoo[11], an online system for identifying galaxies, which received more than 50 million contributions in its first year.

PyBOSSA (PyBOSSA 2014) was released in 2013 and is inspired by BOSSA sharing similar core features to manage users and jobs but with additional advanced features and project templates to solve specific crowdsourcing scenarios, such as image classification, transcription and geocoding. It has an active community, recently updated documentation and tutorials and is used on prominent projects by the British Museum and University College London.

As open source frameworks, BOSSA and PyBOSSA offer the possibility of greater customisability over closed source online services such as Mechanical Turk and CrowdFlower, but they introduce additional challenges to creating crowdsourcing projects. They only provide functionality for the most common tasks shared amongst crowdsourcing projects, such as basic user and job management - the required policies need to be developed on a per project basis. Neither has built-in features to support any gamification elements, paying users and a way to group users and tasks.

3.3. Custom System

A prototype which met the system and user requirements identified in Section 3.1 and addressed the missing features from BOSSA and PyBOSSA was successfully implemented in under a week in less than 300 lines of code. The rest of the experiments extended the prototype with little time and effort, mostly due to an effectively designed architecture, which allowed configuration rather than requiring new features to be developed from scratch.

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

[11] http://www.galaxyzoo.org/

3.3.1. Architecture

Two main factors contributed to the short development time of the custom system. Firstly, the literature review and research questions succinctly outlined the needs of the system. Secondly, the researcher had prior experience working with the chosen technologies. CoffeeScript[12], a language that compiles to JavaScript, and Node.js[13], a server application framework written in JavaScript, were used to write the application logic. CoffeeScript is ideal for rapid prototyping, as it requires significantly fewer characters and lines of code than JavaScript to implement features (CoffeeScript 2014). MongoDB[14], a document database that stores JSON (JavaScript Object Notation) documents in schema-less collections, was used to store the experiment data. All of the technologies work well together as the underlying technology is JavaScript.

[12] http://coffeescript.org/

[13] https://nodejs.org/

[14] https://www.mongodb.org/

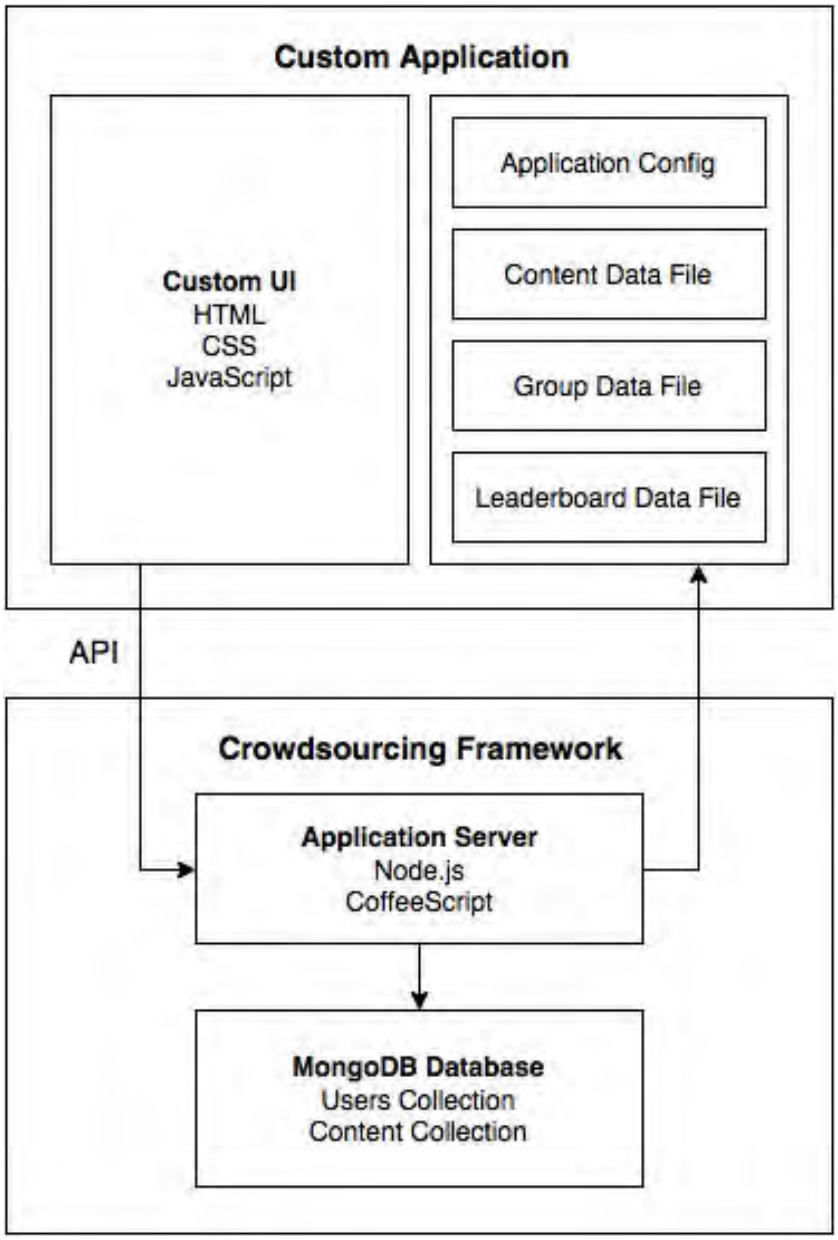

Figure 5: Custom crowdsourcing system architecture

Figure 5 shows an overview of the components a custom application will need to implement when using the developed crowdsourcing system. Each of the four experiments implemented a custom user interface with various HTML (Hyper Text Markup Language) view files and CSS (Cascading Style Sheet) files. The view engine was designed to be flexible enough that all the required functionality could be incorporated into a single page website, as was done for Experiment 1, or across multiple pages, as was done for Experiments 2, 3 and 4. The user interface accessed

the underlying crowdsourcing application server via an Application Programmer Interface (API).

3.3.2. System Configuration

The system was designed to be configurable, allowing different experiments to be setup with minimal effort. The following global settings are configurable by custom applications in the main configuration file:

- **translationRedundancy:** The number of redundant translations required for each sentence.
- **rankingRedundancy:** The number of redundant rankings required for each sentence's translations.
- **translationLimit:** The number of translations a user is allowed to submit.
- **rankingLimit:** The number of rankings a user is allowed to submit.
- **correctQualify:** The number of pre-assessment questions a user needs to get correct to qualify to contribute.
- **rewardType:** A variable to indicate if users are being rewarded per contribution or by their placement on the leaderboard.
- **signupRequired:** A flag to toggle whether or not users had to sign up with an email address and password to contribute.

3.3.3. Application Configuration

The system requires custom applications to provide the following configurable data files (shown in Figure 5):

- **content:** A list of sentences to be translated. The content file could be gathered from external scripts that scraped websites and documents, or manually populated.
- **groups:** A list of groups and their translation and ranking rewards offered.
- **leaderboard:** A list of rewards for leaderboard placement, used if the rewardType was set to leaderboard.

The system allows the sign up process to be turned off and for users to contribute content anonymously; this feature was used in Experiment 1. If the sign up process is enabled, users have to provide their email, display name, mobile number, and student

number (if they were students of the University of Cape Town) and a password. Users were required to provide an active South African mobile number in order to receive payment. A user's display name did not have to be their real name, and would be visible to other users on the leaderboard. Users are required to take the pre-assessment during the sign up process and are given feedback indicating whether they passed or failed.

3.3.4. Data Model

The application server stores content, users and their contributions in a MongoDB database. In MongoDB, JSON documents are stored in collections. A collection is a grouping of documents similar to tables from Relational Database Management Systems (RDBMS) but without any imposed structure (MongoDB 2014).

3.3.4.1. Users

Figure 6 shows an example user JSON document from the users collection. A description of each field follows:

- **id**: The unique ID number given to each user.
- **created:** The date the user joined.
- **email:** The email the user signed up with.
- **displayName**: The name that is shown on leaderboards.
- **mobileNumber**: The South African mobile phone number, which is used for sending mobile payments to the user.
- **studentNumber**: The student identification number if the user is from the University of Cape Town.
- **group**: The group the user belongs to. Some experiments had no groups and therefore all would belong to group 1.
- **correct:** The number of pre-assessment questions the user got correct. Each experiment will configure the required number of correct questions to continue contributing.
- **banned:** A flag to indicate if an admin user has banned the user. Users will be banned if they are found to be cheating.
- **translationCount:** The total translations the user has contributed.
- **rankCount:** The total translation rankings the user has contributed.

- **score:** The total user score the user has achieved from gathering points from translating and ranking. The scoring system can be configured by the experiment.
- **money:** The total amount of money the user has earned. An experiment may round this value up or down before it is finally shown to the user so that they only see denominations that match up to available bank notes.

```
{
  id: ObjectId("546cc7110cc2d9e10960e0a7"),
  created: ISODate("2014-11-19T20:43:35.440Z"),
  email: "sean@seanpackham.com"
  displayName: "translatornator",
  mobileNumber: "0835074196",
  studentNumber: "pcksea001",
  group: 1,
  correct: 4
  banned: false,
  translationCount: 12,
  rankCount: 10,
  score: 1290,
  money: 12.90,
}
```

Figure 6: Example User object

3.3.4.2. Content and Contributions

Figure 7 shows an example document from the content collection. Each sentence from an article will be stored in its own document with its translations and rankings.

```
{
  "id" : ObjectId("546bab7e22c09de53a113248"),
  "group" : 1,
  "index" : 97,
  "translationCount" : 3,
  "rankCount" : 3,
  "text" : "English text...",
  "translations" : [
    { "id" : ObjectId("546cc9820cc2d9e10960e0bf"),
      "created" : ISODate("2014-11-19T16:46:58.686Z"),
      "user" : ObjectId("546cc7110cc2d9e10960e0a7"),
      "translation" : "isiXhosa translation..." },
    { /* translation 2 */ }, { /* translation 3 */ }
  ],
  "rankings" : [
    { "id" : ObjectId("546d00f7636aca507604b9ba"),
      "created" : ISODate("2014-11-19T20:43:35.440Z"),
      "user" : ObjectId("546cc8440cc2d9e10960e0b2"),
      "rank" : {
        "546ccf160cc2d9e10960e119" : "1",
        "546ccefa0cc2d9e10960e116" : "3",
        "546cc9820cc2d9e10960e0bf " : "2" } },
    { /* rank 2 */ }, { /* rank 3 */ }
  ]
}
```

Figure 7: Example Content object with translations and rankings.

The user ID number in Figure 6 has been highlighted and matches the user ID number of the first translation in Figure 7. This linking indicates that the user created this translation. It is also used for various data checks, like making sure that a user cannot submit more than one translation of the same sentence. The same linking structure has been used for connecting rankings with the users who created them. Rankings store scores given by the users with the corresponding translation ID number, also highlighted in Figure 7. A description of each field follows:

- **id**: The unique ID number given to each sentence.
- **group:** The user group the sentence has been assigned to. It can only be translated or ranked by users from the same user group.
- **index:** The index of the sentence in the original article. The index is used to reassemble all translated sentences into translated articles.
- **translationCount:** The number of redundant translations the sentence has received so far. The required translation redundancy can be configured for each experiment.

- **rankCount:** The number of redundant rankings the sentence's translations have received so far. The required sentence redundancy can be configured for each experiment.
- **text:** The sentence in the source language that needs to be translated.
- **translations[i].id:** The unique ID number given to each translation.
- **translations[i].created:** The date the translation was submitted.
- **translations[i].user:** The ID of the user who submitted the translation.
- **translations[i].translation:** The translated sentence submitted by the user.
- **rankings[i].id:** The unique ID number given to each ranking.
- **rankings[i].created:** The date the ranking was submitted.
- **rankings[i].user:** The ID of the user who submitted the ranking.
- **rankings[i].rank:** The ranking object that contains the score given to each of the sentence's translations, linked by the translation ID numbers.

3.3.5. Experiment 1: Pilot Study

Experiment 1 was conducted as a pilot study to test whether participants for future experiments could be gathered from the author's Twitter network without financial rewards and therefor no gamification, payment or user registration features were needed and a simple one-page design was used.

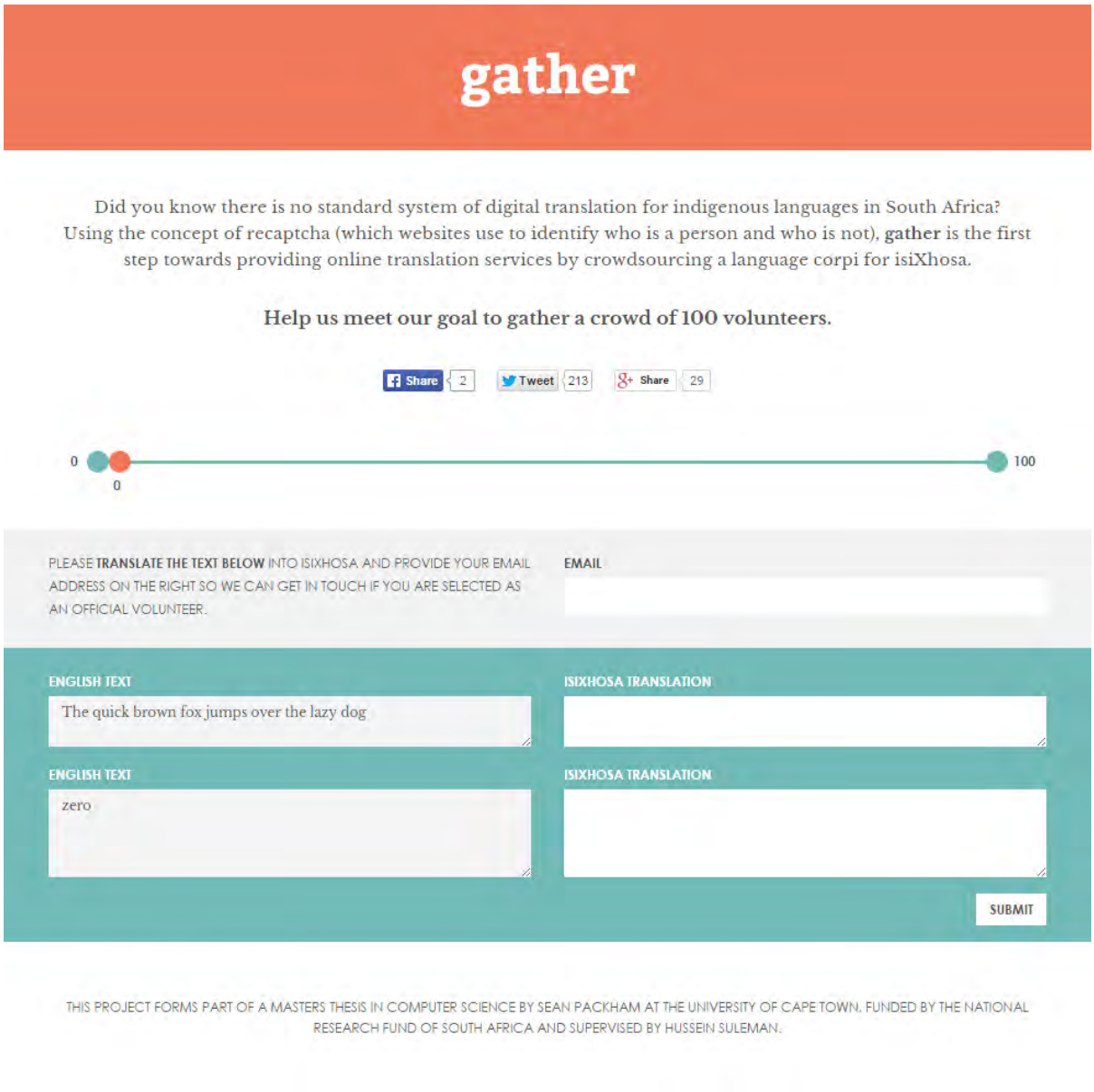

Figure 8: Experiment 1: Screenshot of prototype system

The prototype website for Experiment 1 can be seen in Figure 8. It was assessed with two expert users to identify usability issues. None of the experts spoke isiXhosa so mathematical tasks were used instead of translation tasks and all the other project information remained the same. The experts provided the following comments:

- They felt that the objective of the project could be clearer. They would prefer to know who is running it, what the objective is and what they are required to do to contribute to the objective.

- Both users stated that it was unclear what the progress bar was showing and that it looked like a sliding toggle. Both users suggested using text to indicate progress.
- Both users felt that the design could be simpler and more compact.

The feedback from the expert users was used to create an improved frontend, seen in Figure 9. All aspects of the design were simplified, the heading removed and replaced with large text indicating the progress. The project description was simplified to provide only essential information about who was running the project and its goals.

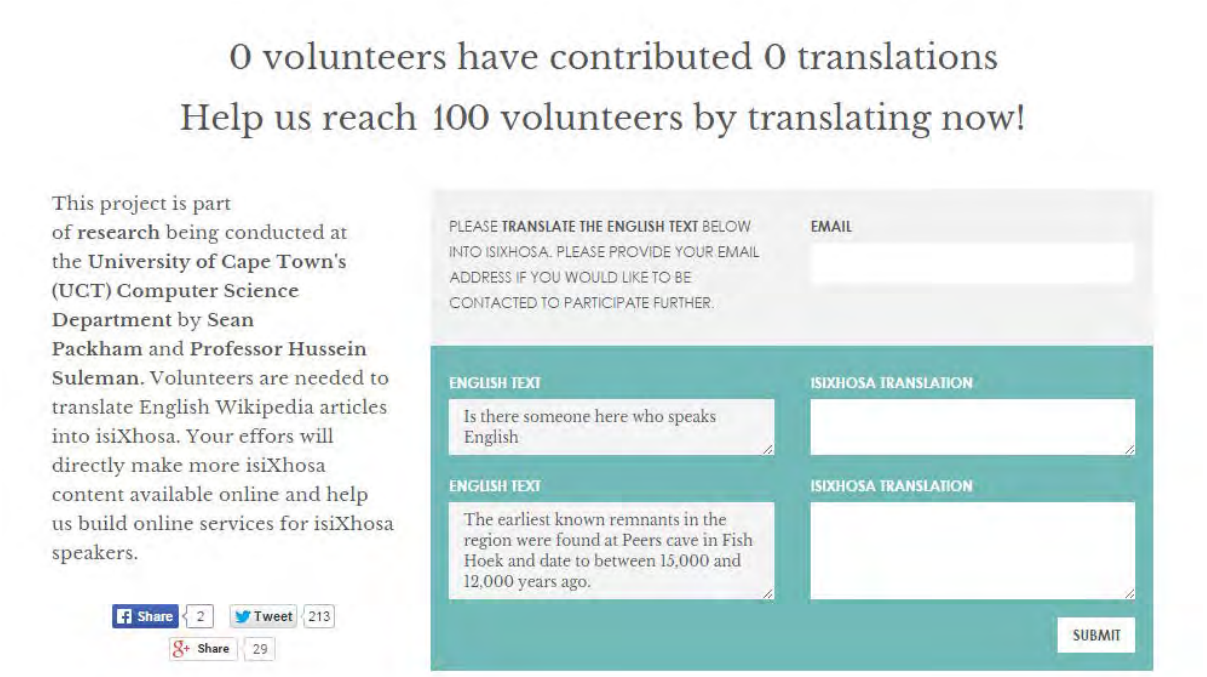

Figure 9: Experiment 1: Screenshot of final system

3.3.6. Experiments 2, 3 and 4

Experiments 2, 3 and 4 shared similar designs configurations and therefore large portions of their user interfaces could also be shared. The three experiments required users to be able to register, contribute translations and rankings and view their standing on the points leaderboard and their earnings.

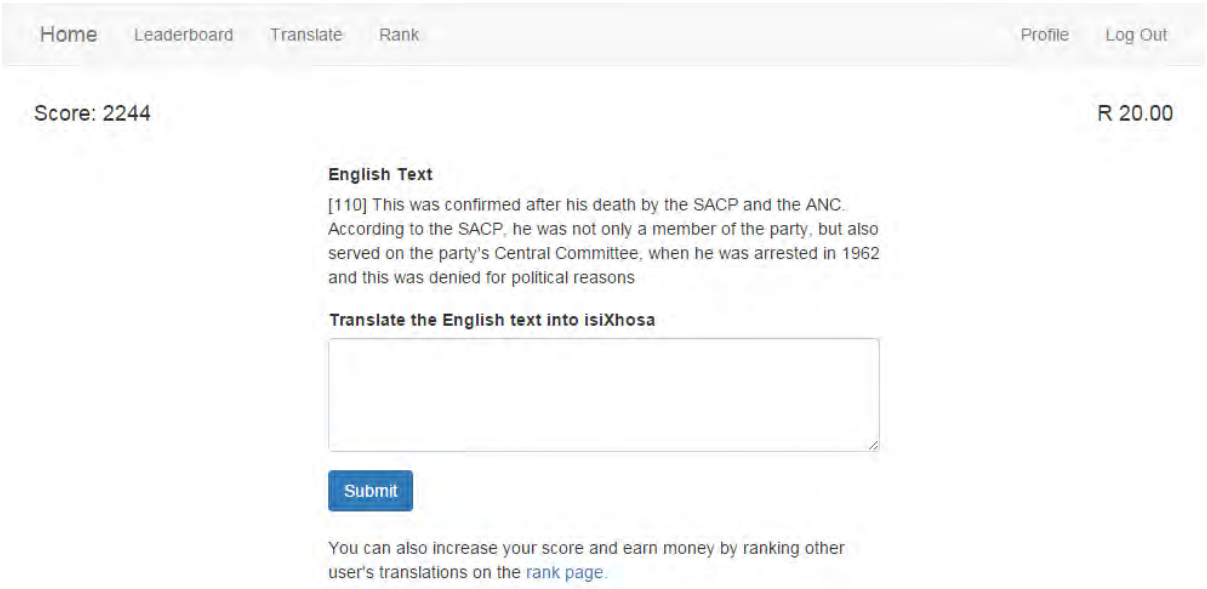

Figure 10: Experiment 2: Screenshot of translate page

Experiment 2's translation page is show in Figure 10 and the ranking page in Figure 11. The user's score is shown to them in the top left corner and their money earned in the top right corner. Each point is equal to 1 South African cent and the score is rounded down to the closest R10. Experiment 4 used exactly the same translation and ranking views and Experiment 3 only removed the money earned label.

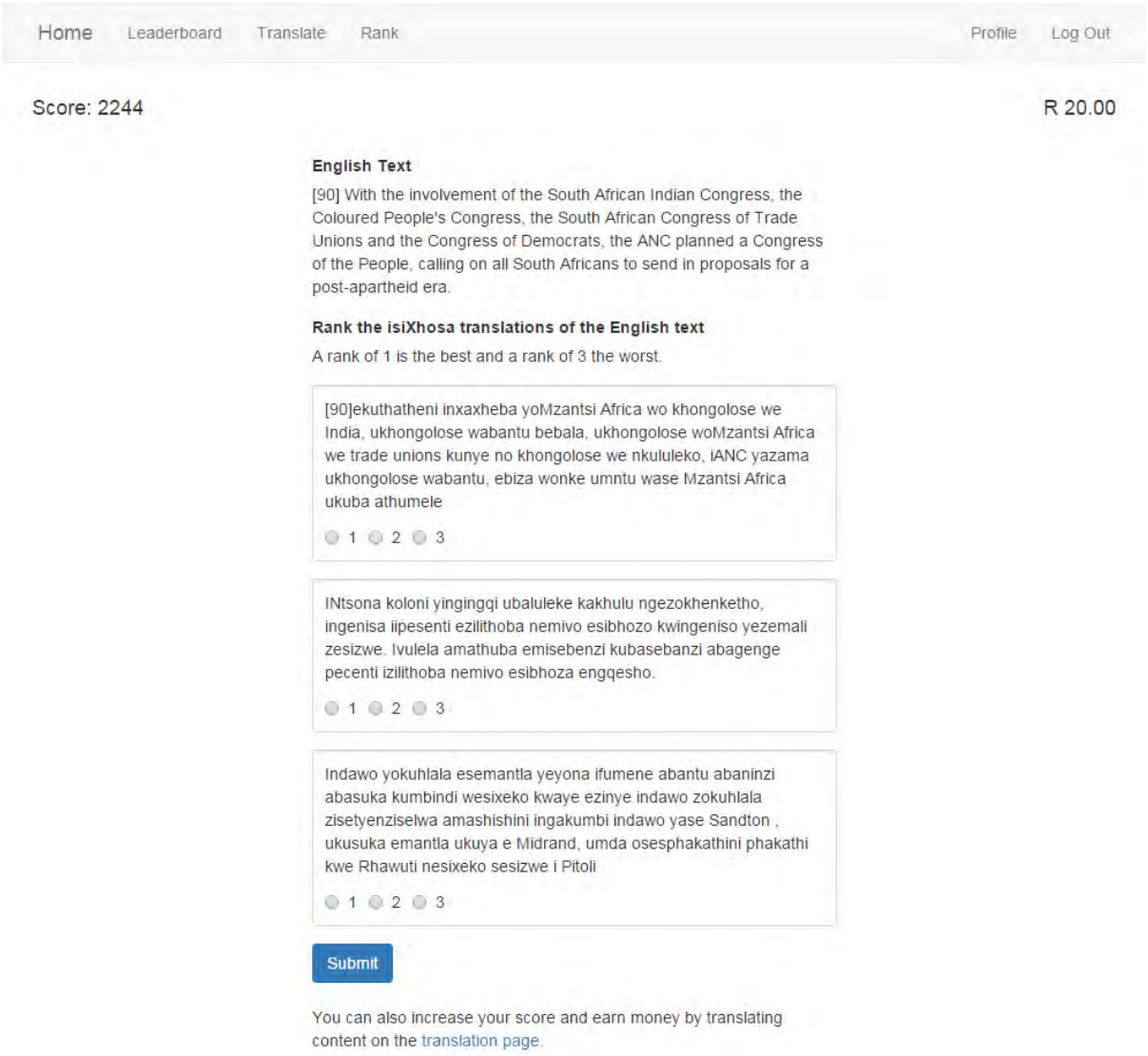

Figure 11: Experiment 2,3 and 4: Screenshot of rank page

The leaderboard page, seen in Figure 12, shows every user's score alongside their display name; the current user's name appears in bold.

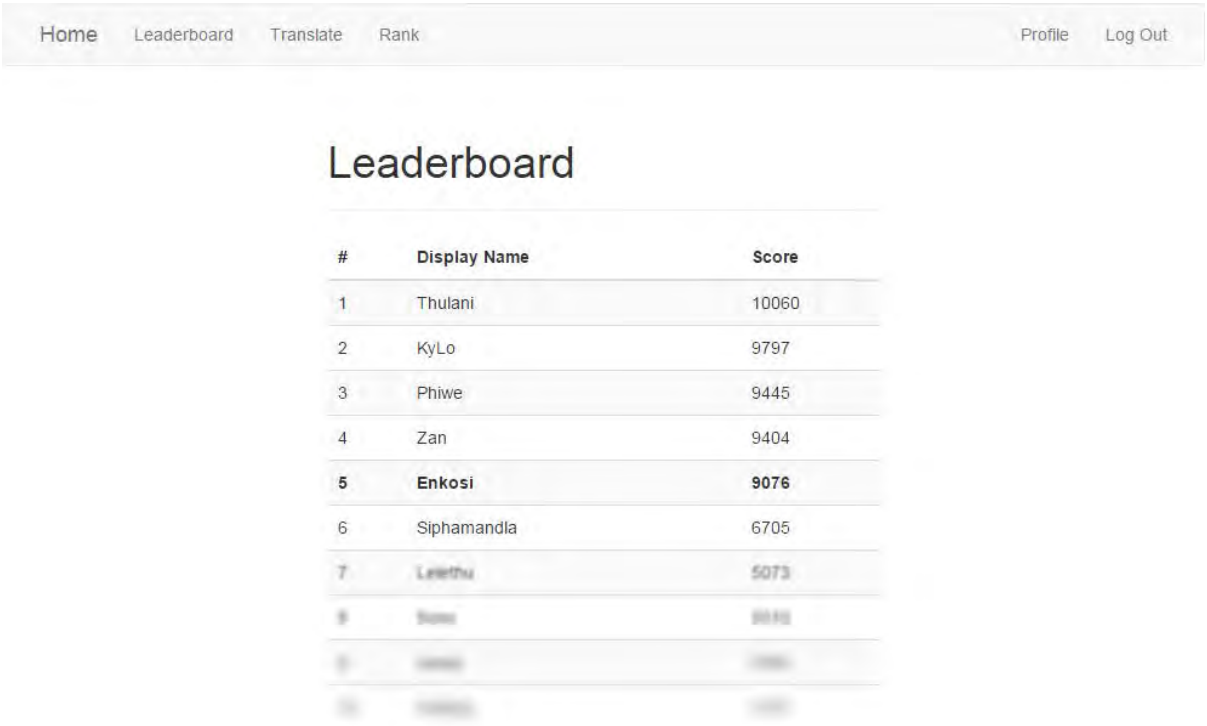

Figure 12: Experiment 2 and 3: Screenshot of leaderboard page

The leaderboard for Experiment 3 had the same design as Experiment 2 but Experiment 4's leaderboard (seen in Figure 13) showed the amount of money the user would earn given their position on the leaderboard.

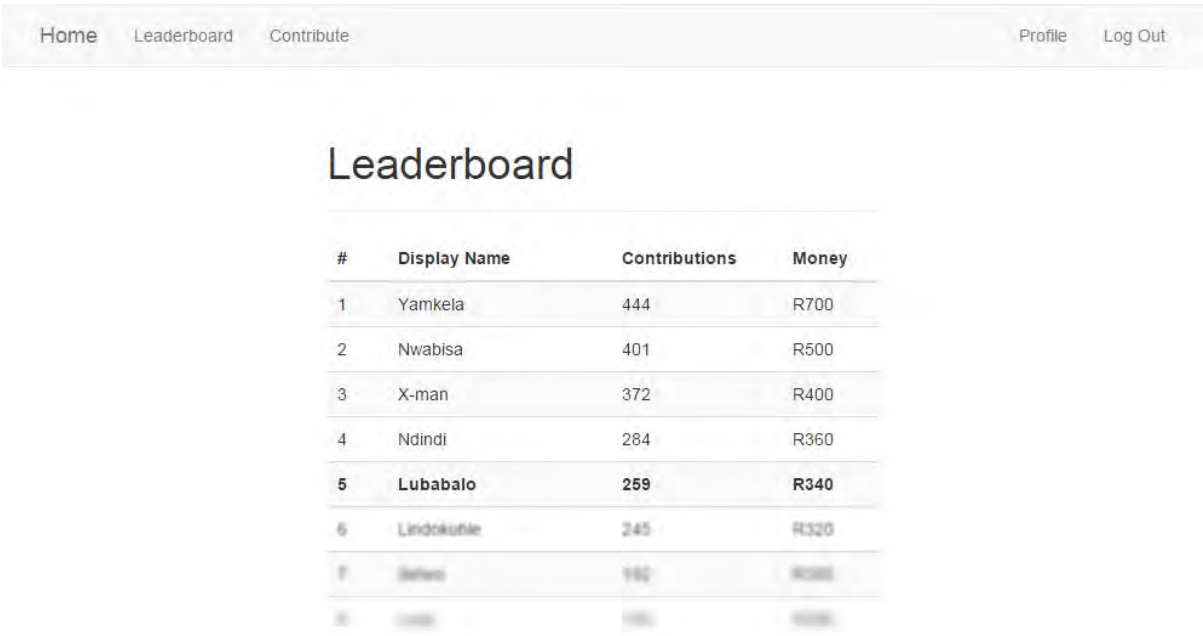

Figure 13: Experiment 4: Screenshot of leaderboard page

3.4. Mobile Payments

It was decided to not use cash or physical coupons to pay users because there could potentially be a large number of students from all over South Africa participating who would return home during the holidays. Mobile payment services, offered by many of South Africa's large banks, have a number of advantages over cash. You only need a valid South African mobile number to receive payments and can withdraw cash from any of the sending bank's branches, ATMs or from a list of authorised partners, such as grocery stores and the recipient is not required to have a bank account. 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 a text message with a transaction summary and instructions on how to withdraw the money.

The mobile payment services offered by the four main banks in South Africa were surveyed and the First National Bank's (FNB) eWallet service was selected (First National Bank 2014). FNB's eWallet has 1 free cash withdrawal a day, no payment fees and the author already had an FNB bank account with online banking to make payments.

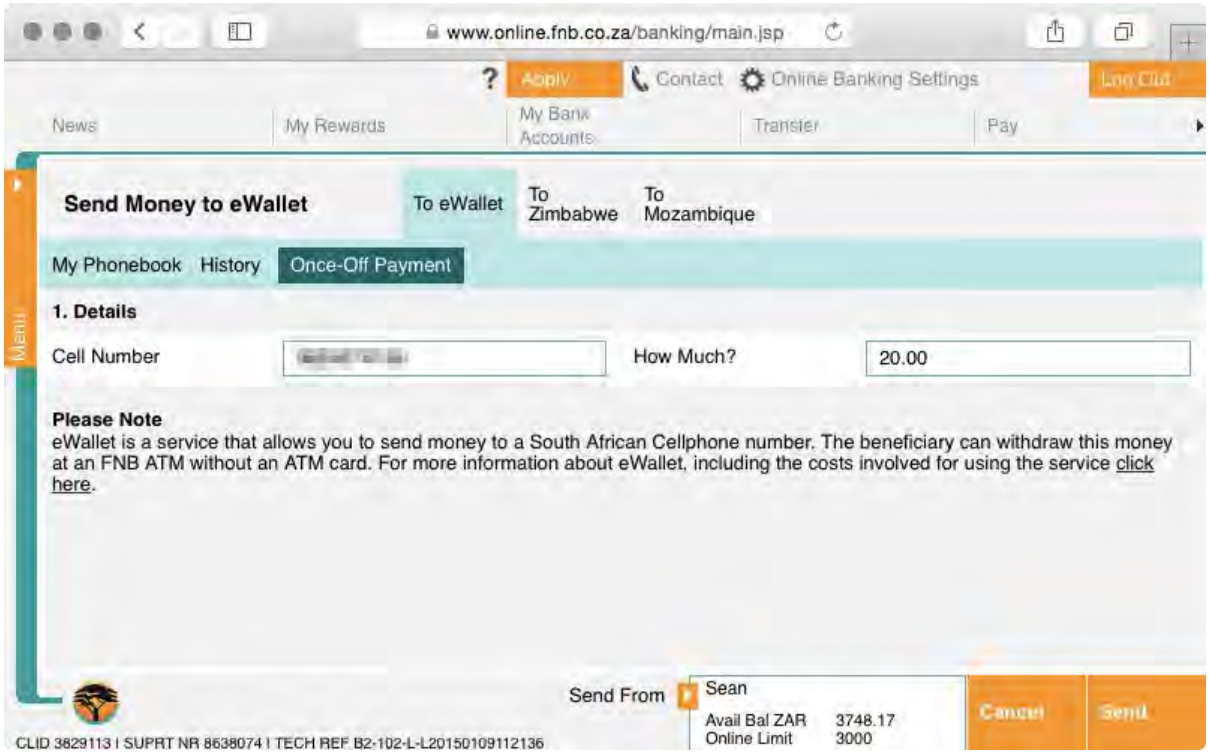

Figure 14: Online banking mobile wallet payment screen.

The payment screen shown in Figure 14 is part of the FNB online banking service. To make a payment you need to enter a valid South African mobile number and an amount. Amounts between R20 and R3000 are allowed in one transaction. Once the payment is sent to the mobile number, the owner will receive a text message and can use a Wireless Application Protocol (WAP) service to check their balance, withdraw cash or even make their own payments to other mobile numbers.

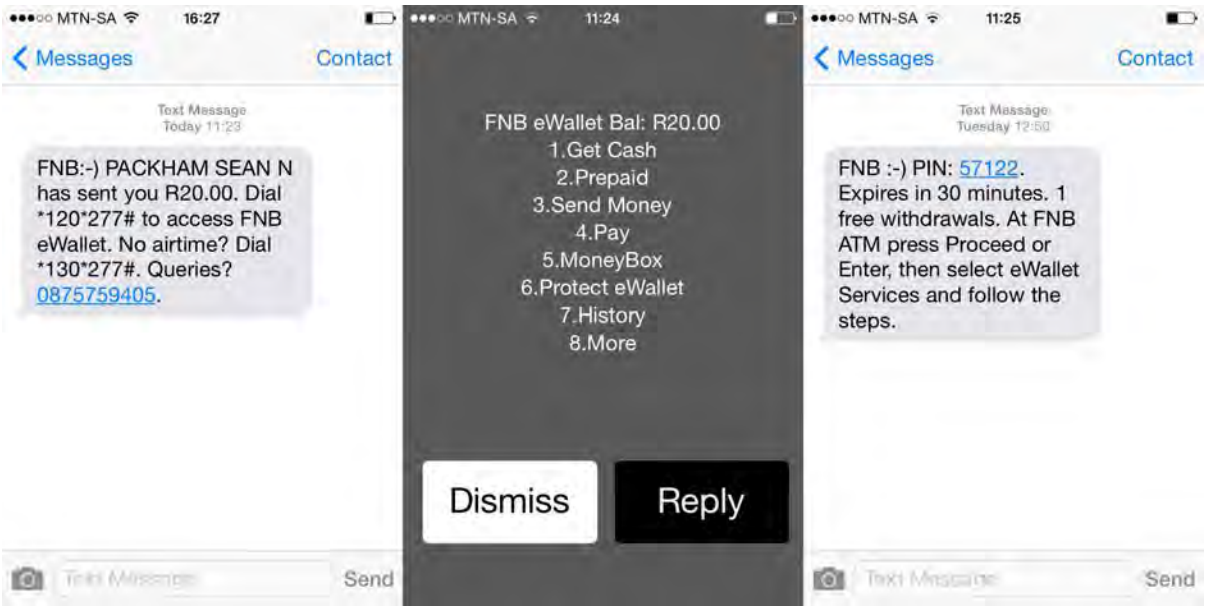

Figure 15: Receiving a mobile payment and withdrawing cash

Figure 15 shows three of the screens a user will see when receiving a payment and withdrawing cash. The first screen on the left shows an example text message detailing the transaction and how to access the money. The middle screen in Figure 15 shows the main menu of the WAP service where users are shown their balance and can perform a number of actions. Selecting option 1 to get cash will send a text message, seen in the last screen on the right in Figure 15. The text message contains a PIN that can be used to withdraw their cash. When withdrawing from an ATM, a user will choose to do a cardless transaction and enter their mobile number to identify their account and the received PIN to authenticate access. They can then choose their withdrawal amount. For security reasons, the PIN is only valid for 30 minutes.

3.5. Summary

Due to the low number of African users on crowdsourcing platforms like Amazon's MTurk and CrowdFlower, alternative systems were explored. Popular open source systems - BOSSA and PyBOSSA - were missing key features required for the experiments so a custom system was built to easily support many scenarios. It was expected that users participating in the experiments would come from all over South Africa and paying rewards in cash would pose a number of administrative and logistical issues, therefore a mobile cardless payment service was chosen as it allowed

rewards to be instantly paid and cash to be easily withdrawn by users from a number of outlets.

Chapter 4 covers the first experiment, run as a pilot study to explore alternative methods for gathering a crowd of bilingual isiXhosa/English speakers. Experiments 2, 3 and 4 are covered in Chapters 5, 6 and 7 respectively.

4. Experiment 1: Pilot Study

Experiment 1 was designed as a pilot study to determine if a crowd of qualified users from Twitter could be motivated to participate purely on the project's intrinsic value rather than by any extrinsic motivators such as financial reward.

4.1. Methodology

The experiment's primary purpose was testing whether users could be gathered from social networks to participate in future crowdsourcing games to translate content for low resource languages like isiXhosa. The experiment website would be shared on Twitter, a social network for sharing short messages called tweets to followers, through various curated tweets at different times of the day. Interested parties would follow the links to the project website, where they could find out more about the project and begin contributing. The custom system covered in Chapter 3.3 was configured in the following way:

- A complete signup process was left out to make contributing as quick and easy as possible. Users who wanted to participate in future experiments could provide their email to be contacted later.
- No pre-assessment test was used. The experiment only wanted to determine interest; translations could be assessed in future experiments.
- A sentence redundancy of 1 was chosen so that each sentence would only be translated once.
- No ranking was used for this experiment.
- No translation limit was set but translations were still tracked to see how many users would contribute.
- The leaderboard was not used.
- A single page design was used so that users could see all the relevant content without having to navigate to additional pages (Seen in Figure 9).
- A goal of obtaining 100 volunteers was set and a custom widget was implemented to show the progress towards achieving the goal (Seen in Figure 8).
- Social network sharing buttons were included for Facebook, Twitter and Google (Seen in Figure 8).

The two expert users used to evaluate Experiment 1's user interface in Section 3.3.5 were also asked to rate five sample tweets on a scale of 1 to 5 to indicate how appealing the wording was to them, where 1 is not appealing at all and 5 is very appealing. The results of the survey can be seen in Table 4. The highest rated tweets were used to share the project website with the author's followers.

Table 4: Experts evaluation of proposed tweets for Experiment 1

Tweet	*A*	*B*
English to isiXhosa translators needed for research to build isiXhosa online services.	1	1
Help researchers build isiXhosa online services by translating English to isiXhosa.	3	2
Need your help translating English to isiXhosa for Exciting new research to build isiXhosa online services.	2	3
Wikipedia in isiXhosa? Make it happen today!	5	4
More online services for isiXhosa speakers? Make it happen today!	4	4

User A commented that tweet 1 sounded like paid work was being offered. User B commented that tweets 1 – 3 were too formal. Both users said they preferred the casual conversational tone of tweets 4 and 5 and that it called them to action.

4.2. Dataset

The dataset for the experiment was assembled by hand picking locally relevant Wikipedia articles of varying lengths and topics. All the articles were broken down into individual sentences, each representing a single translation task. The selected articles and their access dates can be seen in Appendix E.

4.3. Results

Experiment 1 was run from 05/08/2014 – 07/08/14. At the time, the researcher's Twitter network consisted of 132 followers. Table 5 shows the tweets that were sent and the activity they generated. The tweets contain various hashtags, for example #isiXhosa, #UCT, etc. Hashtags are special keywords that label tweets, allowing them to be topically filtered in real-time to enhance discovery. The R column lists the number of retweets a tweet received, the C column lists the number of times the project website link was clicked and the U column lists the number of users who signed up. The tweets that received the most clicks were those that were retweeted and shared with more people. The retweeted accounts had a combined total of +-1500

followers at the time. Although users clicked through to the project website, none attempted to contribute translations.

Table 5: Experiment 1 Tweets and Activity

Date	*Tweet*	*R*	*C*	*U*
15:06 - 5/8/2014	Wikipedia in #isiXhosa? Let's make it happen! http://bit.ly/isiXhosa #isiXhosa #UCT	0	1	0
20:17 - 5/8/2014	Would you like to see more #isiXhosa content online? http://bit.ly/isiXhosa #isiXhosa #UCT #Crowdsourcing	0	0	0
08:54 - 6/8/2014	My 1st experiment, can a crowd of English to #isiXhosa #translators be gathered from Twitter #crowdsourcing https://bitly.com/isiXhosa	3	8	0
15:33 - 6/8/2014	#Wikipedia in #isiXhosa? Let's make it happen! http://bit.ly/isiXhosa #isiXhosa #SouthAfrica #UCT	1	0	0
11:46 - 7/8/2014	Would you like to see more #isiXhosa content online? http://bit.ly/isiXhosa #isiXhosa #SouthAfrica #UCT #Crowdsourcing	0	1	0

There are a number of possible reasons why website visitors did not contribute:

- The researcher's network and the extended networks reached through retweeting did not contain the required demographic of users.
- The Twitter hashtags used were either not popular or were useless in the same way English speakers aren't clicking on the #English hashtag.
- Users who did click through might have done so out of curiosity even though they spoke no isiXhosa.
- None of the tweets were tweeted in isiXhosa.
- The demographic of users being appealed to may not be on Twitter or may not want to contribute for free.

4.4. Summary

We were unable to motivate people to participate Experiment 1 by relying only on the project's intrinsic value and offering no extrinsic motivators such as financial rewards.

5. Experiment 2: Comparing rewards

A crowd of users could not be gathered from the researcher's Twitter network in the pilot study. Bilingual students from the University of Cape Town were the target users for Experiment 2. The experiment was designed to answer the first research question.

5.1. Methodology

The custom system covered in Chapter 3.3 was configured in the following way:

- The full signup process was used.
- Four pre-assessment multiple choice questions were used and users had to get at least three correct.
- 6 user groups were configured (Section 5.1.1).
- Each group had its own high-score leaderboard.
- A translation redundancy of 3 was configured (Section 5.1.3).
- A ranking redundancy of 3 was configured (Section 5.1.4).
- Users were limited to contributing 100 translations.
- Users were limited to contributing 100 rankings.
- The points system equated directly to money earned (Section 5.1.7)
- A multi-page design was used, as users could perform more tasks than was possible in the pilot study and a single page design would be cluttered and difficult to use.

5.1.1. Groups

The experiment was designed to test the effect of paying users consistent, increasing or decreasing rewards per task. The consistent and increasing rewards were chosen to mirror salary/wage and commission based earnings. Offering increasing rewards per task also mirrors the increasing reward schemes used in games, where players receive greater rewards the longer they play. Decreasing rewards per task were added to act as a sanity check to compare the other two schemes against, it was expected that decreasing rewards per task would not motivate people to contribute as well as the other two schemes. There were two types of tasks - translation tasks (Section 5.1.3) and ranking tasks (Section 5.1.4) - each with their own leveling scheme. Two leveling schemes were used to determine when rewards would change (except for users

receiving consistent rewards). The first scheme changed rewards after consistent intervals of work effort, while the second changed rewards after increasing intervals of work effort. Table 6 shows a summary of the resultant six groups with their reward type and work effort leveling scheme.

Table 6: Effort and reward schemes in Experiment 2

Group	*Effort Required*	*Rewards per Level*
Group 1 (CC)	Consistent	Consistent
Group 2 (IC)	Increasing	Consistent
Group 3 (CI)	Consistent	Increasing
Group 4 (II)	Increasing	Increasing
Group 5 (CD)	Consistent	Decreasing
Group 6 (ID)	Increasing	Decreasing

The consistent effort groups required the same number of tasks to be performed at each level to progress to the next level while the increasing effort groups required more tasks to be performed at subsequent levels to progress to the next level. The consistent reward groups received the same reward per task at each level, while the increasing reward groups' reward per task increased when moving to subsequent levels. Groups 1 and 2 were designed as a baseline, offering consistent rewards from consistent and increasing effort. Groups 3 and 4 offered increasing rewards and 5 and 6 decreasing rewards from consistent and increasing effort. Section 5.1.7 details the design of the levels and the final number of tasks required at each level.

All the groups limited user contributions to 100 translations and 100 rankings. Setting a limit allowed predictable payment values to be calculated for each group. Task payment points were first chosen for the consistent reward groups and adjusted for the increasing and decreasing reward groups. The increasing reward groups were adjusted to start at a lower rate and end at a higher rate. The decreasing reward groups were adjusted to start at a higher rate and end at a lower rate. This design allowed all the groups to have the same average reward per task if both the translation and ranking task limits were reached. This design created a fairly predictable reward system where rewards cannot spiral out of control. Furthermore, a user in each of the 6 groups has the potential to earn as much as any other user in another group if they reach the translation and ranking task limits.

5.1.2. Users

There were a number of possible channels to recruit users for the experiment from the University of Cape. The primary channel was the university's "All Students" email list, which contains email addresses for all currently registered students and is often used to share student research projects in need of participants. The number of students registered at UCT in 2014 was 26,332, but the demographic data was only publicly available for 2013. In 2013 a total of 26,116 students were enrolled, of which 6,199 potentially spoke African languages. This number would include speakers of all African languages, not only isiXhosa. Despite isiZulu being spoken by more people nationally, there are a greater number of isiXhosa speakers than isiZulu in UCT's Western Cape Province. The email to all the students called for bilingual English and isiXhosa speakers to participate in an online translation game for 1 week. The email was required to be brief and contain no images. The full email specified various details relating to payment, qualification, time, privacy, the researchers and ethical clearance and is viewable in Appendix A. Finally, it contained a link to an online registration form, which can be seen in Appendix B.

If a sufficient number of users could not be gathered via the "All Students" email list, a similar appeal would be made on the University's learner management portal. The portal allows research projects to be shared with students on its home page via advertisements, which have to be a specific size but are allowed to contain images, potentially giving them an advantage over plain text emails. There are many competing advertisements in the system at one time, all sharing airtime with one another. Furthermore, advertisements are only shown to unauthenticated students. A student might not login for a few days or weeks and when they do they might be served a competing advertisement. Using the "All Student" email option at least guarantees that every student will receive the call for participation at least once. Lastly, the experiment can be marketed directly to the university's languages department. If neither of these options provided a sufficient number of users, it was decided that the experiment would be scaled down.

5.1.3. Translating

Translating English Text into isiXhosa was the first task users could perform to contribute to the project. Users were given one sentence at a time to translate.

Sentences were extracted from chosen Wikipedia articles and inserted into a translation queue. As the sentences were translated, an isiXhosa article was gradually built up. This approach has the advantage of leaving a partially translated article in a useful state, rather than randomly assigning sentences to users and being left with chunks of randomly translated articles. Once all the sentences in an article were translated, a new article was taken from a queue of articles. When a sentence was assigned to a user for the first time, it was tagged with the user's group. A sentence was considered completely translated when it had received the required number of redundant translations by different users from the same group. When a user requested to translate a sentence, they were either given one that had already been assigned to their group and that had not received the required number of redundant translations and had not already been translated by them or they received a new sentence. A user could only translate each sentence once and could not edit their translation. Once a user had translated a sentence, their total score increased. Translations were limited to 100 per user to prevent a small group of users contributing the majority of the corpus and exhausting the budget. Once a user reached the translation limit they could no longer translate, they were thanked for their contributions and encouraged to continue ranking other user translations.

5.1.4. Ranking

Ranking other user translations was the second task users could perform. Ranking required users to assess and order redundant isiXhosa translations in their perceived order of correctness, where 1 is most correct and 3 (redundancy level) is least correct. Ranking was designed in a similar way to translating; when a user requested to rank a sentence, they were given one from their group that had received the required number of redundant translations but had not yet received the required number of redundant rankings and had not been ranked by them already. If no sentence matched the criteria, the user was asked to check back later and encouraged to continue translating. Once a user reached the limit they could no longer rank but were encouraged to continue translating. Once a sentence and its translations had been completely ranked, the translation with the lowest rank total is flagged as the model translation. For example, if all the users agreed and ranked the same translation first, that translation will be the model translation because it will have the lowest rank total of 3 = 1 + 1 + 1. Ranking can produce more than one model answer if two or more translations achieve the same total rank.

5.1.5. Qualifying

Before potential users could participate, they needed to pass four pre-assessment multiple-choice questions that assessed their English to isiXhosa translation skills. The questions were translated by a handful of the first users who signed up for the experiment. They were offered a once-off reward to translate four groups of three sentences. The translations with the most overlap were considered the most correct and were verified with an external isiXhosa speaker. The sentences and their translations used in the experiment can be seen in Table 7.

Table 7: Qualifying assessment sentences and their translations

Group	*English*	*isiXhosa*
1	I like you.	Ndiyakuthanda.
1	I don’t like you.	Andikuthandi.
2	The ball is smaller on Mondays.	Ibhola incinci ngeMivulo.
2	The ball is bigger on Mondays.	Ibhola inkulu ngeMivulo.
3	He gave her his book today.	Umnike incwadi yakhe namhlanje.
3	He gave her his book yesterday.	umnike incwadi yakhe izolo.
3	He will give her his book tomorrow.	Uzakumnika incwadi yakhe ngomso.
4	Green is her favourite colour.	Uluhlazza ngumbala wakhe oyena amthandayo.
4	Green was her favourite colour.	Uluhlazza yayingumbala amthandayo.
4	Green is not her favourite colour.	Uluhlazza ngumbala angamthandiyo.

5.1.6. Payment Model

Professional translation services charge per word for translation jobs but the experiment was designed for users to translate sentences with varying length and get rewarded per sentence. Choosing the optimal payment point for translation and ranking tasks took considerable investigation, a number of scenarios were simulated and a payment model that could easily be adjusted was created. Care had to be taken to ensure that a number of criteria were met for the project and its users. If translation costs were too high, it would be more affordable and possibly more efficient to use professional services.

Reward amounts were chosen that were at least comparable to the minimum hourly rate for various local industries. The minimum hourly rate and the average cost of professional translation services represented a range of acceptable rewards. The mywage.co.za online service was used as the primary source for hourly wage rates in

South Africa. The service covers a wide range of sectors, such as farming, forestry, hospitality, transport, domestic, security and retail. Knowing that the primary audience for the experiment would be undergraduate students from the University of Cape Town, the education level was mapped to the 2014 minimum hourly rate for general admin workers at ZAR 12.71/hour.

Various local and international professional translation services were contacted via email for a quote to translate the English “Cape Town” Wikipedia article into isiXhosa. A summary of the quotations, given as a rate per word, can be seen in Table 8. The “Cape Town” article was one of the larger articles selected for the experiment. A number of the professional services categorised the article as one with above average complexity. The quoted rates ranged from a high of ZAR 4.42 per word from an international service to a low of ZAR 0.85 per word from a local service. At the exchange rate at the time of 11.04 ZAR/USD and 17.72 ZAR/GBP, the average word cost came to ZAR 2.22 per word.

Table 8: Quotations for translating the English "Cape Town" Wikipedia article into isiXhosa

Company	*USD*	*GBP*	*ZAR*
South Africa Translation Service 1			0.95
South Africa Translation Service 2			0.85
Unites States Translation Service	0.40		4.42
United Kingdom Translation Service		0.19	3.35

Experiment 2’s payment model has been split into two figures. Figure 16 contains all the input variables and a translation matrix indicating the total number of translations that could be achieved given the experiment constraints. Input values could easily be adjusted so that a balance could be struck between attractive rewards for users and affordable translation costs for the experiment.

Project	
ZAR/USD	11.04
ZAR/GBP	17.72
Budget	R 20,000.00
T Dups	3
R Dups	3
Words/T	22.47
Cheaper	1675%

Participants	
T Time	150
R Time	90
Ts	100
Rs	100
Groups	6
Ps/Group	34
Ts/Group	1118
Minutes	400
Reward	R 99.36
Per Hour	R 14.90

T USD	T ZAR	T Dup	Translations				
0.01	0.11	0.33	30193	20129	15097	12077	10064
0.02	0.22	0.66	20129	15097	12077	10064	8627
0.03	0.33	0.99	15097	12077	10064	8627	7548
0.04	0.44	1.32	12077	10064	8627	7548	6710
0.05	0.55	1.66	10064	8627	7548	6710	6039
0.06	0.66	1.99	8627	7548	6710	6039	5490
0.07	0.77	2.32	7548	6710	6039	5490	5032
0.08	0.88	2.65	6710	6039	5490	5032	4645
0.09	0.99	2.98	6039	5490	5032	4645	4313
0.10	1.10	3.31	5490	5032	4645	4313	4026
0.11	1.21	3.64	5032	4645	4313	4026	3774
0.12	1.32	3.97	4645	4313	4026	3774	3552
0.13	1.44	4.31	4313	4026	3774	3552	3355
0.14	1.55	4.64	4026	3774	3552	3355	3178
0.15	1.66	4.97	3774	3552	3355	3178	3019
		R Dup	0.33	0.66	0.99	1.32	1.66
		R ZAR	0.11	0.22	0.33	0.44	0.55
		R USD	0.01	0.02	0.03	0.04	0.05

Figure 16: Payment model: Input variables and payment points

The values in the “T USD” column are translation rewards per sentence in USD and the “R USD” row are ranking rewards per sentence in USD. The specific values come from the various literature and online surveys conducted in Section 2.2.2. Additional values were added at the higher end of the range to simulate the effect higher rewards would have on the model’s various outputs. Changing the “ZAR/USD” exchange rate variable under the project section would re-calculate the values in the “T ZAR” column (translation reward per sentence in ZAR) and the “R ZAR” row (ranking reward per sentence in ZAR). This in turn would re-calculate the “T Dup” column (translation reward per sentence in ZAR taking into account the translation redundancy) and the “R Dup” row (ranking reward per sentence in ZAR taking into account the ranking redundancy). Translation and ranking redundancy are set under the Project section in the “T Dups” and “R Dups” cells. With the cost of translation and ranking and the project budget, set in cell “Budget”, and the average number of words per sentence, set in the “Words/T” cell, a translation word cost could be calculated and compared to that of professional translation services. The average word count of sentences from the chosen articles was around 21 words per sentence. Selecting a translation payment point of ZAR 0.66 and a ranking payment point of ZAR 0.33 on the translation matrix would result in 6710 sentences translated. At this configuration of the payment model, approximately 150,000 words would be

translated if the entire budget of ZAR 20,000 were used. This would be 1675% cheaper than the average rate offered by the sampled professional translation services and 640% cheaper than the lowest rate.

The average time it takes a user to translate and rank one sentence can be set in the “T Time” and “R Time” cells respectively. Five lab assistants were asked to translate English sentences, with a similar average word count as the test data, into their second language. Despite the wide range of languages, the times to translate and rank were similar. The average translation and ranking times were around 90 seconds. These times were input into the payment model and an hourly rate above the target minimum wage was achieved. The “Minutes” value is an approximate time for how long it would take a user to translate “Ts” sentences and rank “Rs” translations. The “Reward” value is the total reward a user could earn if they translate “Rs” sentences and rank “Rs” translations. With the calculated “Minutes” and “Reward”, an hourly wage could be calculated. To simulate a slower user, the time to translate was changed to 150 seconds, which resulted in a wage of ZAR 14.97/hour, still above the target minimum wage. The “Ps/Group” cell indicates how many users are needed per group to reach the translation target chosen in the translations matrix. Similarly, “Ts/Group” shows how many translations are needed per group, to reach the translation target.

In Figure 17, the first matrix shows the resulting word cost at the selected translation and ranking rewards chosen in Figure 16. The second matrix shows the number of users required if they all translate the maximum translations allowed at the selected translation and ranking rewards. This is a best-case scenario; realistically the actual number required will depend on the activity of the users. Therefore, to be on the safe side, the target user goal was set at 150% of the minimum user number and would be adjusted in future if necessary.

ZAR / Word					Participants				
0.03	0.04	0.06	0.07	0.09	906	604	453	362	302
0.04	0.06	0.07	0.09	0.10	604	453	362	302	259
0.06	0.07	0.09	0.10	0.12	453	362	302	259	226
0.07	0.09	0.10	0.12	0.13	362	302	259	226	201
0.09	0.10	0.12	0.13	0.15	302	259	226	201	181
0.10	0.12	0.13	0.15	0.16	259	226	201	181	165
0.12	0.13	0.15	0.16	0.18	226	201	181	165	151
0.13	0.15	0.16	0.18	0.19	201	181	165	151	139
0.15	0.16	0.18	0.19	0.21	181	165	151	139	129
0.16	0.18	0.19	0.21	0.22	165	151	139	129	121
0.18	0.19	0.21	0.22	0.24	151	139	129	121	113
0.19	0.21	0.22	0.24	0.25	139	129	121	113	107
0.21	0.22	0.24	0.25	0.27	129	121	113	107	101
0.22	0.24	0.25	0.27	0.28	121	113	107	101	95
0.24	0.25	0.27	0.28	0.29	113	107	101	95	91

Figure 17: Payment model: Word cost and users matrix

At a glance, the model allows various scenarios to be simulated with minimal effort. Selecting the lowest translation and ranking reward of USD 0.1, seen in some of the literature surveys, allows 30,193 sentences to be translated with the available budget. This equates to approximately 6.7 million words translated, at a translation cost per word of ZAR 0.03. These are attractive targets but they come with challenges; the minimum number of users required is raised to 906 and, when increased by 150% as was done previously to estimate a safer user target, it is raised to 1359 users. At these low rewards, the hourly rate drops to ZAR 3.31, way below the target minimum wage, but if users could do the work 4.5 times faster these rates would be acceptable but not attractive.

5.1.7. Final Rewards

The developed payment model was used to choose the final reward for the six groups. The work effort based leveling scheme required users to contribute a number of tasks before progressing to the next payment level. For Groups 1, 3 and 5, the consistent effort required per level was set to 20 translations and 20 rankings, which gave 5 levels with the chosen translation limit of 100. Groups 2, 4 and 6 required their 5 levels to also add up to 100 tasks but each subsequent level had to have more tasks allocated to it. The first level began with 16 tasks and each subsequent level required 2 more tasks. This resulted in a 50% increase in effort required between the first and the last level.

Table 9: Group 1: Consistent effort and consistent rewards.

Contributions/ level (Effort)	*ZAR/Translation*	*ZAR/Ranking*
20	0.66	0.33
20	0.66	0.33
20	0.66	0.33
20	0.66	0.33
20	0.66	0.33

Table 10: Group 2: Increasing effort and consistent rewards.

Contributions/ level (Effort)	*ZAR/Translation*	*ZAR/Ranking*
16	0.66	0.33
18	0.66	0.33
20	0.66	0.33
22	0.66	0.33
24	0.66	0.33

Table 10 and Table 9 show the final leveling system with rewards for the consistent reward Groups 1 and 2. It should be noted that the levelling system had no effect on the consistent reward groups, as users would receive the same reward at each level, but they were designed first so that the increasing and decreasing reward groups could easily be derived from them. At the chosen rates, Group 1 and 2 users could earn a total of ZAR 99.00: ZAR 66.00 for translating and ZAR 33.00 for ranking.

Table 11: Group 3: Consistent effort and increasing rewards.

Contributions/ level (Effort)	*ZAR/Translation*	*ZAR/Ranking*
20	0.58	0.25
20	0.62	0.29
20	0.66	0.33
20	0.70	0.37
20	0.74	0.41

Table 12: Group 4: Increasing effort and increasing rewards.

Contributions/ level (Effort)	*ZAR/Translation*	*ZAR/Ranking*
16	0.58	0.25
18	0.62	0.29
20	0.66	0.33
22	0.70	0.37
24	0.74	0.41

Groups 3 and 4 required subsequent levels to offer increasing rewards while still maintaining an average translation and ranking reward per sentence close to ZAR 0.66 and ZAR 0.33. Table 11 and Table 12 show the final leveling system with rewards for Groups 3 and 4. Rewards began at lower than average amounts and ended at higher than average amounts; translations started at ZAR 0.58 and ended at ZAR 0.74 and rankings started at ZAR 0.25 and ended at ZAR 0.41. Group 3 users could earn a total of ZAR 99.00: ZAR66.00 for translating and ZAR 33.00 for ranking. Group 4 users could earn a total of ZAR 100.60: ZAR66.80 for translating and ZAR 33.80 for ranking.

Table 13: Group 5: Consistent effort and decreasing rewards.

Contributions/ level (Effort)	*ZAR/Translation*	*ZAR/Ranking*
20	0.74	0.41
20	0.70	0.37
20	0.66	0.33
20	0.62	0.29
20	0.58	0.25

Table 14: Group 6: Increasing effort and decreasing rewards.

Contributions/ level (Effort)	*ZAR/Translation*	*ZAR/Ranking*
16	0.74	0.41
18	0.70	0.37
20	0.66	0.33
22	0.62	0.29
24	0.58	0.25

Groups 5 and 6 required subsequent levels to offer decreasing rewards while still maintaining an average translation and ranking reward per sentence close to ZAR 0.66 and ZAR 0.33. Table 13 and Table 14 show the final leveling system with rewards for

Groups 5 and 6. The reward amounts used for Groups 3 and 4 were reversed, starting at a higher than average amount and ending at a lower than average amount. Group 5 users could earn a total of ZAR 99.00: ZAR 66.00 for translating and ZAR 33.00 for ranking. Group 6 users could earn a total of ZAR97.40: ZAR65.20 for translating and R32.20 for ranking.

Group 6 had a slightly lower total of ZAR 97.40 and Group 4 had a slightly higher total of R100.60 than the remaining groups, which each total up to ZAR 99.00. Users were told that, if they reached both the translation and ranking limit, they would have their reward rounded up to ZAR 100.00.

5.1.8. Dataset

The same dataset used in the first experiment was reused and expanded with additional articles. The selected articles and access dates can be seen in Appendix E.

5.1.9. Summary

The development of the payment model greatly improved the efficiency of simulating various reward scenarios. A consistent translation and ranking reward was chosen that struck a balance between achieving an affordable translation cost and offering attractive rewards for users. The rewards for the increasing and decreasing groups were derived from the consistent reward groups; care was taken to ensure a user in each group could earn the same reward but at different rates. Section 5.2 details the running and results of Experiment 2.

5.2. Results

Experiment 2 was run for a week from Wednesday 19 November 2014 to Wednesday 26 November 2014. These dates were selected so that the experiment would begin after the final November exams had finished. The experiment was advertised to all UCT students on 12 November by means of the UCT "All Students" email list. By 19 November 2014, 333 students had signed up for the experiment. The number of users met the minimum user goal so none of the other channels were used. Of the 333 who initially showed interest, only 201 users created an account; 172 identified themselves as UCT students and the rest did not indicate or discovered the experiment by other means. Only one user did not pass the pre-assessment with a score of 2/4, 141 users scored 4/4 and 59 scored 3/4.

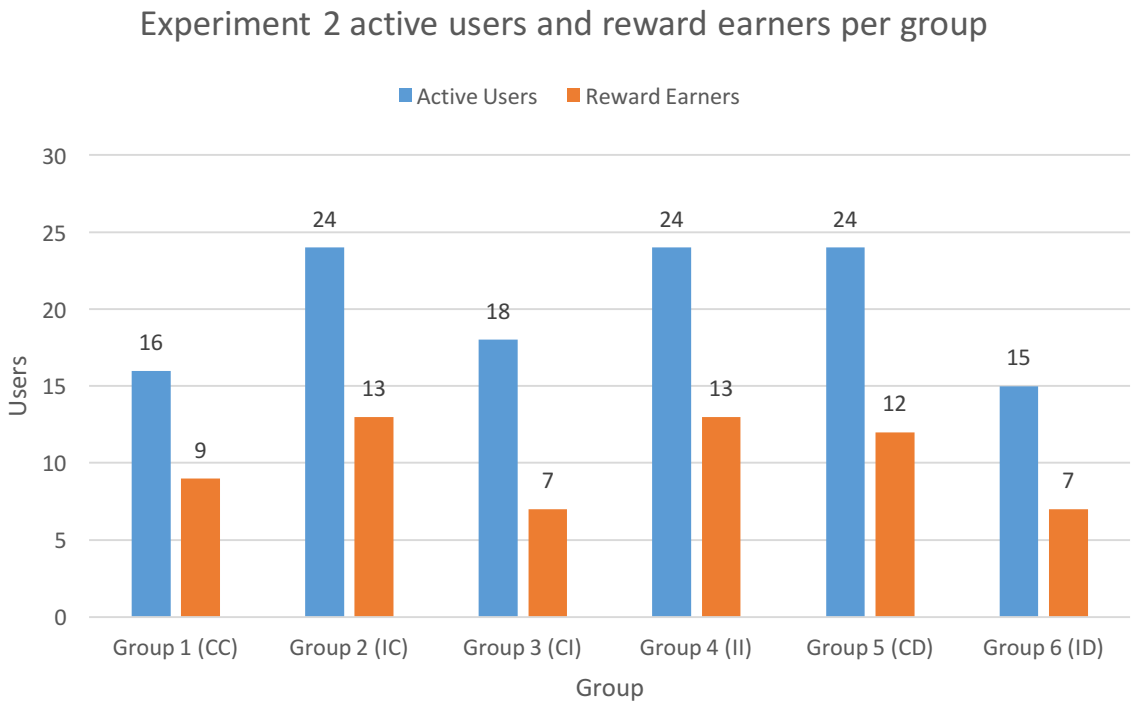

Figure 18: Experiment 2: Active users and reward earners per group

The distribution of users amongst the groups was fairly even; Groups 1, 2, 5 and 6 had 33 users and Groups 3 and 4 had 34 users. The reason Groups 1 and 2 had fewer was because they each had a test user, Group 5 had one user who did not qualify and Groups 5 was the last group to get a 34th user, leaving group 6 at 33. The number of active users and reward earners per group can be seen in Figure 18. All groups

exhibited a similar behaviour, in that half of the active users earned a reward. This gives us a formula to predict future participation irrespective of payment scheme implemented.

A total of 3600 translations and 2589 rankings were contributed; these figures include redundant contributions. A total of 1088 sentences received 3 redundant translations; 734 sentences received 3 redundant rankings and could be reassembled into isiXhosa articles. Of the 200 qualified users, 121 contributed at least one translation or ranking, 105 users translated, 84 ranked and 68 did both. 21 users reached the translation limit, 4 reached the ranking limit and 3 reached both, earning the highest reward tier shown in Figure 19. Over half the active users did not earn a reward and the most earned reward was the lowest tier. Receiving some points is not enough for 50% of the users to be motivated to contribute enough to reach the first reward level. Future studies could reduce the amount of work required per level and increase the number of levels, so that users receive rewards sooner and more frequently, thereby increasing their motivation to continue contributing.

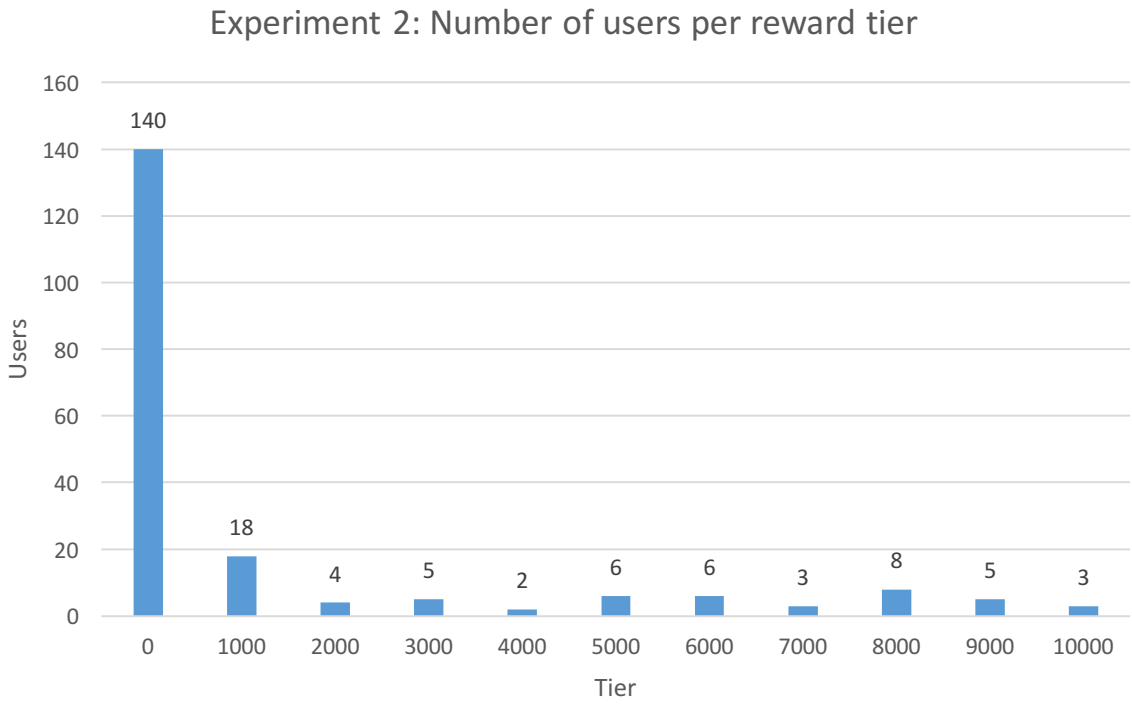

Figure 19: Experiment 2: Number of users per reward tier.

The rest of the chapter will analyse the experiment data to determine the effect the various payment groups had on contribution quantity and user motivation by analysing the interval in-between translations and ranks as an indicator of motivation. For Experiment 2, the system was not designed to explicitly record task duration; this feature was added to the system from Experiment 3 onwards. Before the data could be analysed, it was pre-processed to detect and exclude contributions from suspected cheaters (Section 5.2.1). Afterwards, the legitimate user activity is examined (Section 5.2.2) and finally three comparisons are performed. The first comparison looks at the combined effect of work effort and reward type of the six groups on contribution quantity and user motivation (Section 5.2.3.1). The second comparison ignores the possible effect of reward type, and examines the effect of work effort on contribution quantity and user motivation (Section 5.2.3.2). The last comparison ignores the possible effect of work effort, and examines the effect of reward type on contribution quantity and user motivation (Section 5.2.3.3). The chapter ends with a summary and discussion of the experiment findings (Section 5.3).

5.2.1. Pre-processing Data

It was expected that an experiment that required users to contribute original content while competing for standing and financial rewards would result in cheating but no explicit warning discouraging cheating was given. All 6 groups had translation and rank cheaters. The presence of cheaters lowered the number of legitimate active users for each group. Cheaters may have further impacted the experiment by decreasing the motivation of legitimate users, discouraging them from competing when it appears, despite their legitimate efforts, they cannot catch up to users higher up on the leaderboard. A total of 7 cheaters were found, which meant on average you would get one cheating user for every 9 reward earners.

A two-phase approach was used to detected cheaters. Firstly, users and their content were flagged by searching for duplicate, non-isiXhosa or gibberish translations. Duplicate content was easy to find programmatically but the researcher had to search for non-isiXhosa and gibberish content manually. Secondly, the intervals between successive translations were analysed to look for exceptionally fast translations. Ranking cheaters were detected by searching for repeating ranking patterns e.g. 1,2,3 used for successive rankings. No legitimate translations were found to have intervals

shorter than 20 seconds and no legitimate ranks were found to have intervals shorter than 10 seconds.

Figure 20 shows the total translations for each group, with a breakdown of legitimate and suspected cheat translations. Group 1 had the most cheat translations submitted, and Group 4 had the least. In future experiments, a much stricter and explicit policy against cheaters was given to users.

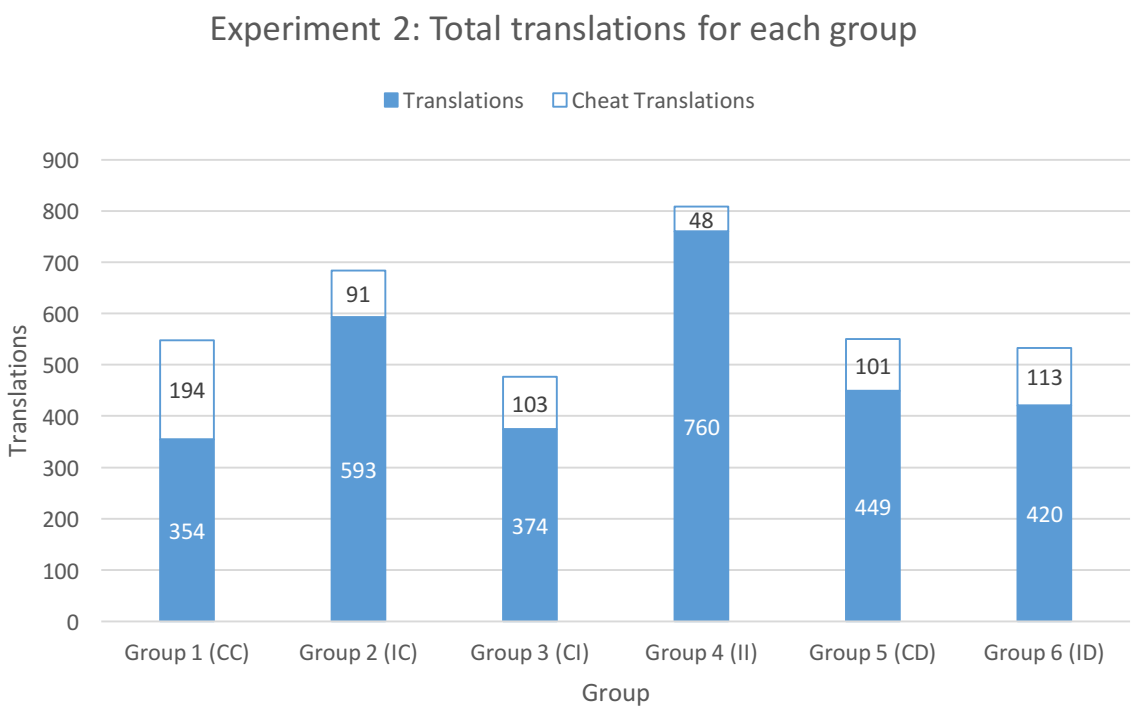

Figure 20: Experiment 2: Total translations for each group.

The total ranks for each group, with a breakdown of legitimate and suspected cheat ranks, can be seen in Figure 21. People were less likely to cheat while ranking, most likely because it still took longer to order three translations than it took to paste in some copied text.

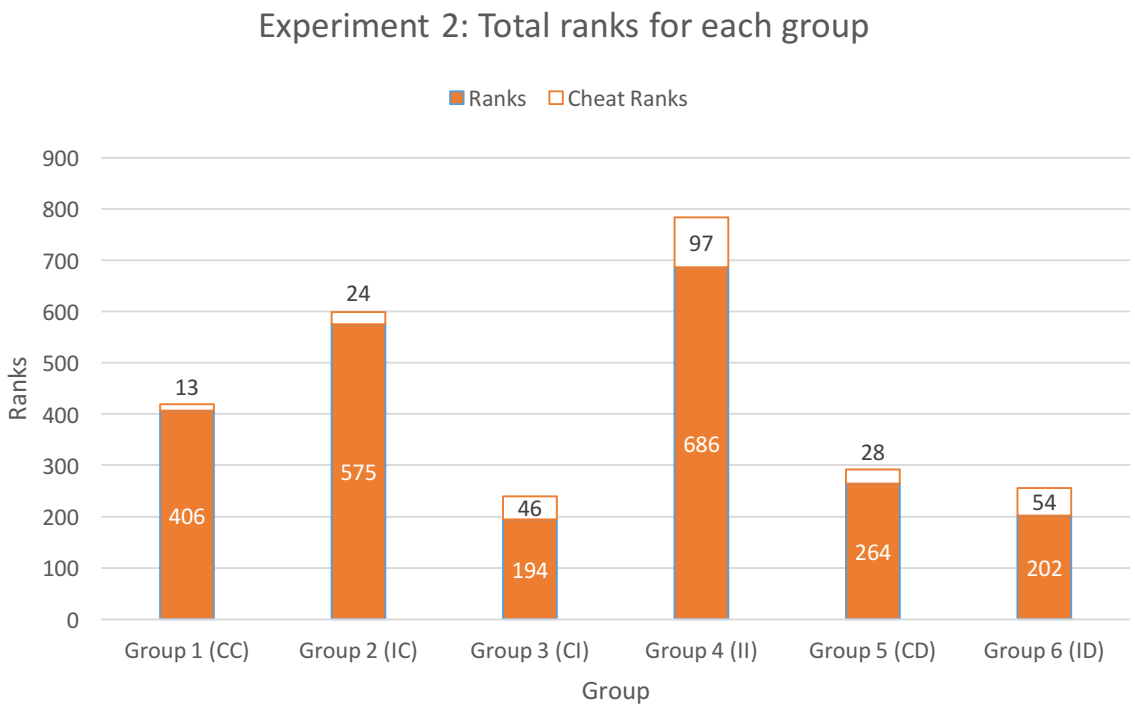

Figure 21: Experiment 2: Total ranks for each group

All the cheat translations and ranks can be seen in Figure 22. Although Group 4 had the most cheat ranks, the researcher was satisfied with the 10 second cut-off point as the minimum amount of time required to read the source sentence and three translations and order them.

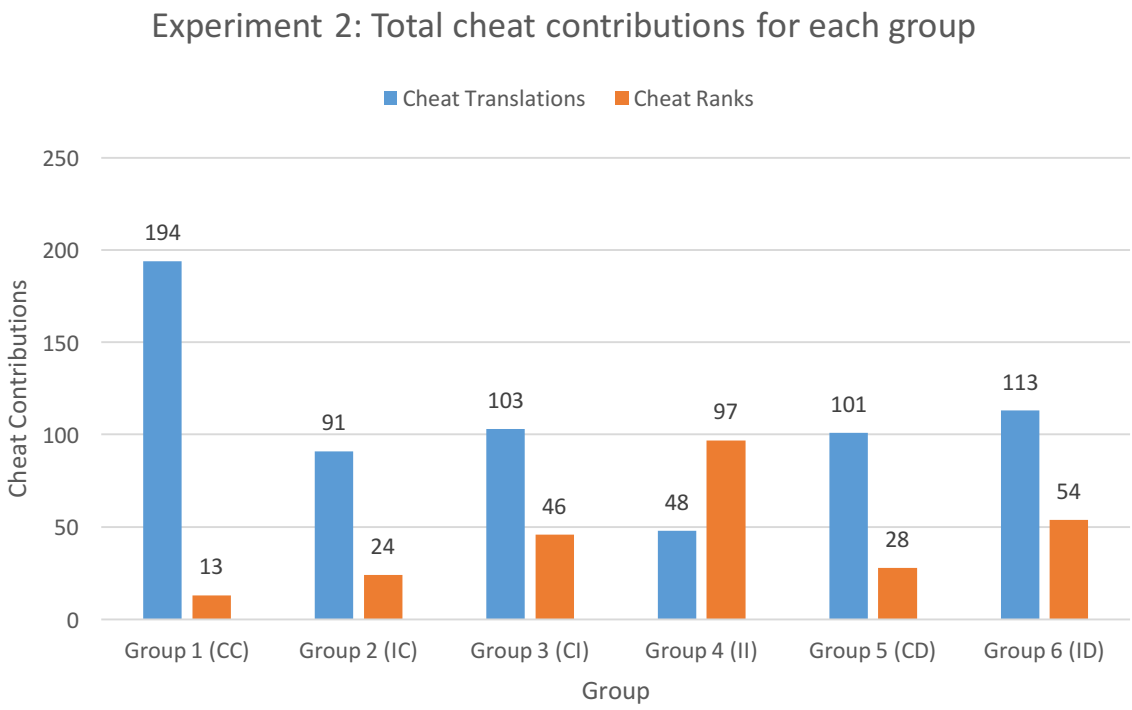

Figure 22: Experiment 2: Total cheat contributions for each group

Figure 23 combines the total legitimate translations and ranks from the previous two charts. Group 1 has more ranks than translations because of the large number of cheat translations contributed. Group 3 users were the least active but not far behind Groups 1, 5 and 6.

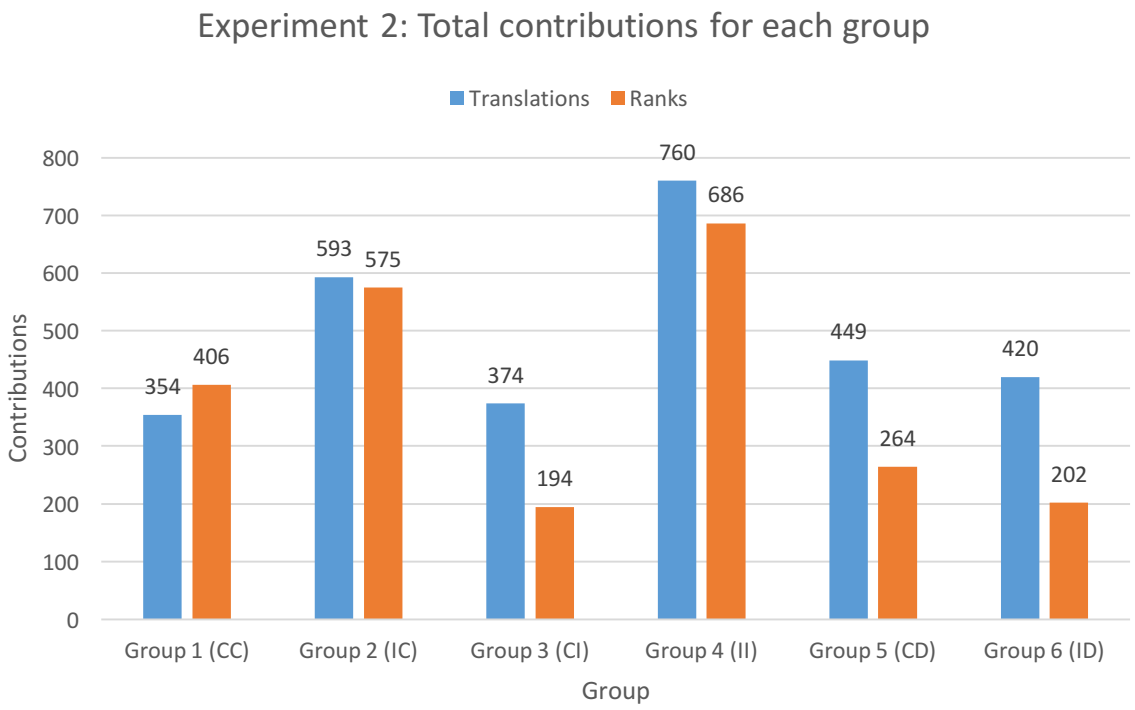

Figure 23: Experiment 2: Total contributions for each group

Group 4 stands out the most from the other groups with the most translations and rankings. It required users to put in increasing work effort and offered increasing rewards but the rewards changed earlier than a consistent work effort scheme. By simply examining the quantity of contributions, a scheme that requires increasing work effort, and offers increasing rewards that rewards users sooner and more frequently in the earlier stages of the competition, was the most successful strategy to motivate users to engage and continue contributing.

5.2.2. Activity

Figure 24 shows the legitimate user activity for all the groups throughout the experiment. Users performed a substantial number of translations in the first two days and ranking took off on the second day when there were translations available to rank. Overall activity was lower on the weekend and picked up at a similar pace on Monday.

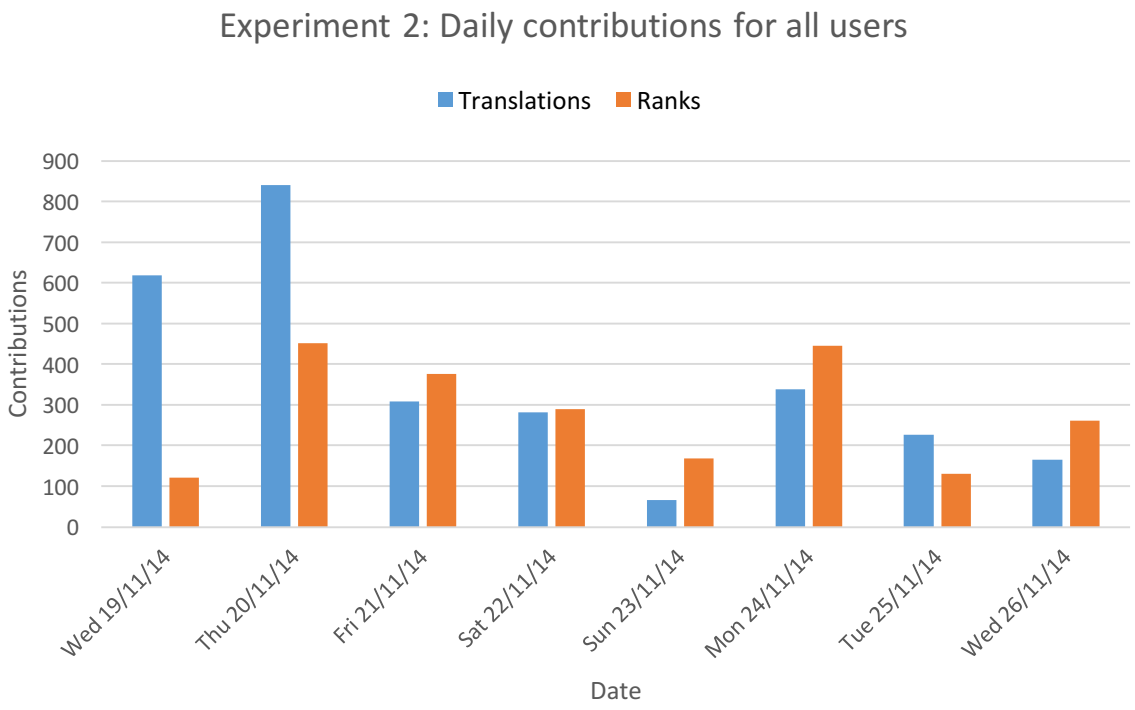

Figure 24: Experiment 2: Daily contributions for all users

Figure 25 shows the number of translations and rankings performed for each hour of the day (GMT +2). With the exception of a few time slots each day, the number of rankings performed each hour is less than translations. From 5 am, activity picks up to the peak in the early evening, after which it drops off sharply for the day.

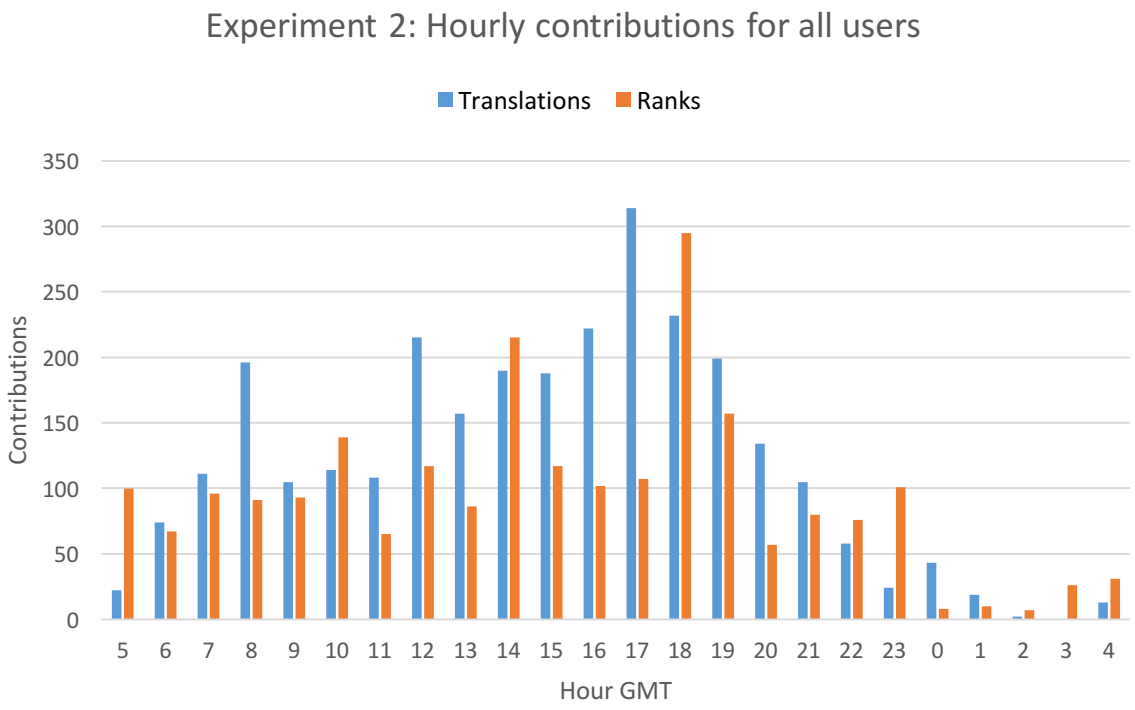

Figure 25: Experiment 2: Hourly contributions for all users

The legitimate translation intervals and ranking intervals for all groups have been plotted in Figure 26 and Figure 27. The charts illustrate that, as the experiment progressed, some users lost motivation and took longer intervals between subsequent translations. At any point during the experiment, an interval can never be more than the duration of the experiment at that point; this creates the imaginary diagonal line that appears in both charts. Figure 28 and Figure 29 take a closer look at the translation and rank intervals that fall under 30 minutes. These charts show that the activity reduced over the experiment and that the intervals converged as users became more proficient at translating and ranking. Translation intervals converged around 10 minutes and rank intervals below 5 minutes. Rank intervals could be lower because ranking takes less time than translating or they are more motivated to return after shorter durations to rank or a combination of the two factors.

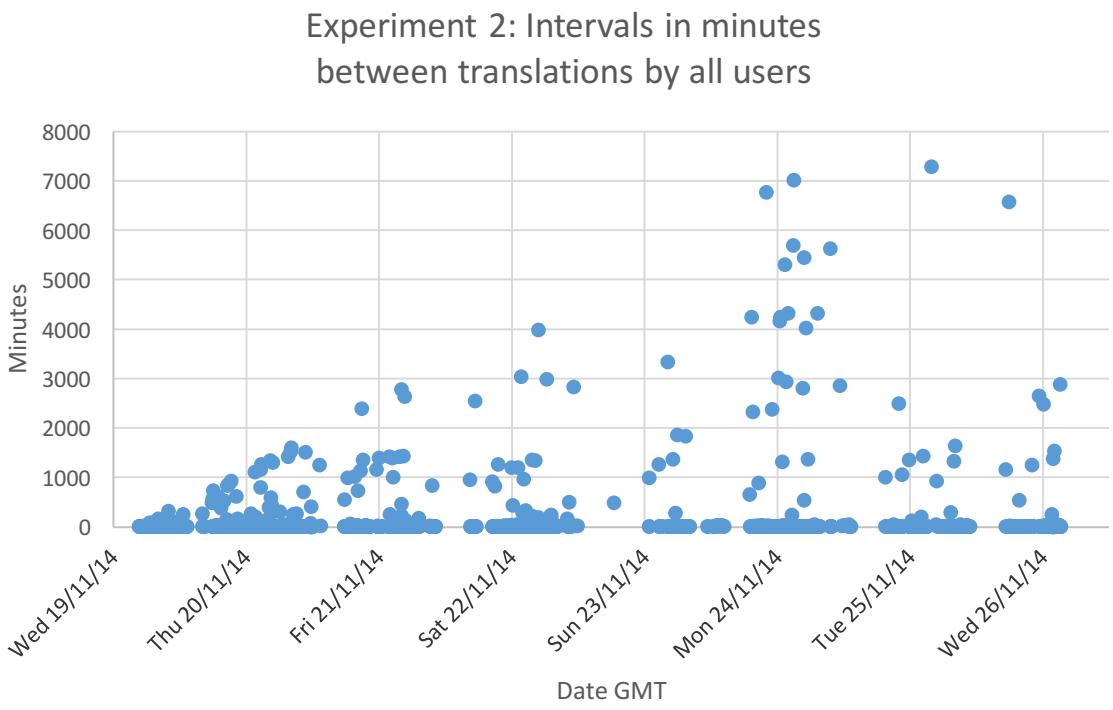

Figure 26: Experiment 2: Intervals in minutes between translations by all users

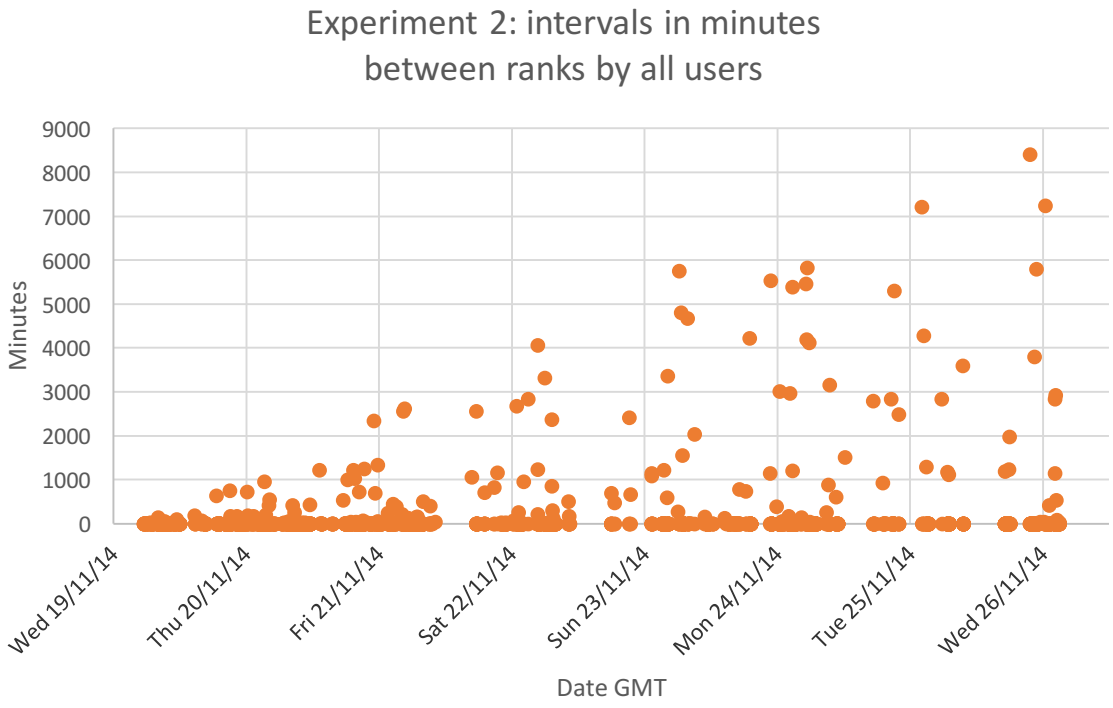

Figure 27: Experiment 2: Intervals in minutes between ranks by all users

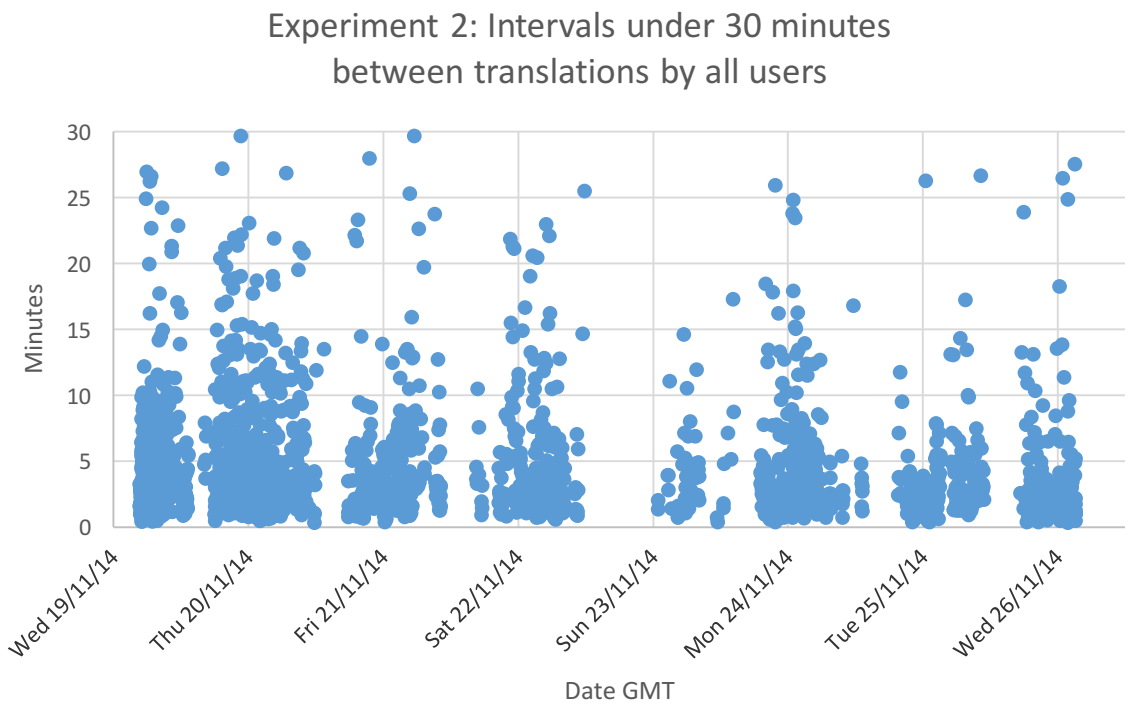

Figure 28: Experiment 2: Intervals under 30 minutes between translations by all users

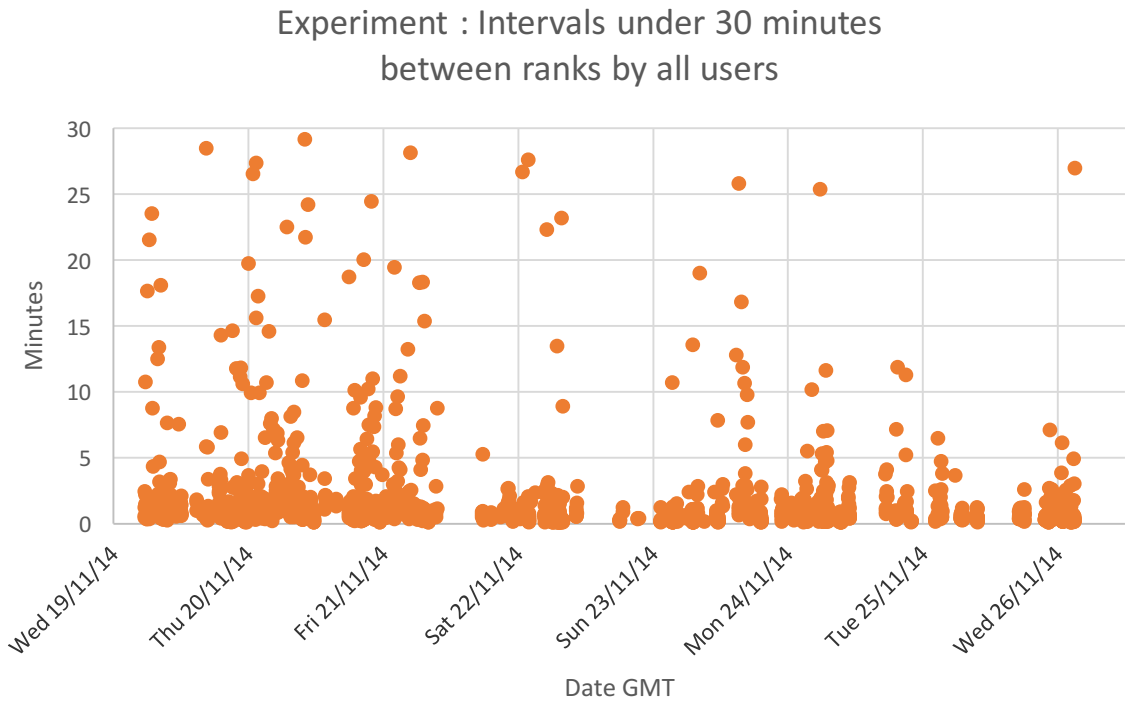

Figure 29: Experiment 2: Intervals under 30 minutes between ranks by all users

5.2.3. Analysis

The R statistical programming language (version 3.2.3 (2015-12-10) - "Wooden Christmas-Tree") was used to identify which schemes had an identifiable effect on user motivation. A Shapiro-Wilk test for normality was conducted on the translation and ranking intervals for each group. The null hypothesis (H0) is the population samples come from a normal distribution; the alternative hypothesis (H1) is that the population samples come from a non-normal distribution. A Shapiro-Wilk test for each translation and ranking interval population gave p-values less than 2.2e-16 (smallest value returned by R). At these p-values we reject the null hypothesis of normality and conclude that the alternative hypothesis of non-normality is true at the 95% confidence level.

All the population samples were found to be from non-normal distributions, therefore a Mann-Whitney-Wilcoxon test for identical populations; which does not require normal populations, was used to compare translation and ranking intervals for each group. The null hypothesis (H0) for the Mann-Whitney-Wilcoxon test for identical populations states that population samples A and B are identical; the alternative hypothesis (H1) states that the population samples A and B are non-identical. All the comparisons were performed with error rate (alpha value) of 0.05; this is a confidence level of 95%. Firstly, a comparison of translation and rank intervals for the six groups was performed in section 5.2.3.1. In section 5.2.3.2, translations and rankings are grouped by work effort, ignoring the effect of reward type, so that the effect of a consistent and increasing work load on user motivation can be compared. Similarly, in section 5.2.3.3, translations and rankings are grouped by reward type, ignoring the effect of work effort, so that the effect of a consistent, increasing and decreasing reward type on user motivation can be compared.

5.2.3.1. Comparison of All Groups

Population statistics for the translation intervals for all groups are available in Table 15 and Box and Whisker plots for all populations in Figure 30; the end point of upper Whisker error lines are not shown due to the high maximum values for all populations seen in Table 15.

Table 15: Experiment 2: Population statistics on translation intervals in seconds for all groups

	1	*2*	*3*	*4*	*5*	*6*
Q1 - Min	102	69	58	65	100	81.5
Min	21	20	27	20	22	23
Q1	123	89	85	85	122	104.5
Median	215.5	157.5	151.5	134.5	226.5	172
Q3	454	309	317.5	290.75	428.75	281.5
Max	437283	406088	139143	394330	420846	341842
Max - Q3	436829	405779	138825.5	394039.25	420417.25	341560.5

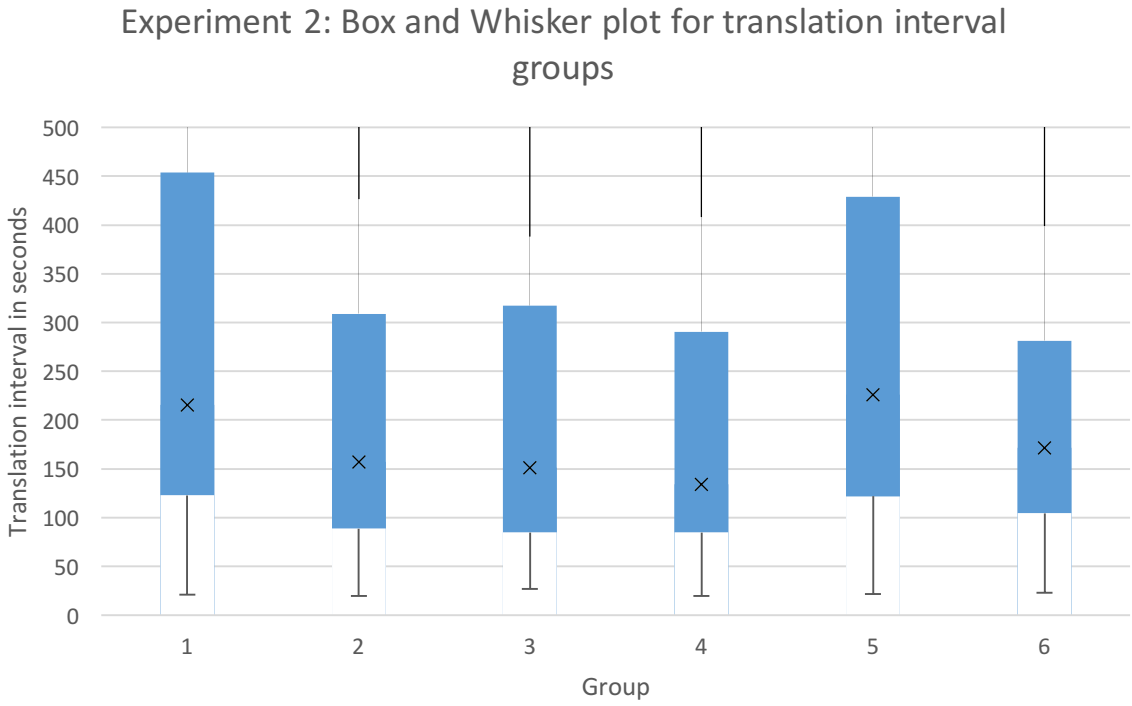

Figure 30: Experiment 2: Box and Whisker plot for translation interval groups

The p-values from the Mann-Whitney-Wilcoxon test for all possible pairings between the translation interval groups can be seen in Table 16, with significant p-values below an alpha value of 0.05 highlighted. For these comparisons there is sufficient evidence to reject the null hypothesis of identical populations and conclude that the alternative hypothesis of non-identical populations is true at the 95% confidence level. The users in these comparison groups behaved differently. The Box and Whisker plots in Figure 30 support these conclusions. The area and position of the boxes of Groups 2, 3, 4 and 6 are very similar, with third quartiles much lower than Groups 1 and 5.

Groups 1 and 5 also share a similar box area and position, both used a consistent work effort scheme and therefore change rewards later than the increasing work effort Groups 2, 4 and 5. Although Group 3 used a consistent work effort, it gave increasing rewards, which can explain the similar user behaviour. From these results it can be concluded that an increasing work effort scheme that changes rewards sooner and more frequently in the the early stages or an increasing reward scheme are better at motivating users to translate again after less time than a consistent work effort or decreasing payment scheme.

Table 16: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups

Group	*2*	*3*	*4*	*5*	*6*
1	1.13E-06	1.79E-06	5.29E-11	0.9375	0.0001025
2		0.5672	0.07798	6.58E-08	0.1812
3			0.4311	1.58E-07	0.09363
4				2.54E-13	0.001185
5					1.59E-05

Population statistics for the rank intervals for all groups are available in Table 17 and Box and Whisker plots for all populations in Figure 31; again the end point of upper Whisker error lines are not shown due to the high maximum values for all populations seen in Table 17.

Table 17: Experiment 2: Population statistics on rank intervals in seconds for all groups

	1	*2*	*3*	*4*	*5*	*6*
Q1 - Min	9	18	27.75	17	17.5	11
Min	10	10	10	10	10	10
Q1	19	28	37.75	27	27.5	21
Median	34	51	61.5	44	46	43
Q3	72	94	123.25	78.75	98	109
Max	434349	332414	140866	347697	323009	504186
Max - Q3	434277	332320	140742.75	347618.25	322911	504077

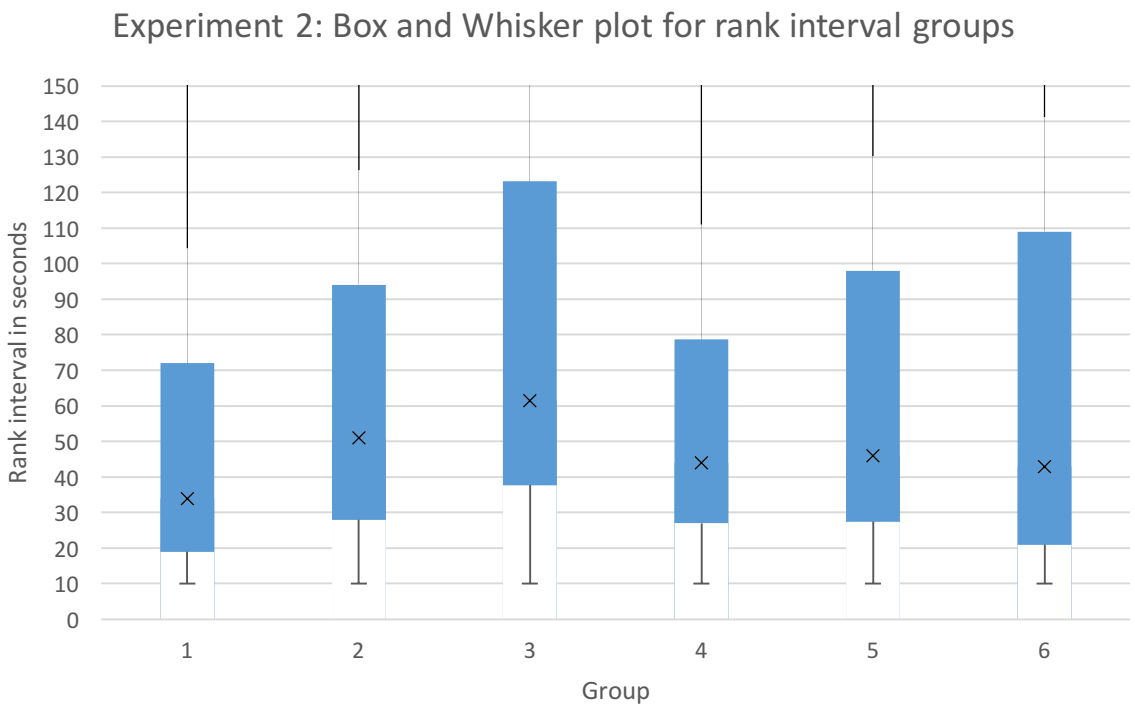

Figure 31: Experiment 2: Box and Whisker plot for rank interval groups

The p-values from the Mann-Whitney-Wilcoxon test for all possible pairings between the rank interval groups can be seen in Table 18. Significant p-values below an alpha value of 0.05 have been highlighted. For these comparisons, there is sufficient evidence to reject the null hypothesis of identical populations and conclude that the alternative hypothesis of non-identical populations is true at the 95% confidence level. The users in these comparison groups behaved differently. Group 4 with its increasing work effort and increasing rewards had a low third quartile, indicating its strength at motivating users to rank again after less time than the other groups performing well as it did with translating. Group 3 with its consistent work effort and Groups 5 and 6 which offered decreasing rewards, were the least effective schemes for motivating users to rank again sooner rather than later, apparent from their higher third quartile values.

Table 18: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on rank interval groups

Group	*2*	*3*	*4*	*5*	*6*
1	2.26E-07	5.61E-09	9.20E-05	0.0003448	0.1433
2		0.009356	0.05703	0.5367	0.091
3			3.56E-05	0.005132	0.002281
4				0.4763	0.4835
5					0.257

Across both translating and ranking, an increasing work effort that rewarded users early on and more frequently but steadily required more effort performed better than consistent work effort at motivating users to contribute again after less time. Increasing rewards performed better than consistent or decreasing rewards at motivating users to contribute after less time. Consistent work effort and decreasing rewards resulted in users taking more time between translating and ranking and is considered the worst combination of work effort and reward type.

5.2.3.2. Comparison of Work Effort Groups

Population statistics for the translation intervals for groups combined on work effort are available in Table 19 and Box and Whisker plots for theses grouped populations in Figure 32. Again the end point of upper Whisker error lines are not shown due to the high maximum values for all populations seen in Table 19.

Table 19: Experiment 2: Population statistics on translation intervals in seconds for groups combined on work effort

	Consistent Effort: 1, 3, 5	*Increasing Effort: 2, 4, 6*
Q1 - Min	88	70
Min	21	20
Q1	109	90
Median	195.5	151
Q3	404.5	296
Max	437283	406088
Max - Q3	436878.5	405792

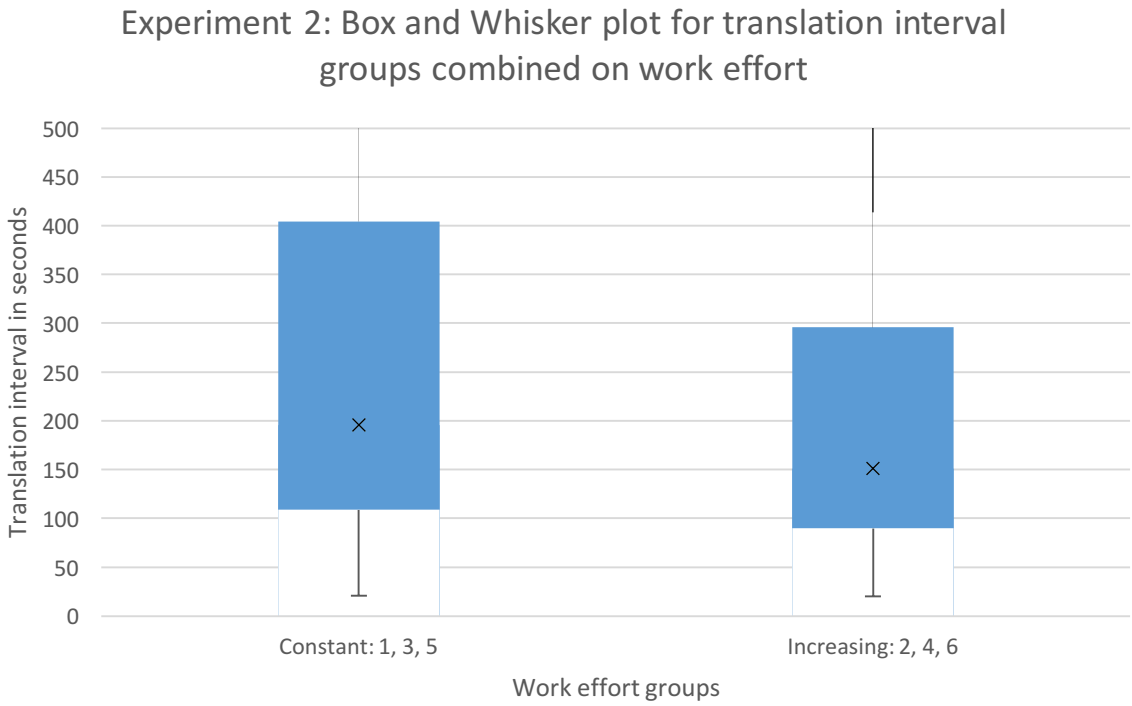

Figure 32: Experiment 2: Box and Whisker plot for translation interval groups combined on work effort

Figure 32 shows how the combined increasing work effort groups have a lower third quartile, meaning users in an increasing effort group were more likely to contribute again in less time than users in a consistent effort group and the result of the Mann-Whitney-Wilcoxon test, seen in Table 20, confirms that these populations are non-identical. This finding reinforces the previous conclusion that increasing work effort that rewards user sooner and more frequently during the early stages of progression performs better than a consistent work effort at motivating users to contribute again after less time.

Table 20: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups combined on work effort

Group	*Increasing: 2, 4, 6*
Consistent: 1, 3, 5	1.19E-10

Population statistics for the rank intervals for groups combined on work effort are available in Table 21 and Box and Whisker plots for theses grouped populations in Figure 33. Again the end point of upper Whisker error lines are not shown due to the high maximum values for all populations seen in Table 21.

Table 21: Experiment 2: Population statistics on rank intervals in seconds for groups combined on work effort

	Consistent Effort: 1, 3, 5	*Increasing Effort: 2, 4, 6*
Q1 - Min	14	17
Min	10	10
Q1	24	27
Median	44	46
Q3	88.75	89
Max	434349	504186
Max - Q3	434260.25	504097

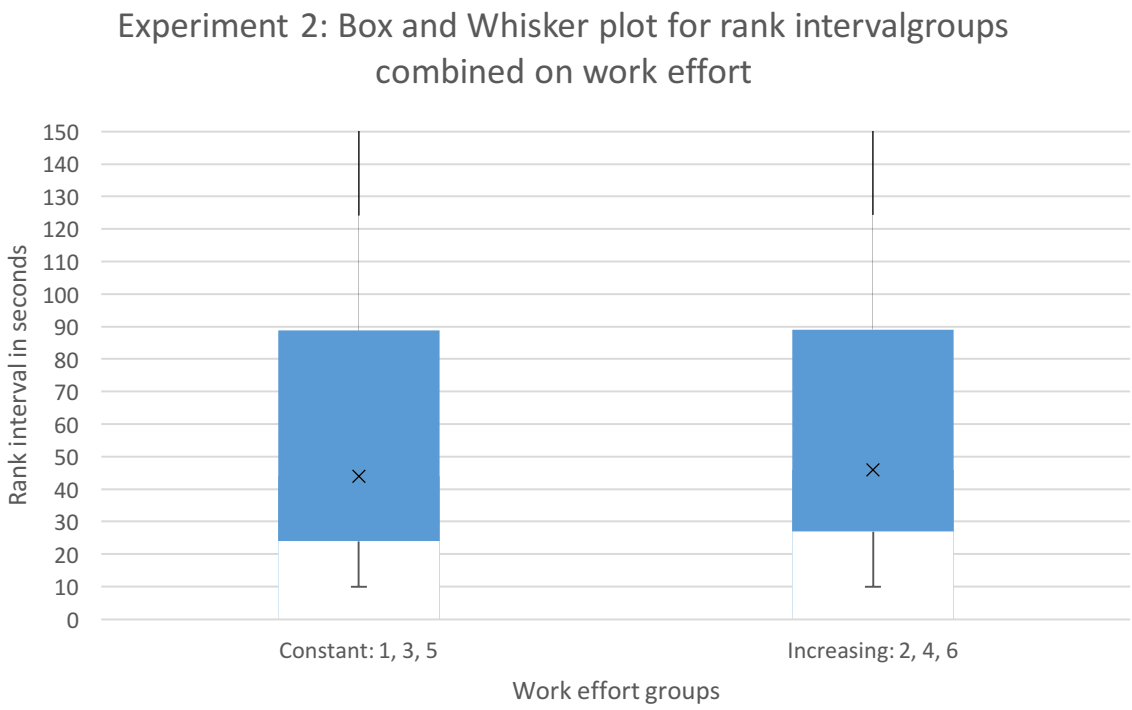

Figure 33: Experiment 2: Box an Whisker plot for rank interval groups combined on work effort

The result of the Mann-Whitney-Wilcoxon test, seen in Table 22, and the positon of both boxes in Figure 33 confirm that there is insufficient evidence to reject the null hypothesis that the work effort schemes are identical. Work effort does not appear to affect user motivation to rank sooner rather than later as it does translating.

Table 22: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups combined on work effort

Group	*Increasing: 2, 4, 6*
Consistent: 1, 3, 5	0.193

5.2.3.3. Comparison of Reward Groups

Population statistics for the translation intervals for groups combined on reward type are available in Table 23 and Box and Whisker plots for these grouped populations in Figure 34. Again the end point of upper Whisker error lines are not shown due to the high maximum values for all populations seen in Table 23.

Table 23: Experiment 2: Population statistics on translation intervals in seconds for groups combined on reward type

	Consistent Reward: 1, 2	*Increasing Reward: 2, 3*	*Decreasing Reward: 5, 6*
Q1 - Min	80	65	93
Min	20	20	22
Q1	100	85	115
Median	178	139.5	190
Q3	351.25	301	355
Max	437283	394330	420846
Max - Q3	436931.75	394029	420491

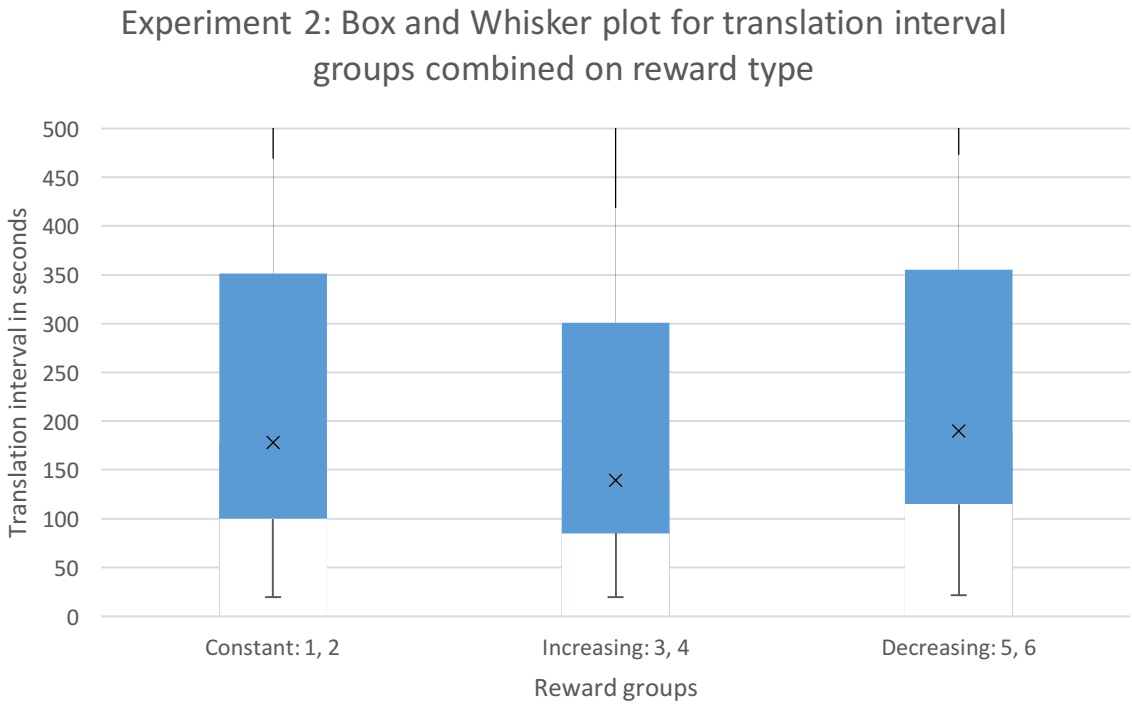

Figure 34: Experiment 2: Box an Whisker plot for translation interval groups combined on reward type

The result of the Mann-Whitney-Wilcoxon test, seen in Table 22, show that there is sufficient evidence to conclude that increasing reward groups behaved differently to consistent and decreasing reward groups and that there is insufficient evidence to reject the null hypothesis that the consistent and decreasing reward groups are identical. Figure 34 shows that the increasing reward groups had a lower first and third quartile, meaning users in increasing reward groups were more likely to translate again in less time than users in a consistent or decreasing reward group. Furthermore,

there does not appear to be a difference in user motivation between consistent and decreasing reward groups for translations.

Table 24: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation interval groups combined on reward type

Group	*Increasing: 3, 4*	*Decreasing: 5, 6*
Consistent: 1, 2	6.98E-06	0.06942
Increasing: 3, 4		5.15E-11

Population statistics for the rank intervals for groups combined on reward type are available in Table 25 and Box and Whisker plots for these grouped populations in Figure 35. Again the end point of upper Whisker error lines are not shown due to the high maximum values for all populations seen in Table 25.

Table 25: Experiment 2: Population statistics on rank intervals in seconds for groups combined on reward type

	Consistent Reward: 1, 2	*Increasing Reward: 2, 3*	*Decreasing Reward: 5, 6*
Q1 - Min	14	19	14
Min	10	10	10
Q1	24	29	24
Median	44	48	45.5
Q3	85	88	104
Max	434349	347697	504186
Max - Q3	434264	347609	504082

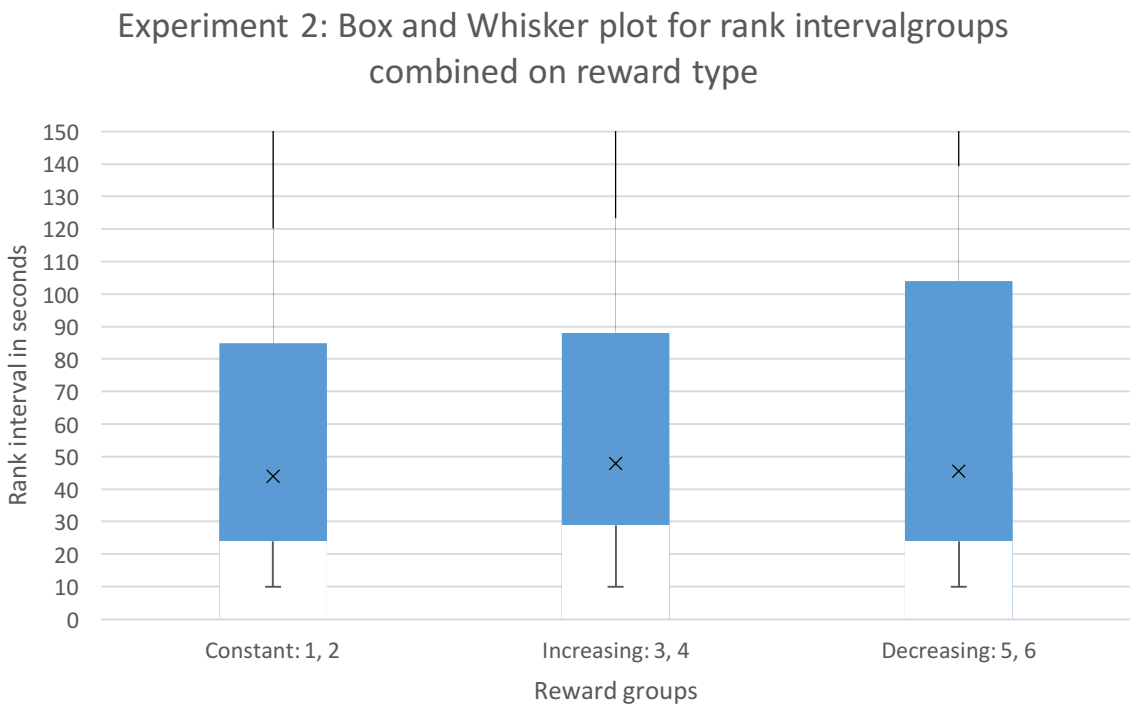

Figure 35: Experiment 2: Box an Whisker plot for rank interval groups combined on reward type

The result of the Mann-Whitney-Wilcoxon test, seen in Table 26, show that there is sufficient evidence to reject the null hypothesis and conclude that the consistent and increasing reward groups are non-identical populations and that there is insufficient evidence to reject the null hypothesis that the consistent and increasing reward groups are identical to the decreasing reward groups. Consistent and increasing rewards were better at motivating users to rank after less time than decreasing rewards, seen by their lower third quartile values.

Table 26: Experiment 2: p-values from Mann-Whitney-Wilcoxon test for identical populations on rank interval groups combined on reward type

Group	*Increasing: 3, 4*	*Decreasing: 5, 6*
Consistent: 1, 2	0.02566	0.4967
Increasing: 3, 4		0.3121

5.2.3.4. Average Active User Contributions

The finding from the various comparisons are reinforced when looking at the average number of contributions of active user per group (Figure 36). On average, Group 4

were motivated to contribute the most. The average contributions of users in both the decreasing reward groups - Groups 5 and 6 - is explained by their decreased motivation identified in the previous comparisons. Consistent work effort and decreasing payments is the least effective combination for motivating users to contribute again sooner rather than later. Users in Groups 1 and 2 behaved similarly as they received consistent rewards and were not affected by the leveling scheme.

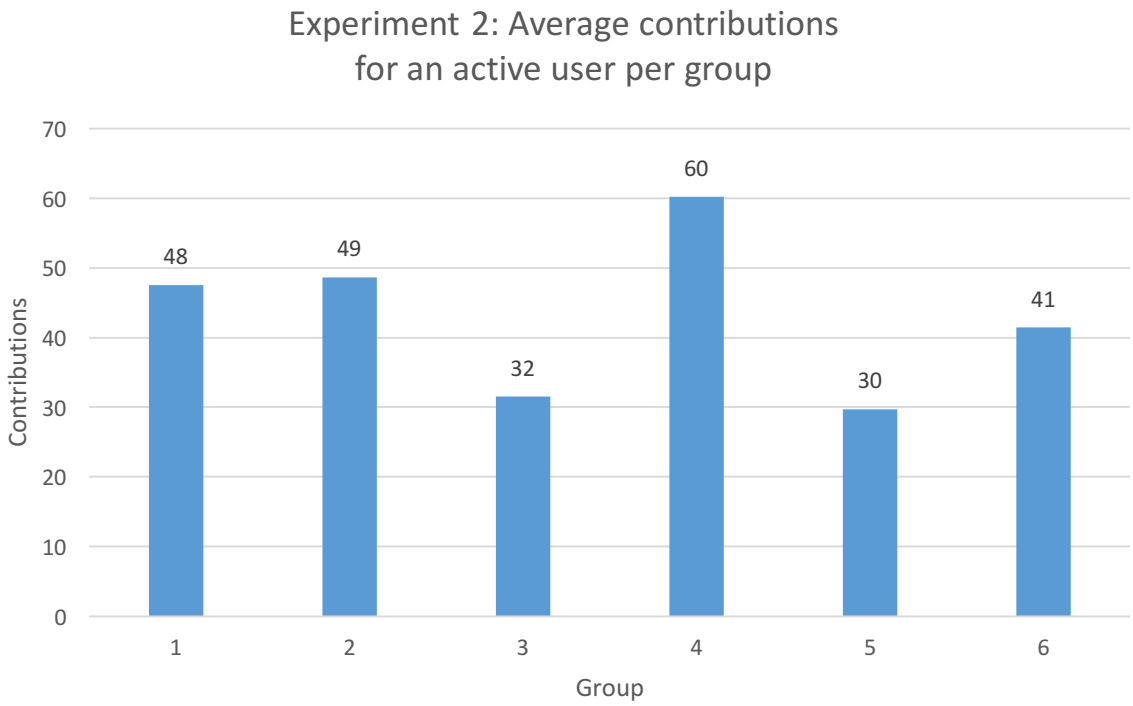

Figure 36: Experiment 2: Average contributions for an active user per group

5.3. Summary

Translating was more popular than ranking; an alternative design could largely have removed this issue. Ranking could either be rewarded more points than translating or tasks could be automatically allocated to users by the system instead of allowing them to choose. Experiment 2 had a high percentage of cheaters when compared to the active money earners; a more active stance warning users against cheating and disqualifying cheaters was taken in later experiments.

By comparing translation and ranking intervals, and grouping intervals by work effort and reward type, it was found that user motivation to translate again sooner rather than

later is affected by both the work effort required and the reward type. Motivation to rank again sooner rather than later seems to only be affected by the type of reward schemes employed and not the type of work effort required.

When comparing the individual six groups without grouping any of the variables, a scheme that requires increasing work effort or offers increasing rewards was better at motivating users to translate again in less time than consistent work effort or a decreasing payment scheme. The combination of increasing work effort and increasing rewards used in Group 4 performed the best at motivating users to translate again. The combination of consistent work effort and consistent rewards used in Group 1, along with the combination of consistent work effort and decreasing rewards used in Group 5, performed the worst. For ranking, the combination of increasing work effort and increasing payment scheme used in Group 4 performed well at motivating users to rank again sooner rather than later, as it did with translating. Unexpectedly, Group 1 with consistent work effort and consistent reward also performed well, contrary to the translation findings, but otherwise consistent work effort or decreasing payment groups such as Groups 3, 5 and 6 performed the worst at motivating users to rank again sooner rather than later.

Combining the six groups on work effort showed that groups that employed an increasing work effort were different to consistent work effort groups and that users in these groups were more likely to translate again after less time than users in a consistent effort group. Work effort did not appear to affect ranking; the consistent and increasing interval populations appear to be equal.

When combining the groups by reward type, users in increasing reward groups were motivated to translate again after less time than users in consistent and decreasing reward groups, which appear to have an identical effect on user motivation. For ranking offering, consistent or increasing rewards was better at motivating users to rank again after less time than offering decreasing rewards.

By looking at the individual variables of work effort and reward type, it was confirmed that overall increasing rewards motivate users more than decreasing rewards and rewarding increasing work effort motivates users more than rewarding work effort consistently.

6. Experiment 3: Removing rewards

Experiment 3 was designed to address the second research question. It specifically tested if the intrinsic value of the project alone was enough to motivate the same users who were appealed to in Experiment 2 to contribute if financial incentives were removed.

6.1. Methodology

The custom system covered in Chapter 3.3 was configured in the following way for the experiment:

- The full signup process was used.
- Four pre-assessment multiple choice questions were used and users had to get at least three correct.
- 1 user group was configured with a leaderboard.
- A translation redundancy of 3 was configured.
- A ranking redundancy of 3 was configured.
- No translation limit was set.
- No ranking limit was set.
- Only the scoring system was used.
- A multi-page design was used for the same reasons as Experiment 2.

6.1.1. Users

Experiment 3 used the same approach for gathering users as Experiment 2 but appealed to students at the start of the 2015 academic year. The new pool of students would include a new year of first year students and exclude the previous year's alumni. The call for participant's email was similar to that used in Experiment 2 but it specifically mentioned that no monetary reward would be given. The full email can be seen in Appendix C.

The different payment groups from Experiment 1 were dropped and replaced with a single user group. To make ranking more attractive, users were rewarded 20 points for ranking other user translations and only 10 points for translating.

6.1.2. Dataset

The same dataset from Experiment 2 was used but content already translated was skipped.

6.2. Results

The experiment began on Monday 24 February 2015 and finished after 9 days when the users had completely stopped contributing. 47 users registered but only 12 contributed at least one translation or ranking. Translating was again more popular than ranking, with a total of 37 translations and 10 rankings submitted, despite ranking offering double the points than that offered for translating. The user activity 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 translated and ranked 3 times. Offering a monetary reward was considerably more successful at attracting and engaging users.

No users were found to have cheated but with the low number of users this cannot be reliably concluded to be the result of offering no monetary reward. Experiment 2 saw an average of 1 cheater for 9 active users, but Experiment 3 only had 11 active users.

6.3. Summary

The gamification elements implemented, points and a high-score leaderboard, and intrinsic value of the project did not motivate users to engage when financial incentives were removed. Experiment 3 has given additional support to the idea that, in the context of sourcing content for low resource languages, users are less inclined to contribute without financial incentives, as was show in Experiment 1.

7. Experiment 4: Leaderboard rewards

Experiment 4 was designed to incorporate aspects of Experiment 2 and 3; it used a gamified payment scheme to attract users and a single group. The gamified payment scheme would be assessed on its potential to attract more users and generate more content, thereby addressing the third research question.

7.1. Methodology

The custom system covered in chapter 3.3 was configured in the following way for the experiment:

- The full signup process was used.
- Four pre-assessment multiple choice questions were used and users had to get at least three correct.
- 1 user group was configured with a leaderboard.
- A translation redundancy of 3 was configured.
- A ranking redundancy of 3 was configured.
- No translation limit was set.
- No ranking limit was set.
- The scoring system was removed and a user's total contributions determined their leaderboard position (Section 7.1.2).
- A multi-page design was used for the same reasons as Experiments 2 and 3.

7.1.1. Users

The experiment was designed to run for 2 weeks and to coincide with the first term vacation. An appeal was made to UCT students via the All Students email list on 25 March 2015; the full email can be viewed in Appendix D. The email detailed the payment scheme and explicitly warned students that anyone caught cheating would be disqualified and forfeit any rewards.

7.1.2. Payment Model

The payment scheme was designed to be simple but incorporate aspects of the original gamified payment schemes implemented for Experiment 2. Users would be paid based

on their leaderboard placement and not directly for each contribution. The top leaderboard position was allocated the largest reward and each subsequent position less. A budget of ZAR 6,000.00 was allocated to the top 40 positions on the leaderboard. Figure 37 illustrates the breakdown of rewards. For the exact rewards, see Appendix F. Experiment 2 showed that only 61 out of the 200 users earned money and therefore this scheme was designed to offer rewards to only the most active users. A simpler scoring method was used where points were abandoned and the individual user contributions (translations and rankings) were counted.

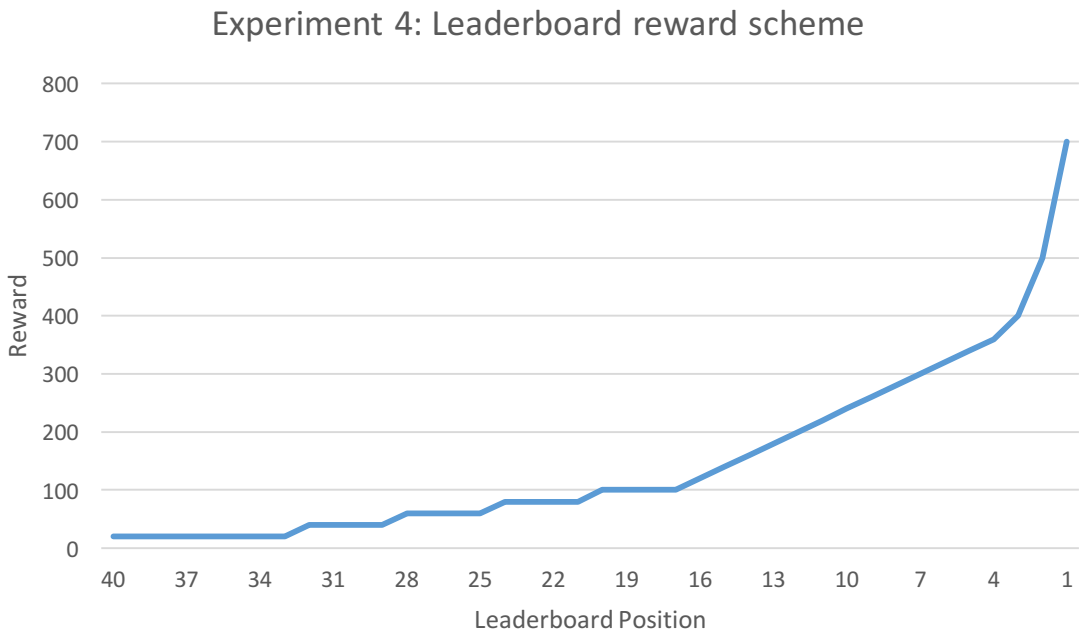

Figure 37: Experiment 4: Leaderboard payment scheme

7.1.3. Dataset

The same dataset from Experiment 2 and 3 was used but content already translated was skipped.

7.1.4. Results

Experiment 4 received 147 users but only 57 users contributed. A total of 1865 individual translations were contributed and 617 sentences received 3 translations. A total of 1767 rankings were contributed and 584 sentences received 3 rankings. Experiment 4 achieved a translation cost of R0.22 per word, almost double the rate of

Experiment 2. Experiment 4's task allocation policy and single group resulted in a noticeable improvement on the number of sentences that were ranked; only 5.8% were not completely ranked while Experiment 2 had 32.5% that were not completely ranked.

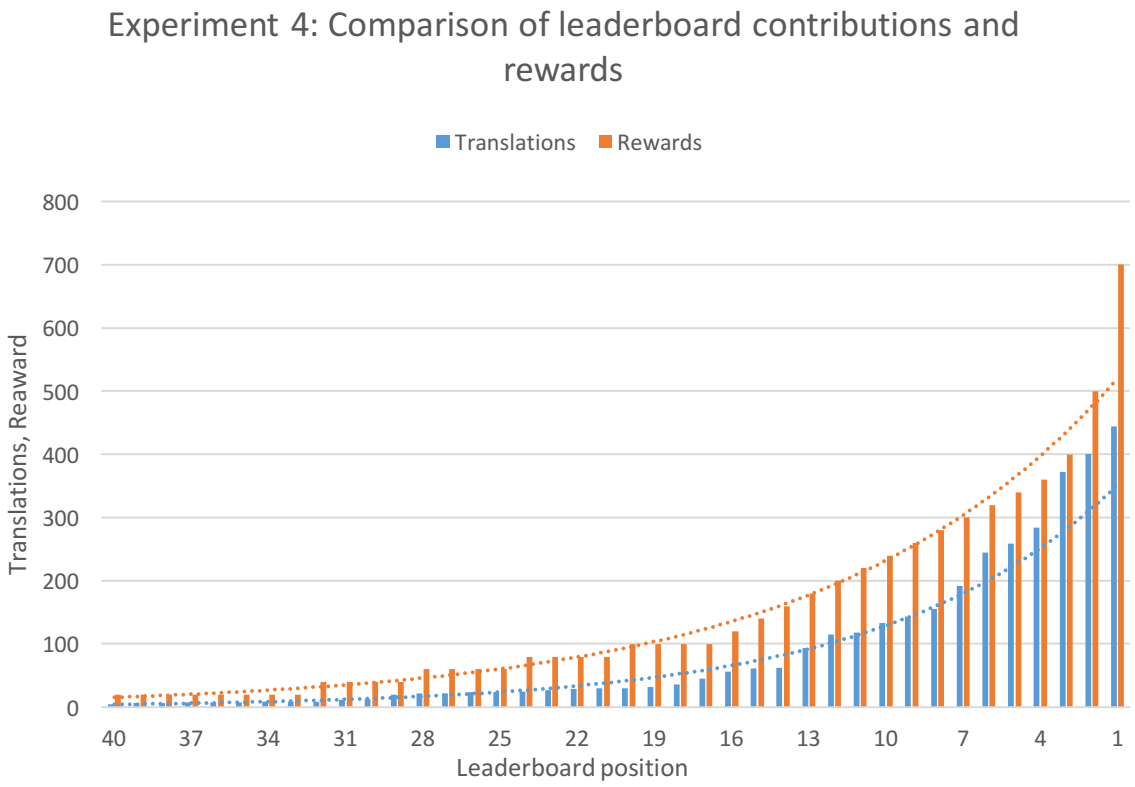

Figure 38: Experiment 4: Comparison of leaderboard contributions and rewards

Figure 38 shows how contributions closely matched the increasing rewards, although at a slightly slower rate. A reward system based on leaderboard position motivated the top users to contribute proportionally more than those lower on the leaderboard.

7.1.5. Pre-processing Data

The same cheating detection techniques used in Experiment 2 were used and no users were caught explicitly cheating but the same cut-off point of 20 seconds for translations and 10 seconds for ranks were used, which resulted in a marginal percentage of contributions being excluded, as seen in Figure 39. Explicitly warning users against cheating had a positive effect on their behaviour.

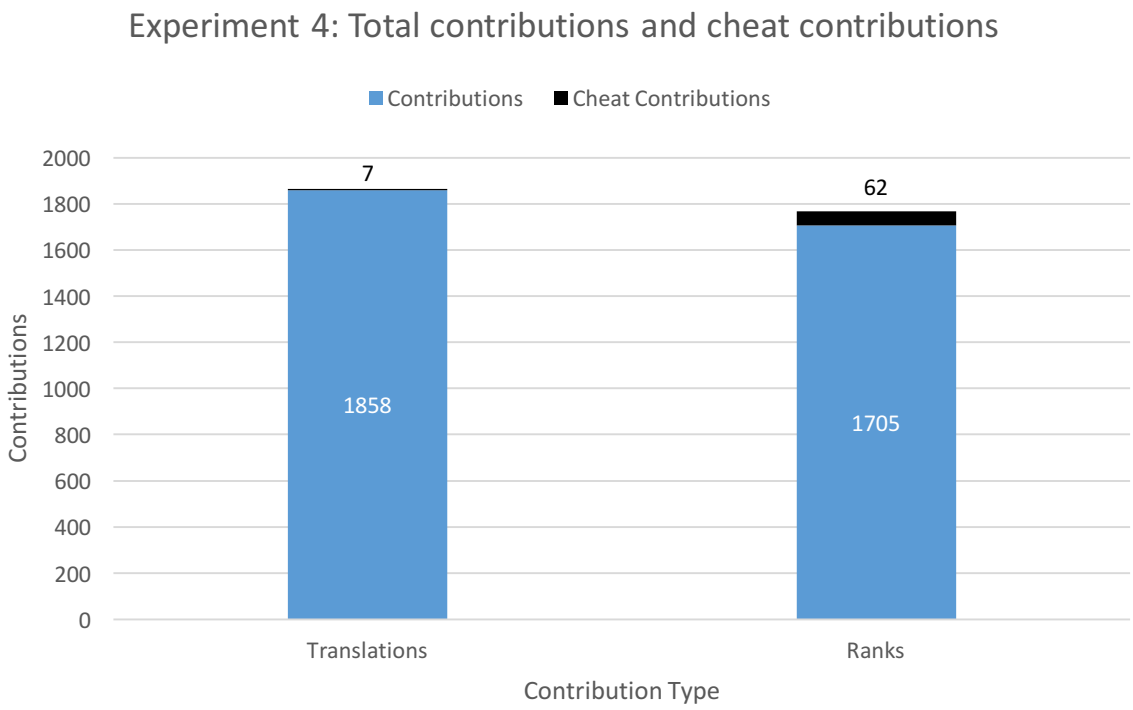

Figure 39: Experiment 4: Total contributions and cheat contributions

7.1.6. Activity

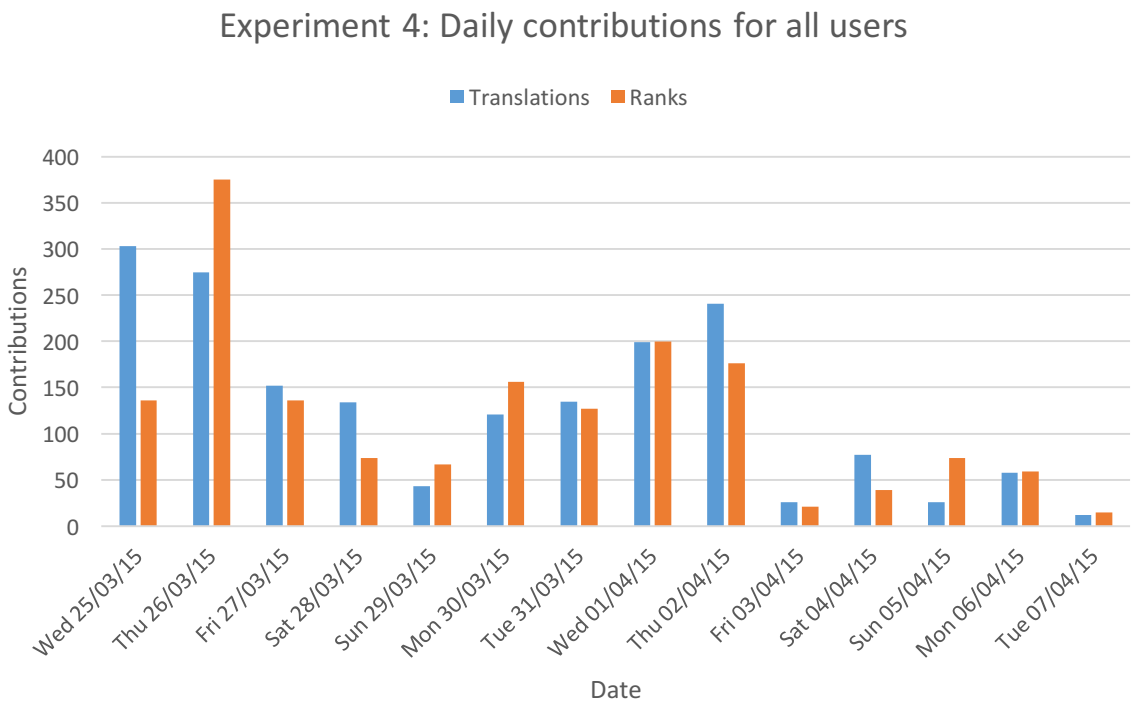

Figure 40: Experiment 4: Daily contributions for all users

It was expected that running the experiment for longer would result in an equivalent number of contributions, but it appears users were most active during the first 9 days of the experiment and lost interest by the second weekend, as shown in Figure 40. Perhaps users were also less likely to contribute during the middle of their short vacation if it looked like they weren't going to have a chance to reach and keep a higher leaderboard position. Figure 41 shows that users in Experiment 4 rose to their peak activity more quickly than users in Experiment 2, reaching their peak activity around mid-day instead of the early evening. This is most likely due to the Experiment 4 overlapping more with the student vacation than Experiment 2, so students now had more of their days available, unlike in Experiment 2.

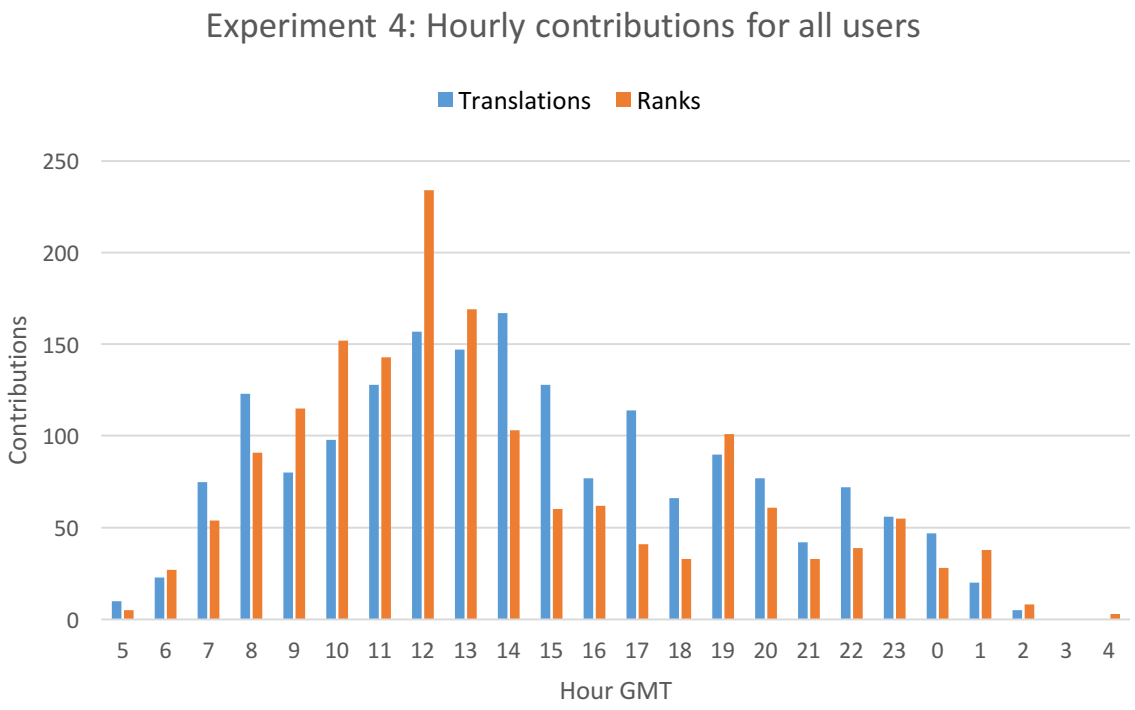

Figure 41: Experiment 4: Hourly contributions for all users

The new task allocation scheme, which allocated translations and ranks to ensure users maintained an equal balance between the number of each task complete, successfully reduced the number of completed translations that remained unranked, an issue with Experiment 2. Unfortunately, by removing the freedom of choice, users were allocated tasks in a fairly predictable fashion, usually a translation followed by a ranking if one was available, which resulted in the intervals between translations being more dependent on rank intervals and vice versa than they were in Experiment 2. This is more apparent when examining the translation intervals plotted in Figure 42 and Figure 44 and ranking intervals plotted in Figure 43 and Figure 45. A specific example that illustrates this phenomenon can be seen in Figure 42 and Figure 43 on Thursday 02/04/15 around 10,000 minutes. At this point, 3 users returned to the experiment after having not contributed since the beginning, all three translated and ranked, identifiable by the three similar data points in both charts.

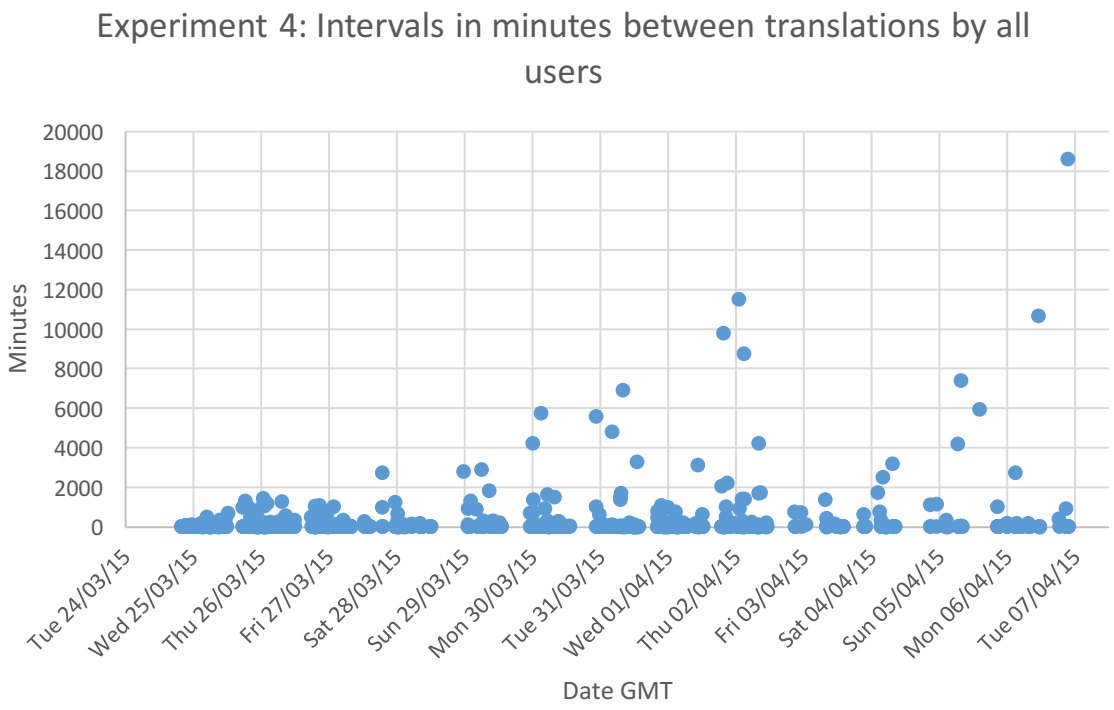

Figure 42: Experiment 4: Intervals in minutes between translations by all users

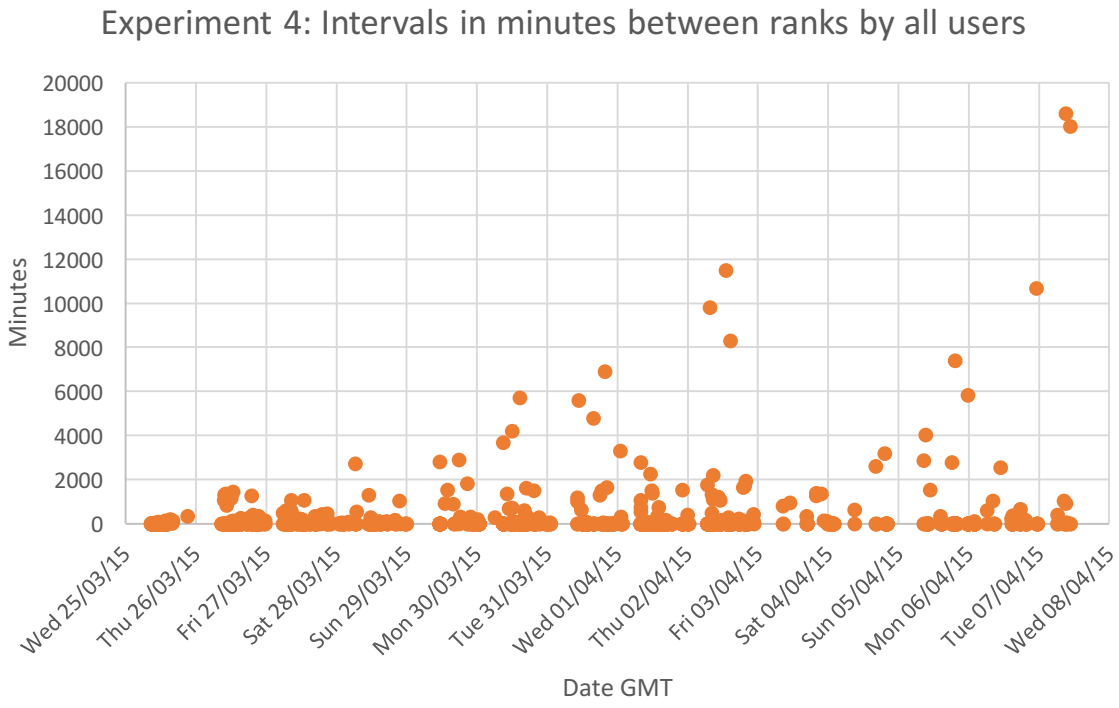

Figure 43: Experiment 4: Intervals in minutes between ranks by all users

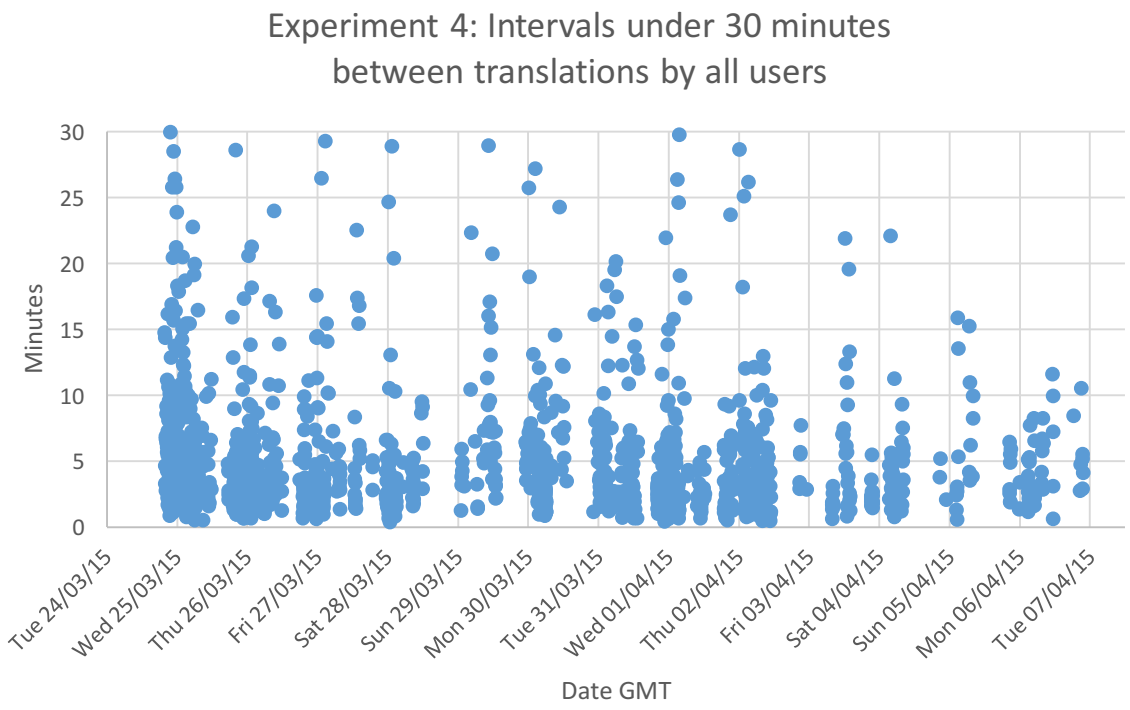

Figure 44: Experiment 4: Intervals under 30 minutes between translations by all users

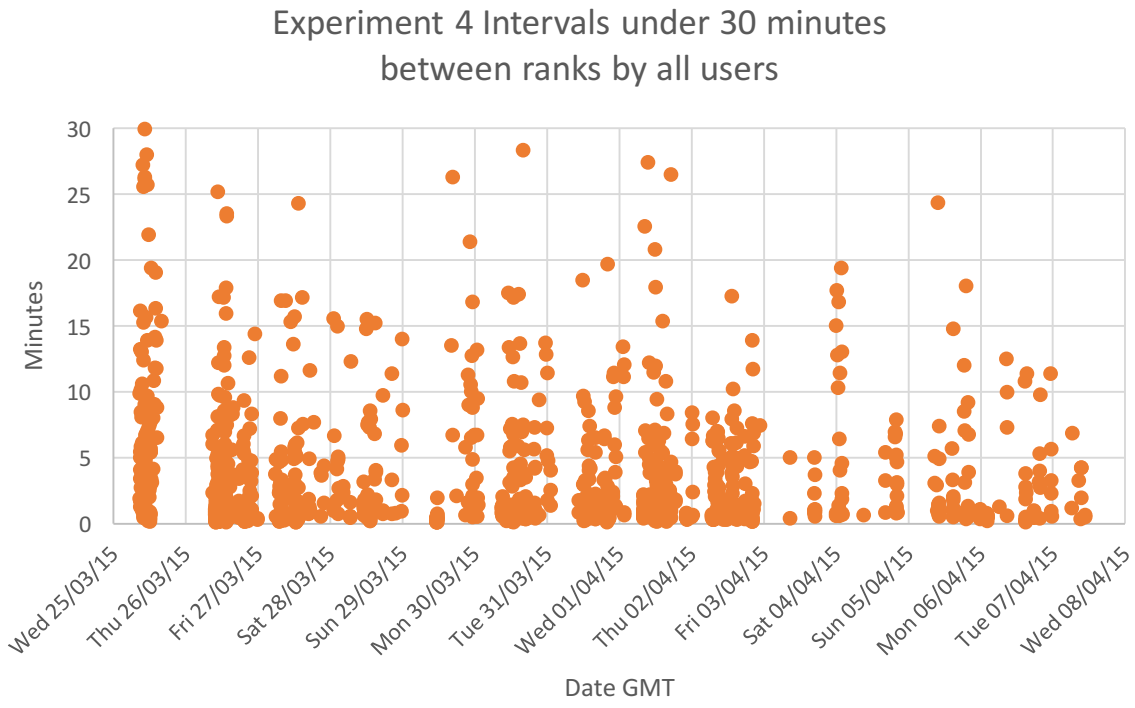

Figure 45: Experiment 4: Intervals under 30 minutes between ranks by all users

7.1.7. Analysis

The dependency of translation intervals on ranking intervals and vice versa can be seen again when plotting the translation and rank intervals for Experiment 4 alongside the equivalent Box and Whisker plots from Experiment 2, seen in Figure 46 and Figure 47 respectively, and together in Figure 48. Compared to Experiment 2, ranking intervals are considerably longer, almost as long as translation intervals. This is because there was a high chance that users would have translated before each rank.

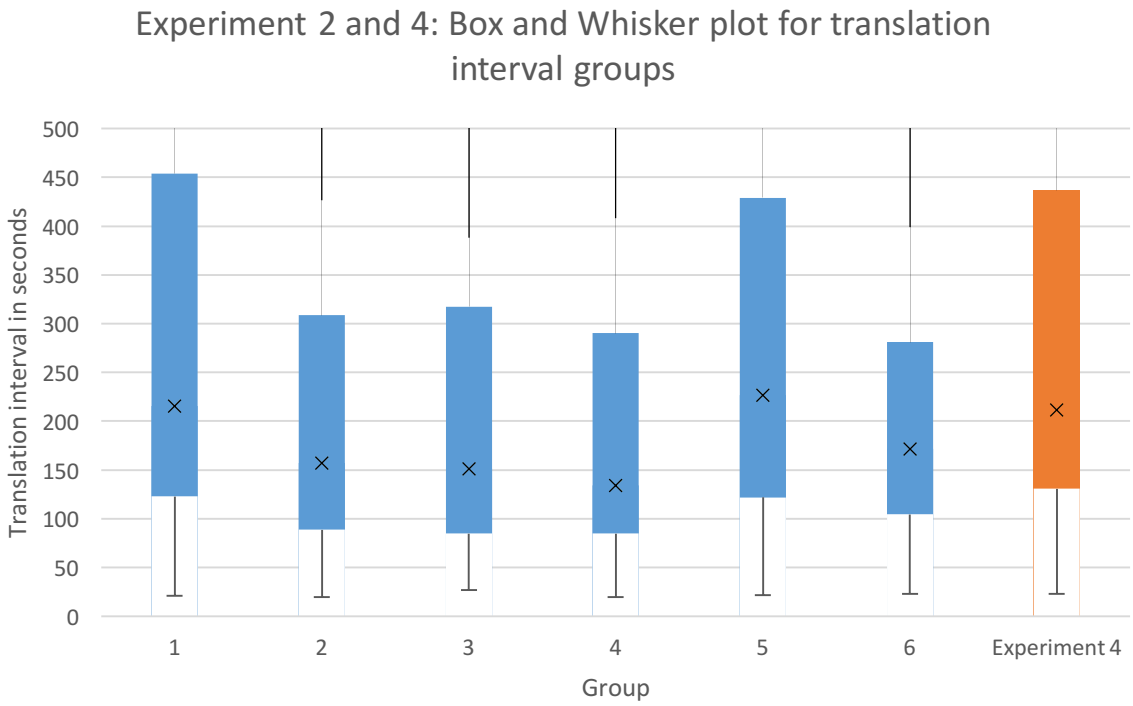

Figure 46: Experiment 2 and 4: Box and Whisker plot for translation interval groups

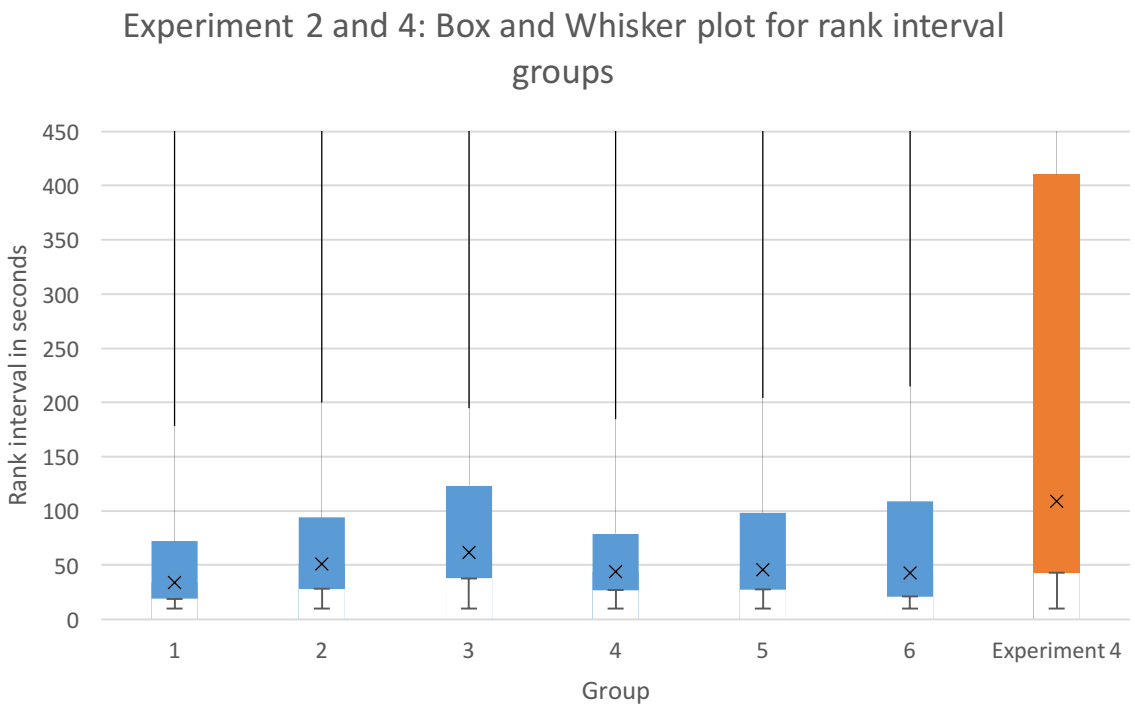

Figure 47: Experiment 2 and 4: Box and Whisker plot for rank interval groups

This deep dependency between translation and rank intervals makes it impossible to compare user motivation across Experiment 4 and Experiment 2 in the same manner as was used in Experiment 2.

Table 27: Experiment 4: Population statistics on translation and rank intervals in seconds for all users

	Translations	*Ranks*
Q1 - Min	108	33
Min	23	10
Q1	131	43
Median	212	109
Q3	437	410.5
Max	1116177	1117116
Max - Q3	1115740	1116705.5

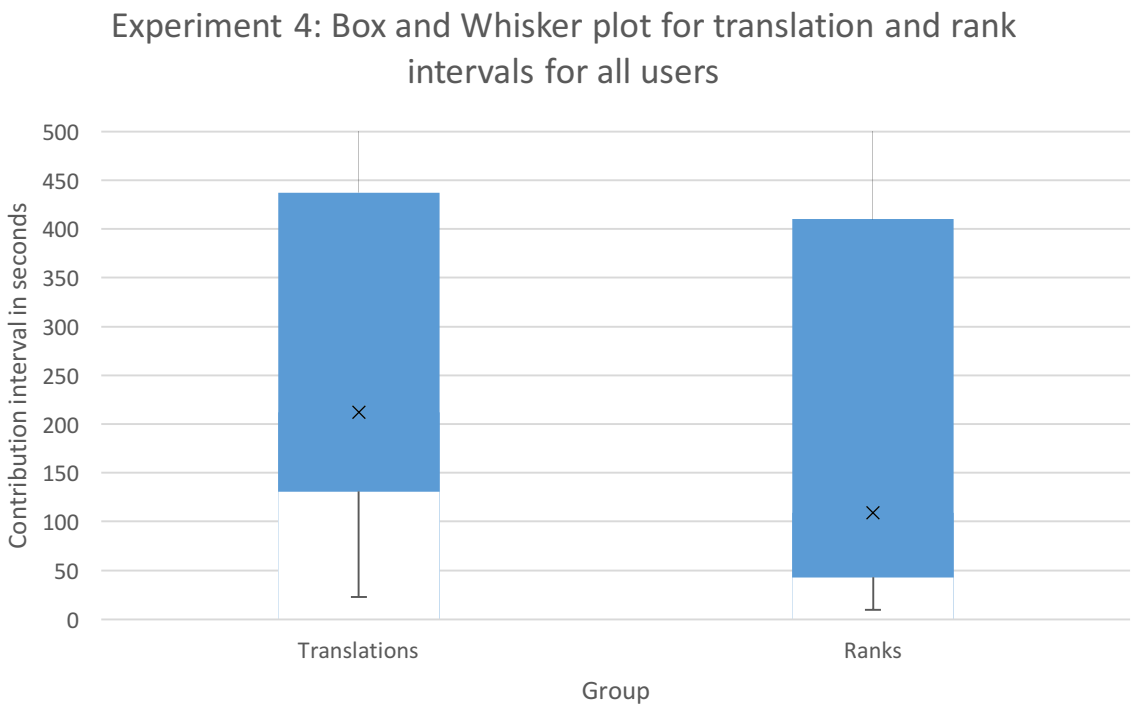

Figure 48: Experiment 4: Box and Whisker plot for translation and rank interval for all users

Although the translation and rank populations from Experiment 4 are not identical, seen in the population statistics in Table 27 and the result of a Mann-Whitney-Wilcoxon test in Table 28, they have influenced each other, creating higher first quartile, median and third quartile values.

Table 28: Experiment 4: p-values from Mann-Whitney-Wilcoxon test for identical populations on translation and rank intervals

	Ranks
Translations	< 2.2e-16

We are still able to compare the average contribution of active user between Experiment 2 groups and Experiment 4. Figure 49 shows that, Experiment 4 users were as active as the most active user group from Experiment 2 - Group 4. Both shared design similarities by requiring increasing work effort and offering increasing rewards, but Group 4 from Experiment 2 offered guaranteed incentives and Experiment 4 offered the possibility of incentives, yet both were equally successful at motivating users to contribute. An average active user in Group 4 from Experiment 2 contributed 60 times and an average active user in Experiment 4 contributed 63 times.

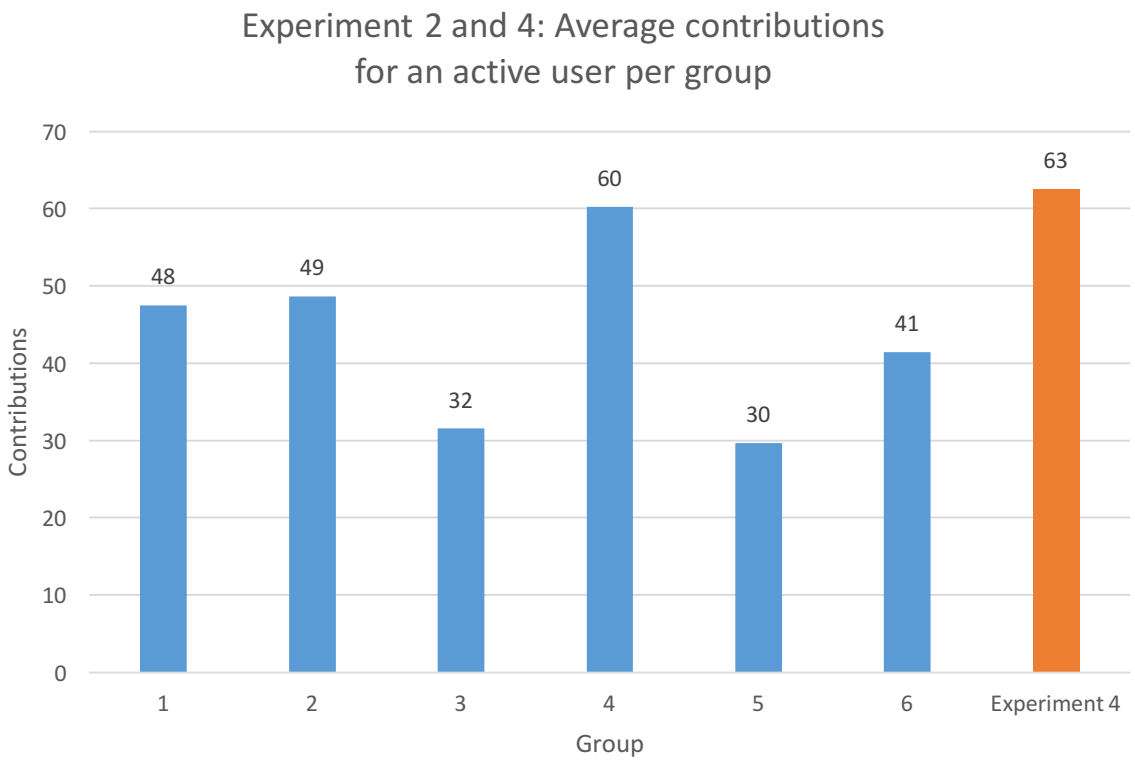

Figure 49: Experiment 2 and 4: Average contributions for an active user per group

7.2. Summary

Despite Experiment 4 using a single group and offering more rewards, it did not generate as much content as Experiment 2. The resultant translation cost was almost double that of Experiment 2. In future the leaderboard rewards could be adjusted based on past user activity to have greater predictability on the resultant translation cost. Users at the top of the leaderboard were substantially more active than users at the bottom. The resultant effect of the leaderboard reward scheme used for Experiment 4 matches the design scheme of requiring increasing work effort and offering increasing rewards, used by Group 4 from Experiment 2. Furthermore, Experiment 4 users were as active as Group 4 users, the most active group in Experiment 2, reinforcing the findings from Experiment 2, that requiring an increasing work effort and offering an increasing reward scheme was the most successful at motivating users to engage more and engage again sooner rather than later. Experiment 4 also showed that the leaderboard reward scheme, which offered the possibility of incentives, was as attractive as the guaranteed incentives offered by Group 4 from Experiment 2.

8. Conclusions

Four experiments were run to test whether people would participate in various online crowdsourcing games to translate English content into isiXhosa. Experiments 1 and 3 showed that people do not volunteer for free or would not continue contributing if payments were taken away.

Experiment 2 showed that paying for contributions can generate content at rates considerably cheaper than professional translation services. The experiment also confirmed that offering rewards in a crowdsourcing project can result in a high percentage of users cheating. For translating, groups that required an increasing work effort or offered increasing payments motivated users to translate more and to keep translating sooner rather than later. The combination of requiring an increasing work effort and offering increasing rewards was the most motivating. User motivation for ranking was not affected by the type of work effort required but a consistent or increasing reward type motivated users to rank more and rank again sooner rather than later. Overall, increasing work effort and increasing rewards are preferable to requiring consistent work effort and offering consistent or decreasing rewards.

Experiment 4 had a higher percentage of completed translations and rankings because of an improved allocation algorithm. A strict warning against cheating reduced cheating almost entirely. Over-estimating an increased activity for Experiment 4 and offering generous reward tiers resulted in a higher translation word cost than Experiment 2. Still, Experiment 4 users were as active as the most active and motivated group from Experiment 2 - Group 4. Both these groups required increasing work effort and offered increasing rewards but Group 4 from Experiment 2 offered guaranteed incentives while Experiment 4 only offered the possibility of incentives, but both were equally attractive for users. If the user activity for Experiment 4 is compared to the other groups from Experiment 2, the possibility of incentives is more attractive than directive incentives, but these groups did not share design similarities with a leaderboard based payment scheme as Group 4 did.

8.1. Research Questions

8.1.1. Question 1

> *In the context of crowdsourcing content for low resource languages, what is the effect on user engagement and contribution quantity when paying users consistent, increasing or decreasing rewards for subsequent contributions?*

Experiment 2 showed that users were more motivated to contribute when they were rewarded for their increased work effort or when offered increasing rewards for subsequent contributions than when work effort was consistently rewarded or consistent or decreasing rewards were offered. Furthermore, ranking was not affected by the type of work effort required and decreasing rewards performed the worst for both translating and ranking.

8.1.2. Question 2

> *In the context of crowdsourcing content for low resource languages, will users continue to contribute if payments are taken away and only intrinsic motivators remain?*

Experiment 1 was not able to attract users by only relying on the intrinsic value of the project and, when removing payments in Experiment 3 and appealing to the same pool of users used in Experiment 2, substantially fewer users participated. Only 47 contributions were made, compared to 6189 in Experiment 2. The intrinsic value of the project was not as strong a motivator as offering financial rewards for getting people to contribute to a crowdsourcing game for a low resource language.

8.1.3. Question 3

> *In the context of crowdsourcing content for low resource languages, is the possibility of future rewards more attractive to users than direct guaranteed incentives?*

The activity of users from Experiment 4 and from similarly designed Group 4 from Experiment 2 show that people find the possibility of future rewards and direct

guaranteed incentives equally attractive; an average user in both groups contributed nearly the same number of times.

8.2. Contributions and implications

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 true in some cases. The experiments have illustrated that in the context of low resource environments, monetary payment is still a stronger motivation factor than intrinsic motivation and gamification but people prefer gamified monetary payment to non-gamified monetary payment and the possibility of incentives was as attractive as guaranteed incentives.

9. Limitations and future work

9.1. Further explore sourcing participants from social networks

Experiment 1 tried sourcing participants from the author's Twitter network but was unsuccessful because the author and majority of his immediate network did not speak isiXhosa. Twitter and other social networks could still be a powerful tool to recruit participants for crowdsourcing instead of relying on crowdsourcing market places. Future research into this area should appeal to users who speak the same language.

9.2. Further develop gamified payment schemes

The payment schemes used in Experiment 2 can be developed further through larger or more focused experiments to strengthen the findings. Decreasing rewards seem to be less attractive and could be replaced with another scheme such as random rewards.

For the leaderboard payment scheme, it would be useful to determine the point where non-guaranteed incentives become unattractive by continuously lowering them closer to the point of contributing for free. Alternatively, a leaderboard payment scheme that takes into account past user performance may be able to achieve more predictable and affordable translations rates. Additional leaderboards such as daily leaderboards and leaderboards that segment users based on past activity could also be explored.

9.3. Further explore non-financial rewards

Although the intrinsic value of the project and implemented gamification elements were not successful at motivating users to engage and contribute without financial rewards, it would be worthwhile exploring what other motivation factors other than financial rewards are important to the demographic of users required for low resource language crowdsourcing projects. Community reputation, offering coupons, discounts and free necessities such as clothes and groceries are some of the other motivation factors that could be explored.

9.4. Repeat the experiments in other countries

It would be valuable to know how the experiments perform with other low resource languages and if the economic environment of a country affects user motivation. Are financial rewards more important in developing countries and less important in developed countries? Will the leaderboard payment scheme with its promise of reward outperform a guaranteed reward in developed countries?

9.5. Crowdfund crowdsourcing

This research, along with numerous past studies, shows that although content can be sourced at more affordable rates than those offered by professional translation services, it is still a costly affair to crowdsource enough content for low resource languages from users who are only motivated by financial reward. Crowdfunding, a form of crowdsourcing used to raise funds, could be used to support the crowdsourcing of workers. What motivators will the crowdfunding arm of the project use to attract funders? Will most of the funders come from the same country or abroad? Is this model sustainable in the long term?

9.6. Further develop IR algorithms and tools

With more resources and improved payment schemes, additional content could be crowdsourced and combined with content crawled from the Web by other studies to further develop IR algorithms and tools with the ultimate goal of assembling parallel language corpora that are sufficient to train machine translations systems.

Bibliography

Allwood, Jens, Leif Grönqvist, and A. P. Hendrikse. 2003. “Developing a Tagset and Tagger for the African Languages of South Africa with Special Reference to Xhosa.” *Southern African Linguistics and Applied Language Studies* 21 (4): 223–37.

Allwood, Jens, Harald Hammarström, Andries Hendrikse, Mtholeni N. Ngcobo, Nozibele Nomdebevana, Laurette Pretorius, and Mac van der Merwe. 2010. “Work on Spoken (multimodal) Language Corpora in South Africa.” http://bada.hb.se/handle/2320/7401.

Ambati, Vamshi, and Stephan Vogel. 2010. “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*, 62–65. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1866706.

Anderson, David P. 2004. “Boinc: A System for Public-Resource Computing and Storage.” In *Grid Computing, 2004. Proceedings. Fifth IEEE/ACM International Workshop on*, 4–10. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1382809.

Callison-Burch, Chris. 2009. “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*, 286–95. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1699548.

“CoffeeScript.” 2014. Accessed October 17. http://coffeescript.org/.

Croft, W. Bruce, Donald Metzler, and Trevor Strohman. 2010. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Reading. http://library.mpib-berlin.mpg.de/toc/z2009_2465.pdf.

de Klerk, Vivian. 2006. “Codeswitching, Borrowing and Mixing in a Corpus of Xhosa English.” *International Journal of Bilingual Education and Bilingualism* 9 (5): 597–614.

De Saussure, Ferdinand. 2011. *Course in General Linguistics*. Columbia University Press. https://books.google.co.za/books?hl=en&lr=&id=n6VFhwfLs0gC&oi=fnd&pg=PR9&dq=F.+de+Saussure:+Course+in+General+Linguistics&ots=G8kSwKfjp0&sig=geVLfMEYlD2tK9uECJq63l0o4U0.

Deterding, Sebastian, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. "Gamification. Using Game-Design Elements in Non-Gaming Contexts." In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2425–28. ACM. http://dl.acm.org/citation.cfm?id=1979575.

Drummer, Aurelia. 2013. "Phrase-Based Machine Translation of Under-Resourced Languages." http://people.cs.uct.ac.za/~bsharwood/downloads/AureliaDrummer_Report.pdf.

Eickhoff, Carsten, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. 2012. "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*, 871–80. ACM. http://dl.acm.org/citation.cfm?id=2348400.

Eiselen, E., and M. Puttkammer. 2014. "Developing Text Resources for Ten South African Languages." In *Proc. LREC*. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf.

"First National Bank." 2014. Accessed October 22. https://www.fnb.co.za/.

Garside, Roger, Geoffrey N. Leech, and Tony McEnery. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Taylor & Francis.

Hamari, Juho, Jonna Koivisto, and Harri Sarsa. 2014. "Does Gamification Work?–A Literature Review of Empirical Studies on Gamification." In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 3025–34. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6758978.

Harter, Susan. 1978. "Effectance Motivation Reconsidered. Toward a Developmental Model." *Human Development* 21 (1): 34–64.

Havenga, M., K. Williams, and H. Suleman. 2012. “Motivating Users to Build Heritage Collections Using Games on Social Networks.” In . http://pubs.cs.uct.ac.za:1081/archive/00000807/.

Howe, Jeff. 2008. *Crowdsourcing: How the Power of the Crowd Is Driving the Future of Business*. Random House.

Htonl. 2013. “South Africa 2011 Xhosa Speakers Proportion Map.” https://commons.wikimedia.org/wiki/File:South_Africa_2011_Xhosa_speakers_proportion_map.svg.

Johnson, Kristine K. 2011. “Xhosa-English Machine Translation: Working with a Low-Resource Language.” http://www.cra.org/Activities/craw_archive/dmp/awards/2011/Johnson/kkjohnson_report.pdf.

Jurafsky, Dan, and James H. Martin. 2000. *Speech & Language Processing*. Pearson Education India.

Kaufmann, Nicolas, Thimo Schulze, and Daniel Veit. 2011. “More than Fun and Money. Worker Motivation in Crowdsourcing–a Study on Mechanical Turk.” http://aisel.aisnet.org/amcis2011_submissions/340/.

Koehn, Philipp. 2005. “Europarl: A Parallel Corpus for Statistical Machine Translation.” In *MT Summit*, 5:79–86. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.5497&rep=rep1&type=pdf.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies*. Routledge London. http://www.mersindilbilim.info/wp-content/uploads/2011/09/McEnery-Xiao-Tono2.pdf.

“MongoDB.” 2014. Accessed October 17. https://www.mongodb.org/.

Munro, Robert. 2010. “Crowdsourced Translation for Emergency Response in Haiti: The Global Collaboration of Local Knowledge.” In *AMTA Workshop on Collaborative Crowdsourcing for Translation*. http://amta2010.amtaweb.org/AMTA/papers/7-01-01-Munro.pdf.

Negri, Matteo, and Yashar Mehdad. 2010. “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*, 212–16. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1866730.

Pavlick, Ellie, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. “The Language Demographics of Amazon Mechanical Turk.” *Transactions of the Association for Computational Linguistics* 2: 79–92.

Post, Matt, Chris Callison-Burch, and Miles Osborne. 2012. “Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing.” In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 401–9. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2393073.

Pretorius, Laurette, and Sonja Bosch. 2009. “Exploiting Cross-Linguistic Similarities in Zulu and Xhosa Computational Morphology.” In *Proceedings of the First Workshop on Language Technologies for African Languages*, 96–103. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1564526.

“PyBOSSA.” 2014. Accessed October 13. http://pybossa.com/.

Ross, Joel, Lilly Irani, M. Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. “Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk.” In *CHI’10 Extended Abstracts on Human Factors in Computing Systems*, 2863–72. ACM. http://dl.acm.org/citation.cfm?id=1753873.

Ryan, Richard M., and Edward L. Deci. 2000. “Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being.” *American Psychologist* 55 (1): 68.

Salton, Gerard. 1968. “Automatic Information Organization and Retrieval.” http://www.citeulike.org/group/896/article/500181.

Sharwood, Brett. 2013. “Machine Translation of Under-Resourced Languages.” http://people.cs.uct.ac.za/~bsharwood/downloads/BeeSharwood_Report.pdf.

Silvertown, Jonathan. 2009. “A New Dawn for Citizen Science.” *Trends in Ecology & Evolution* 24 (9): 467–71.

Statistics South Africa. 2012. *Census 2011 Census in Brief*. Private Bag X44, Pretoria 0001: Statistics South Africa. http://www.statssa.gov.za/Census2011/Products/Census_2011_Census_in_brief.pdf.

Webb, Victor N. 2000. *African Voices: An Introduction to the Languages and Linguistics of Africa*. Oxford University Press.

Wiggins, Andrea, and Kevin Crowston. 2011. “From Conservation to Crowdsourcing: A Typology of Citizen Science.” In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, 1–10. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5718708.

Wynne, Martin, John Sinclair, Geoffrey Leech, Lou Burnard, Anthony McEnery, Richard Xiao, and Paul Thompson. 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Vol. 92. Oxbow Books Oxford. http://www.uam.es/proyectosinv/woslac/DOCUMENTS/Libros%20proyecto.doc.

Zaidan, Omar F., and Chris Callison-Burch. 2011. “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*, 1220–29. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2002626.

Appendices

Appendix A Experiment 2: Call for participants email UCT

Dear students,

We invite all bilingual English/isiXhosa speakers to participate in a competitive online game to translate English content into isiXhosa. Your reward is capped at a maximum of R100 for participating. Please note that your participation is entirely voluntary and you are free to decline to participate in this game. All identifying information will be kept confidential.

To register for the game please click on the link below and complete a registration form.

English/isiXhosa Translation Game Registration (link).

If you register you will be contacted via email when the game starts on Wednesday, the 19th of November at 17:00. You will need to complete an online consent form to begin playing the game. You will compete with other users on translation score and can view your ranking on the leaderboard. Translate more to beat the other users to the top. The game will run for 1 week until Wednesday the 26th of November at 17:00. When the game ends users will be paid within 1 week by a cardless bank transaction to a South African mobile number. You will receive a SMS with a virtual voucher, which can be withdrawn from any South African FNB ATM at no charge. The game is not limited to UCT students. Anyone with conversational language skills in both English and isiXhosa may participate.

If you have any questions or concerns about this study, please feel free to contact Sean Packham at pcksea001@myuct.ac.za. I am UCT MSc student in the Department of Computer Science being supervised by Professor Hussein Suleman. This study has been approved by the Faculty of Science Ethics Committee at The University of Cape Town (ref no: FSREC 055–2014).

Appendix B Experiment 2: Online registration form

English/Xhosa Translation Game Registration

We invite all bilingual English/Xhosa speakers to participate in a competitive online game to translate English content into Xhosa. Your reward is capped at a maximum of R100 for participating. Please note that your participation is entirely voluntary and you are free to decline to participate in this game. All identifying information will be kept confidential.

If you register you will be contacted via email when the game starts on Wednesday the 19th of November at 17:00. You will need to complete an online consent form to begin playing the game. You will compete with other participants on translation score and can view your ranking on the leaderboard. Translate more to beat the other participants to the top. The game will run for 1 week until Wednesday the 26th of November at 17:00. When the game ends participants will be paid within 1 week by a cardless bank transaction to a South African mobile number. You will receive a SMS with a virtual voucher, which can be withdrawn from any South African FNB ATM at no charge. The game is not limited to UCT students. Anyone with conversational language skills in both English and Xhosa may participate.

If you have any questions or concerns about this study, please feel free to contact Sean Packham at pcksea001@myuct.ac.za. I am UCT MSc student in the Department of Computer Science being supervised by Professor Hussein Suleman. This study has been approved by the Faculty of Science Ethics Committee at The University of Cape Town (ref no: FSREC 055–2014).

* Required

Name *

Email *

So we can notify you when the game begins

Submit

Never submit passwords through Google Forms.

Appendix C Experiment 3: Call for participants email UCT

Dear students,

We invite all bilingual English/isiXhosa speakers to participate in a competitive online game to translate English content into isiXhosa (Users will receive no reward or remuneration). Please note that your participation is entirely voluntary and you are free to decline to participate in this game. All identifying information will be kept confidential.

To take part in the game, please click on the link below to register and begin playing.

English/isiXhosa Translation Game Registration (link).

If you register you will be contacted via email when the game starts on Wednesday, the 19th of November at 17:00. You will need to complete an online consent form to begin playing the game. You will compete with other users on translation score and can view your ranking on the leaderboard. Translate more to beat the other users to the top. The game will run for 1 week until Wednesday the 26th of November at 17:00. When the game ends users will be paid within 1 week by a cardless bank transaction to a South African mobile number. You will receive a SMS with a virtual voucher, which can be withdrawn from any South African FNB ATM at no charge. The game is not limited to UCT students. Anyone with conversational language skills in both English and isiXhosa may participate.

If you have any questions or concerns about this study, please feel free to contact Sean Packham at pcksea001@myuct.ac.za. I am UCT MSc student in the Department of Computer Science being supervised by Professor Hussein Suleman. This study has been approved by the Faculty of Science Ethics Committee at The University of Cape Town (ref no: FSREC 055–2014).

Appendix D Experiment 4: Call for participants email UCT

Dear students,

We invite all bilingual English/isiXhosa speakers to participate in a competitive online game to translate English content into isiXhosa Please note that your participation in the game is entirely voluntary and you are free to decline to participate. All identifying information will be kept confidential.

To take part in the game, please click on the link below to register and begin playing.

English/isiXhosa Translation Game Registration (link).

Participating will help contribute South Africa isiXhosa content to Wikipedia. As a player in the game you will translate English sentences into isiXhosa and rank other user translations. At the end of the game the highest ranked sentences are assembled back into full articles and submitted to the isiXhosa Wikipedia.

The top 40 users will earn a reward (total prize pool is R6000, largest individual prize is R700) which will be paid out in the form of an FNB eWallet payment, to their South African mobile number (supplied during registration), which can be withdrawn as cash from any FNB ATM. Payment will be made within 3 weeks, after the game has ended. The closer you are to the top of the leaderboard the more you will win. Keep contributing as moving up 1 spot can boost your rewards significantly.

The game will run for 2 weeks, starting on the 25 March 2015 and ending at 22:00 8 April 2015.

Deliberately contributing false translations or rankings will get you flagged for cheating, forfeiting any rewards.

If you have any questions or concerns about this study, please feel free to contact Sean Packham at pcksea001@myuct.ac.za. I am UCT MSc student in the Department of Computer Science being supervised by Professor Hussein Suleman. This study has been approved by the Faculty of Science Ethics Committee at The University of Cape Town (ref no: FSREC 055–2014).

Appendix E Experiment dataset

Table 29: Wikipedia articles used as the dataset for all experiments

Wikipedia Article	*Accessed*
Cape_Town	02/08/2014
Johannesburg	02/08/2014
Nelson_Mandela	18/11/2014
South_Africa	18/11/2014
Rugby	19/11/2014
University_of_Cape_Town	19/11/2014
Lions_Head	19/11/2014
Table_Mountain	19/11/2014
MeerKAT	19/11/2014
Greenmarket_Square	19/11/2014
MyCiTi	19/11/2014
The_Noon_Gun	19/11/2014
The_Groote_Schuur_Zoo	19/11/2014
Quagga	19/11/2014
Workers_Day	20/11/2014
Youth_Day	20/11/2014
Freedom_Day	20/11/2014
Human_Rights_Day	20/11/2014
Heritage_Day	20/11/2014
The_Day_of_Reconciliation	20/11/2014
Steve_Biko	20/11/2014
Desmond_Tutu	20/11/2014
Tutu_House	20/11/2014
Graca_Machel	20/11/2014
John_Langalibalele_Dube	20/11/2014
Chris_Hani	20/11/2014
Albertina_Sisulu	20/11/2014
Walter_Sisulu	20/11/2014
Oliver_Tambo	20/11/2014

Appendix F Experiment 4: Leaderboard payment scheme

Position	*Reward*	*Position*	*Reward*
40	20	20	100
39	20	19	100
38	20	18	100
37	20	17	100
36	20	16	120
35	20	15	140
34	20	14	160
33	20	13	180
32	40	12	200
31	40	11	220
30	40	10	240
29	40	9	260
28	60	8	280
27	60	7	300
26	60	6	320
25	60	5	340
24	80	4	360
23	80	3	400
22	80	2	500
21	80	1	700