


UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Project Final Paper 2022

Title: A Gamified Approach to Boosting Crowdsourced Data Collection for Abstract Wikipedia's Multilingual Content Generation

Author: Zahraa Hoosen

Project Abbreviation: WING

Supervisor(s): Associate Professor Dr Maria Keet

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	0
Theoretical Analysis	*0*	*25*	0
Experiment Design and Execution	*0*	*20*	15
System Development and Implementation	*0*	*20*	20
Results, Findings and Conclusions	*10*	*20*	15
Aim Formulation and Background Work	*10*	*15*	10
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	0
Total marks		**80**	

A Gamified Approach to Boosting Crowdsourced Data Collection for Abstract Wikipedia's Multilingual Content Generation

Zahraa Hoosen
University of Cape Town
Cape Town, South Africa
HSNZAH008@myuct.ac.za

ABSTRACT

The Abstract Wikipedia project aims to make knowledge more accessible by automatically generating articles across multiple languages. This requires extensive lexicographical data, especially in the case of underrepresented African languages. Crowdsourcing through a 'Game with a Purpose' (GWAP) format is a possible solution. This paper introduces two versions of a GWAP called "Word Safari", which uses gamification to boost user participation in crowdsourcing lexicographical data for 3 Niger-Congo B languages - isiZulu, isiXhosa and Shona. One version included one gamification element of a points system and the other version had various gamification elements - such as a points system, a leaderboard, daily streaks and feedback screens. The research question investigated is how do gamification elements motivate users to contribute data and how do these gamified interfaces compare with the current Wikidata interface in terms of users contributing data. Preliminary design began with a high-fidelity prototype, which was refined through user feedback. Android applications were then created for both interfaces. Two experiments were run with each of the interfaces with 10 participants each over five days. Participants were offered monetary incentives for their participation. Both interfaces did collect more lexicographical data compared to Wikidata and the interface with more gamification elements collected more data than the simpler gamified interface. Although gamification did motivate participants to contribute more data, monetary motivation was their primary motivation and this is in line with other literature.

CCS Concepts: • **Information systems** → *Crowdsourcing*; • **Human-centered computing** → *Human-computer interaction (HCI)*; *Collaborative and social computing*.

Keywords: Wikipedia, Wikidata, interface, lexicon, lexicography, gamification, crowdsourcing, low-resourced languages, microtasking, validation

1 INTRODUCTION

Wikipedia seeks to address the current disparity in knowledge across different languages with the Abstract Wikipedia project, which will utilise a language-independent structure that allows for the automatic generation of articles [24]. This approach is expected to be more scalable than the current method of manually written articles and will enable the same article to be available in multiple languages [24]. The Abstract Wikipedia project aims to achieve this by creating a new type of multilingual knowledge repository that will utilise natural language generation (NLG) and renderer functions stored in Wikifunctions to generate articles from abstract data and language-specific lexicographic data stored in Wikidata [1]. The project's multilingual article generation requires extensive lexicographical data that it does not have, especially for low-resourced African languages such as the Niger-Congo B languages. This hinders the project's ability to identify concepts and similarities across languages, making it challenging to produce articles in different languages.

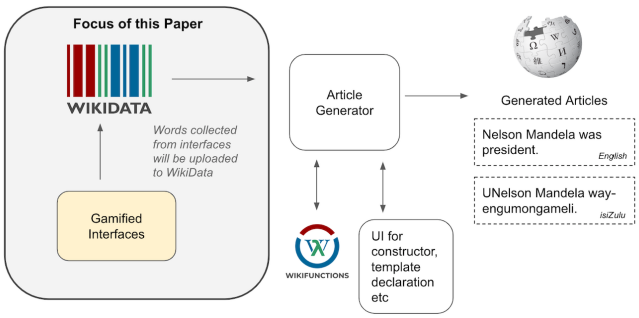


Figure 1. Where the gamified interfaces fit in the Abstract Wikipedia project

The current interface that collects lexicographical data for Wikidata is aimed at expert contributors and is not accessible to most speakers of Niger-Congo B languages and this limits who can contribute to Wikidata. A solution to this is expanding the contributor pool to include ordinary speakers of these languages through gamified interfaces that use crowdsourcing.

This paper will explore the accessibility of two gamified interfaces compared to the existing Wikidata interface, as well as identify the gamification elements that could increase user engagement. To guide this investigation, the research question posed is: 'What specific gamification elements, drawn from two different gamified interfaces, can motivate users to contribute lexicographical data for Niger-Congo B languages,

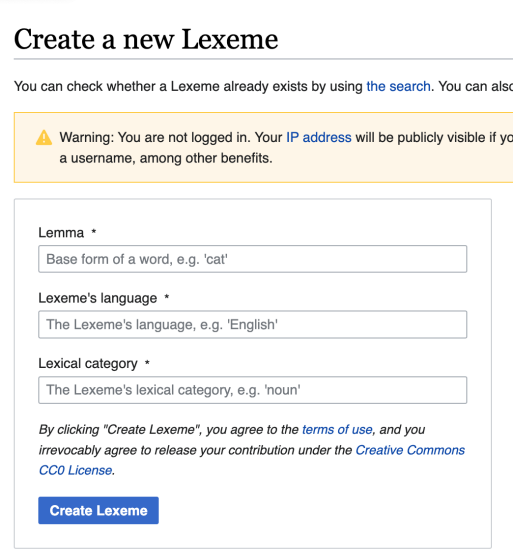


Figure 2. Current Wikidata interface aimed at expert users

and how do they compare in effectiveness to the existing interface used for data collection on Wikidata?'

To investigate the research question, two gamified interfaces were created. One interface is a basic version of the game with just a points system. The other interface is an advanced version of the game that has the same points system and also features various gamification elements such as a leaderboard, feedback screens and daily streaks.

Similar to other gamified interfaces (which are reviewed in the related works section), the two interfaces explored in this study ask for word translations. However, they also ask users for noun class classifications, which have not been explored in existing literature. Given that Niger-Congo B languages are agglutinative, noun classes play a critical role in sentence formation [12], and are important pieces of information to collect for the sentence generation part of the Abstract Wikipedia project.

To highlight the importance of this research, it is worth looking at the existing disparities in lexicographical data on Wikidata between different languages. English boasts over 73,000 lexemes [19] while languages such as isiZulu, isiXhosa, and Shona have lexeme counts of 92, 5, and 3, respectively [20–22]. Gamified interfaces offer a potential solution for this issue.

We begin with a review of relevant literature, followed by a design and implementation section that has a detailed explanation of the gameplay and interface design, as well as the architecture used. The experimental setup is then outlined, leading to a discussion of the results and their implications, along with recommendations for future research. The paper concludes with a summary of the key findings.

2 RELATED WORK

There are several works that discuss the use of crowdsourcing through gamified interfaces to collect lexicographical data. The concept of crowdsourcing using a gamified interface relates to the concept of a "game with a purpose" (GWAP) [6]. Crowdsourcing involves dividing lexicographical data collection into smaller tasks which are called microtasks [4, 14]. Microtasks are easy and simple tasks to do and can be completed by non-experts [4]. For example, a language expert could be tasked with translating many paragraphs from multiple stories while a micro-task could be as simple as translating individual words one by one by multiple contributors who are not experts. This expands the pool of contributors as there are very few language experts around, especially for low-resourced languages (LRLs) such as African languages [4].

GWAPs differ from other interfaces used to collect lexicographical data (for example Excel sheets and the current Wikidata interface) through their use of entertainment [6]. It makes data collection into a game that focuses on user engagement while collecting lexicographical data in the background [6]. This makes contributing lexicographical data more appealing to people beyond the typical contributors - such as linguistic experts, researchers and language enthusiasts.

In the literature surveyed about GWAPs that collect lexicographical data, two prominent themes emerged - the validation of user input and maintaining participant motivation (especially over extended periods of time). GWAPs use several validation techniques to check the accuracy of inputted data by users as inaccurate data inputted by users is a serious risk to the quality of the data [10]. This includes gold standard validation, where user input is checked against sample expert input as a test to see if the user is inputting correct answers, and inter-annotator agreement validation, where multiple annotators have to agree on the same input for it to be correct [7].

For participant motivation, most GWAPs primarily rely on social motivation - such as leaderboards, rankings, and other competitive elements - as well as psychological motivation - such as altruism [7]. Maintaining long-term contributor motivation is crucial, especially for LRLs with limited contributor pools, so participant motivation is important for a project's long-term success [7]. Another complementary issue to this is attracting users to play the game initially.

A study on a GWAP for collecting isiXhosa lexicographical data [18] found that monetary incentives strongly drove motivation and without it, there was little participation from users. It also found that leaderboards did not have a predicted positive effect on user engagement as this intimidated users which reduced user engagement [18]. This was the only related study conducted in South Africa, which is where the

experiments of this study will be conducted. Therefore, its findings are highly relevant to this research project.

There are several other gamified interfaces for collecting lexicographical data - the most notable ones being the COMO app [13], MachtSinn [2] and the Distrbuted Game [3] which all focus on collecting lexicographical data for European languages for Wikidata. Beyond the GWAP for collecting isiXhosa lexicographical data [18] mentioned earlier, the only other literature related to collecting lexicographical data for African languages is the Kamusi project [5].

Other notable crowdsourcing platforms include Amazon's Mechanical Turk and Crowdflower [6]. They differ from GWAPs as users are paid to do microtasks (monetary incentives are the main drivers of motivation) [6]. Amazon's Mechanical Turk is good for tasks that require English speakers and are non-linguistic [18] and the same applies to Crowdflower. These platforms are not feasible options for linguistic tasks for African languages as they lack speakers of these languages.

3 DESIGN AND IMPLEMENTATION

The two games were developed in three phases - first by designing the gameplay, then the frontend interfaces and then the software architecture. The basic gamification game is called "Word Safari Explorers" while the advanced one is called "Word Safari Champions".

3.1 Gameplay Design

During the gameplay design phase, the features and rules of the games were chosen.

3.1.1 Game Questions as Microtasks. Both games ask users the same three types of questions: translating, checking translations and identifying noun classes. All three questions are micro-tasks as they require simple inputs from the user and can be completed by non-experts. For the translation questions, the user is presented with a word and then must type the translated word in their chosen language. The checking translation questions show the user a translation (that another user has submitted) and then they must tap yes or no to agree or disagree with the translation. The use of yes or no questions reduces the cognitive effort required by the user and allows for words to be validated quickly [4]. The identifying noun class question shows the user a translated word and they must select the correct noun class from a drop-down list.

3.1.2 Gamification Elements and Motivation. Gamification elements were selected to increase user motivation. Positive feedback increases intrinsic motivation while negative feedback reduces it [18] and elements were selected with this in mind.

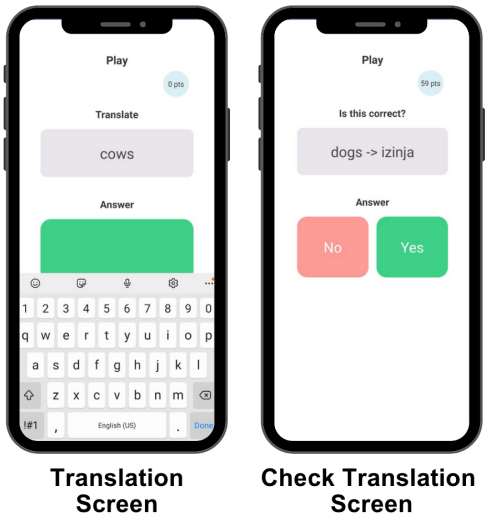


Figure 3. Translation and check translation interfaces showing how these questions are asked as microtasks (the questions asked are in isiZulu)

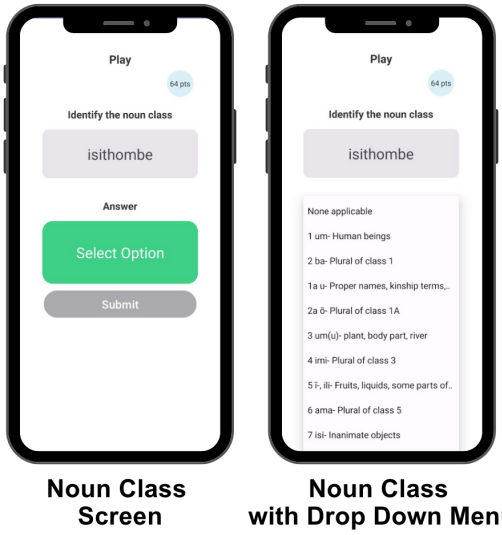


Figure 4. Noun class interface with and without the drop-down list showing how this question is asked as a microtask (the question asked is in isiZulu)

Both games have the same points system (which provides positive feedback). The more advanced game has a leaderboard, daily streaks and feedback screens. A literature review that summarised 24 studies found that points and leaderboards are some of the most common elements used for gamification and that leaderboards are effective [11]. Another study by Packham and Suleman [18] found that leaderboards were not effective as they intimidated users and reduced

motivation. To investigate this further, a leaderboard was included in the advanced gamification interface as the main gamification feature.

Another study about gamification elements found that visible status (this refers to the public display of a user's achievements) and rapid feedback are widely used [9] and this was applied to the advanced gamification game. Visible status, which allows users to feel a sense of social credibility and recognition [9], is implemented through the use of a leaderboard with avatars. Feedback screens and daily streaks were also included as rapid feedback elements which act as instant rewards for users. Two types of feedback screens were included one that appears each time users answer 25 questions and another one when they answer their first 100 questions of the day (which is the requirement to meet the daily streak).

It is important to note that users who are extrinsically motivated by money often prioritise quantity of answers over quality which should be accounted for in the design of the game by thoroughly validating the collected data [4, 5].

3.1.3 Points System. The points system, common to both versions, operates as follows: users receive three points for each successful translation, one point for verifying an existing translation, and two points for accurately identifying a noun class. This tiered point structure acknowledges that some questions demand greater cognitive effort than others (for example, checking a translation is easier than doing a translation).

3.1.4 Validating Data Inputted by Users. Two validation techniques were used to ensure the accuracy of users' answers: gold standard validation and inter-annotator agreement validation. Gold standard validation is implemented in the tutorial, where users are required to answer a series of questions: 10 translation questions, 6 checking translation questions, and 6 noun classification questions. Each category generates a score, and users with insufficient scores are flagged and their future answers are checked to see if they are correct.

Inter-annotator agreement validation requires agreement from at least five users for a particular contribution to be accepted. It's important to note that a lack of agreement does not necessarily mean an answer is incorrect; it may highlight a term with multiple meanings or a word that is vague. It is important to note that if users don't agree that it doesn't mean a word is wrong, it might be an ambiguous word.

3.1.5 Gameplay Mode. There is only one gameplay mode available in both games so as not to overwhelm users with different gameplay options. Since an element of gamification is randomness [7], users are asked different types of questions in a randomised order. Noun and check questions have a 5/11 chance of each being asked approximately and translation questions have a 1/11 chance of being asked approximately. These ratios were selected as five checks are required to validate a translation and five noun classifications to validate a noun class. There are three exceptions to this randomised order described. Firstly, when the game is initially played by users - the first 150 questions are only translation questions in order to build a base for generating other question types (since they are dependent on the answers of the translation questions). Secondly, a user cannot receive a check translation question for a word they've translated so if a user has checked all possible words available, they will be given a translation question instead. Thirdly, when a user first starts the gameplay, they will always get a translation question.

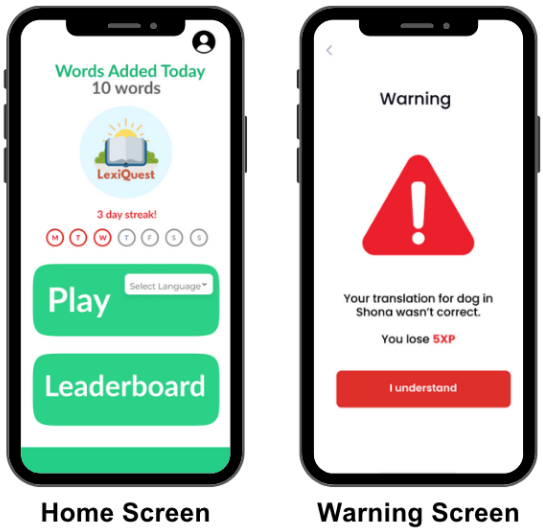


Figure 5. Initial designs for the home and warning interfaces were discarded

3.1.6 Languages and Data for Questions. IsiZulu, isiXhosa and Shona were selected as the languages the games can be played in as it will be feasible to get participants for the experiments who speak these languages. The games will ask users to translate and classify nouns from a list of 1000 of the most frequent nouns from the English Wikipedia Corpus [8] as these would be valuable lexemes to add to Wikidata. Once this initial list is exhausted, additional nouns will be sourced from alternative lists for continued gameplay.

3.2 Interface Design

During the interface design stage, the frontend interfaces of the games were developed with user feedback from a prototype.

3.2.1 Prototype Evaluations. A high-level prototype of the game was initially developed in Figma to visualise the interface and gameplay elements. Studies have shown that the most effective number of participants for testing prototypes ranges from 3 to 5 [15], hence the prototype was evaluated by 3 participants who spoke isiZulu. Ethics clearance was first received before the 3 participants evaluated the prototype and participants also signed consent forms. they were asked to look at the prototype screens and voice their thoughts over a call or voice note and then to fill in a form with their thoughts after doing this. The former is called the "thinking aloud" method which is an easy-to-implement usability engineering method used to spot usability issues [17]. These participants provided feedback on how the gameplay mechanics and design elements can be improved in terms of usability. In response to this feedback, iterative changes were made during the subsequent development phase in Android Studio.

3.2.2 Changes Implemented. Changes included redesigning the home screen, warning screen and some design elements along with other changes. The home screen was redesigned as the original one was cluttered according to users. The warning screen for incorrect answers was found too harsh so it was redesigned into a less harsh version and added to the tutorial to reduce the amount of negative feedback directed at users (as discussed earlier such feedback reduces intrinsic motivation). Users who need warnings for inputting incorrect answers will be sent a friendly email to try to reduce the negative impact of this feedback. A tutorial section was added to show users how to play the game, as some feedback brought up concerns that the gameplay was not intuitive in the initial prototype. Other design changes were using better colours and icons as some of it was flagged in the feedback as inconsistent with what users are used to. Informal feedback after the prototype phase was also collected later from potential users during the software development phase of the games. This resulted in the original name for the games being changed from LexiQuest (referring to the lexicographical data collection of the games) to Word Safari as it is more understandable for users. Users were also not aware of the skip button for questions and a screen about it was added to the tutorial to make it more clear.

3.2.3 Usability Concepts Applied. In the design of the interface, several usability concepts were applied such as Nielsen's 10 Usability Heuristics [16]. For example the "recognition rather than recall" principle was used to minimise the user's memory load by including the drop-down list for noun class questions (instead of asking users to type the noun class). This also relates to one of Schneiderman's Eight Golden Rules called reducing short-term memory load [23]. The "match between system and the real world" principle was also applied as jargon was substituted for more relatable concepts for users (for example there is no mention of lexemes, the Abstract Wikipedia explanation is simplified in the tutorial and the name of the games were changed to Word Safari from LexiQuest). Other principles applied include "aesthetic and minimalist design", "visibility of system status" and "help and documentation".

Schneiderman's rule of 'offer simple error handling' was applied as users receive pop-up messages if they're trying to perform an action incorrectly with feedback on what they need to do (for example, not selecting a language option in the tutorial or inputting password that is too short - users will receive error messages with the correct action). Schneiderman's rule of 'strive for consistency' is also applied in several ways - for example, through the use of a hamburger icon for the menu and consistent colour schemes, such as green for 'yes,' red for 'no,' and orange on feedback screens, all of which are familiar to users.

3.2.4 Final Interfaces. All of the above is evident in the final interfaces for both games - which are included in the appendix for further reference and the interfaces are briefly described here.

Word Safari Explorers (the basic gamification application) has welcome, sign up and create account screens. Once a user has created an account, they are taken to the tutorial screens where the game is explained, they can select the language they want to play the game in and they then do the tutorial questions (gold-standard validation happens here to check their language proficiency). They are then taken to the home screen where there is a play button (which goes to the gameplay mode) and a menu button. In the gameplay mode, they are presented with translation, checking translation and noun classification questions which all have individual screens as shown earlier. The menu screen has 3 options to go back home, a help section and a logout option.

Word Safari Champions (the advanced gamification application) includes the same screens as the Explorers version and additionally has a leaderboard screen (where users are ranked by their points), avatar selection screens (both in the tutorial and menu screens), two types of feedback screens (one for every 25 questions answered and another one for when they reach their daily streak by answering 100 questions) and a daily streak on the status bar on the home screen.

3.3 Architecture

The architecture of both applications is quite similar - with the backend database structure being largely the same. The frontend is also similar with the exception that the Explorers version does not have all the features that the Champions version has such as the leaderboard, feedback screens, daily streak and avatar screen.

3.3.1 Architecture Choices. Both applications were developed in Android Studio, using Java for the frontend as the researcher was most familiar with this approach compared

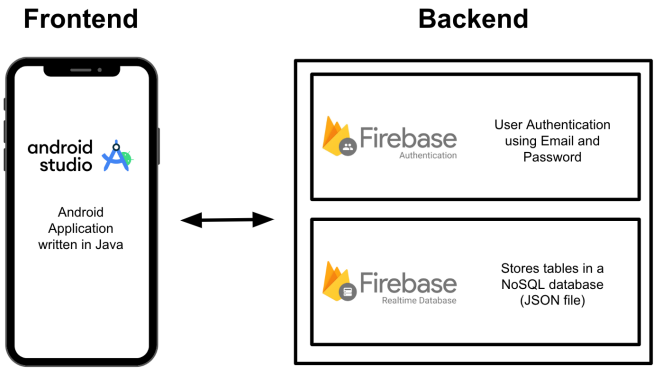


Figure 6. Diagram showing the frontend and backend of both applications

to other ones. For backend functionality, Firebase Authentication and Realtime Database were used. Firebase Authentication allows for easy user verification by asking users for their email and password and was simple to integrate into the Android Studio application. Firebase Realtime Database enables real-time updates for multiple users, something that is important for the overall game concept. This real-time functionality is essential when considering the gameplay mechanics - users need to receive different words to translate (the same word in a language cannot be given to multiple users), points need to be refreshed quickly and the leaderboard must also be updated quickly. Firebase Realtime Database allows for the quick response time needed and this along with its ease of integration into Android Studio made it ideal for the backend for the two applications.

The database for both applications employs a NoSQL architecture, storing data in a JSON (JavaScript Object Notation) format, which is simple to read and allows for flexibility. This flexibility does not require a fixed schema (unlike SQL) which was advantageous as this meant tables could start with a simple structure and fields could be added on later when needed. For example, when a user submits a word translation, the initial 'Translations' object for the selected language is populated. When this same translation is asked as a noun classification question, a "nounMap" field (which contains a list of noun class answers selected by users and how often they were selected by users) is added to the existing 'Translations' object to reflect this new noun class information.

3.3.2 Activities. There are two types of activities in both applications: interface activities (which may optionally engage with the Firebase Realtime Database or Authentication depending on the specific activity in question) and classes that model tables (this serves as a blueprint for how the data is structured in the Firebase Realtime Database). Given that a NoSQL database is used, custom classes are necessary to define the objects that map to database tables (for example, the isiZuluTranslations Class generates the relevant object for the database).

For the Explorers application, there are a total of 35 activities. Out of these, 20 of these are interface-related and 15 are classes for the database. The interface-related activities can be categorised as 10 for the tutorial screens, 3 for the gameplay screens, 1 each for the home, menu and help screens, and 3 for the welcome, login, and sign up screens.

The Champions application has 44 activities, with 28 of these being interface-related and 15 are database-related. The extra 8 activities are for implementing the leaderboard, feedback, and avatar features of the Champions application.

3.3.3 Database Structure. The database schema for Word Safari is designed to account for various needs such as user information storage to tracking answers for the different language tables. The most important thing to note is how the Translation table for each language summarises all user input from all three types of questions for that one word (as shown in Figure 8).

An overview of the tables in the database is given below:

- Users: stores user-related information such as the user's email, first name, last name, points, word count, if they are flagged (from getting a low tutorial score) and selected avatar (only for Champions version)
- UserTutorialScore: stores users' tutorial scores for the three types of questions in the tutorial - translate score, check score and noun score
- isiZuluData: stores questions to be asked
- isiZuluTranslations: stores users' answers to translation questions and also stores corresponding totals for yes/no and total checks submitted by other users along with noun class answers for that word
- isiZuluChecks: stores users' individual answers to checking translation questions
- isiZuluNounClass: stores users' individual answers to noun classification questions
- Metadata: stores numbers of words for each type of language table

For Shona and isiXhosa, the same four tables exist as shown for isiZulu above and store their language data - their table descriptions are omitted for brevity.

The tables do not store the user's daily streak and current selected language as these were coded as "Shared Preferences" that are stored on the user's phone (and not in the database). Shared Preferences provides fast access for small data and is also easy to implement, making it well-suited to store the selected language and daily streak of users.

The only difference between the Explorers and Champions databases is that the Champions Users table has an avatar field while the equivalent table in the Explorers version does not.

```
"Users": {
       "user1@gmail,com": {
              "avatar": "avatar6",
              "email": "user1@gmail.com",
              "firstName": "User",
              "flagged": true,
              "lastName": "One",
              "points": 83,
              "wordCount": 21
}
```


Figure 7. Example of a Users Object


```
"isiZuluTranslations": {
      "bag-isikhwama": {
             "ansWord: "isikhwama",
             "checksCount: 5,
             "engWord: "bag",
             "noCount: 0,
             "nounCount: 2,
             "nounMap: {
                    "7 isi- Inanimate objects": 3
             },
             "yesCount: 5
      }
}
```


Figure 8. Example of an isiZuluTranslations Object

3.3.4 Processing Data to Find Valid Lexemes. To find the valid lexemes to upload to Wikidata, the Translation Tables for each language will be filtered to select words with high numbers of yes counts (which are generated from the checking translation questions). Since noun class classification is a new feature being tested, lexemes and their associated noun classes will be manually inspected to see if they have the correct noun class before they are uploaded.

4 EXPERIMENTS

Two experiments were carried out to assess the impact of gamification on the amount of lexicographical data collected.

4.1 Aims

There are two aims of the experiments. The first aim is to evaluate if including more gamification elements in an interface increases user motivation to contribute more lexicographical data. The second aim is to see how much lexicographical data can be collected for isiZulu, isiXhosa, and Shona compared to the current Wikidata interface.

4.2 Hypotheses

To test the aims of the study, the following two hypotheses are formulated:

1. The more gamified interface will result in higher levels of user motivation and will collect more lexicographical data than the basic gamified interface.
2. Both gamified interfaces will collect more lexicographical data for isiZulu, isiXhosa, and Shona compared to the current Wikidata interface.

4.3 Procedure

Two experiments were conducted, as described below, to test the two hypotheses.

4.3.1 Materials. The Android Package Kit (APK) files for the two gamified interfaces and the participants' mobile phones were used for this study. Consent forms, from the ethics clearance that was done before the experiments were run, were also filled in by participants agreeing to participate in the study.

4.3.2 Participant Recruitment. Twenty participants were recruited from the researcher's WhatsApp through snowball sampling, which was chosen for its efficiency in quickly acquiring participants. Participants consisted of isiZulu and isiXhosa speakers. Most participants were university students at the University of Cape Town and the University of the Witwatersrand, while a minority were employed as service and industrial workers.

4.3.3 Incentives. To encourage participant engagement, two incentive schemes were implemented across the two experiments. In experiment one, a random selection of five participants (who play for a minimum of 10 minutes a day) will receive a R200 cash prize. In experiment two, which included the interface with the leaderboard, the top five participants by game score will be rewarded with a R200 prize each.

4.3.4 Methods. The twenty participants were split into two groups (of ten participants each) for the two experiments. Experiment one used the basic gamification interface while experiment two used the advanced gamification interface (with the leaderboard).

Participants were emailed or messaged on WhatsApp (depending on their personal preference) the game instructions along with a link to download the respective APK for the experiment in which they were participating in. The study was run over 5 days from 5 to 9 September. Participants for

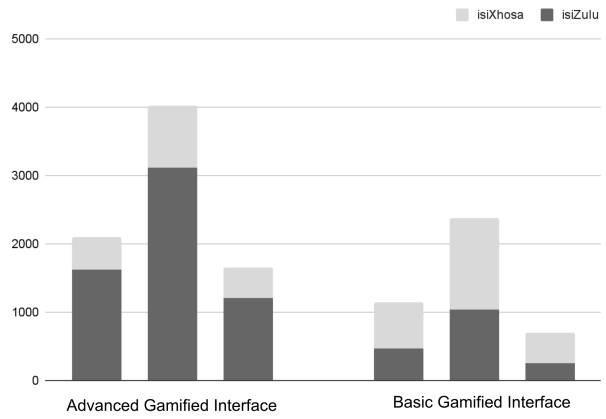


Figure 9. Graph showing how many different types of questions were answered by users for each interface

both experiments were asked to play for at least 10 minutes a day.

Participants' answers to the three different types of questions (translation, checking translation and noun class classification questions) were collected and stored in the databases as outlined in the architecture section above. This data was then analysed.

4.3.5 Data Analysis. Using the data collected from users during the two experiments, data analysis was done to compare how many questions were answered between the different interfaces.

Table 1. Number of questions answered (categorised by type) in the basic gamified interface

	isiZulu	isiXhosa	Total
Translate	1623	480	2103
Check	3116	913	4029
Noun Class	1208	445	1653

Table 2. Number of questions answered (categorised by type) in the advanced gamified interface

	isiZulu	isiXhosa	Total
Translate	474	671	1145
Check	1029	1327	2376
Noun Class	256	442	698

The ratio of translation questions to the other two questions is higher than expected because many users had completed all the check translation questions available to them and were then given more translation questions.

4.3.6 Qualitative Analysis. In addition to the data analysis, qualitative analysis was also done to note interesting observations.

Upon manual inspection, most of the translated words are accurate even if the full numbers of 5 "yesCounts" hasn't been met fully. Words with a "noCount" of 1 or more, usually meant the translation was wrong (there were very few vague terms in the data asked to users as questions and it did not appear any of the no answers were due to ambiguity around the word translated). For the noun class questions, a large numbers of answers inputted by users were incorrect, especially by users with the lowest tutorial scores.

The noun class classification question confused participants as they used descriptions on the drop-down list to try to classify words instead of using the prefix. For example, a user who classified the noun "isityalo" (which means plant in isiXhosa) selected "3 um- plant, body part, river" instead of "7 isi- Inanimate objects" since the given description matches better even though the prefix is isi-. Along with this, it also seems some users randomly guessed some of their noun class answers.

Tutorial scores were analysed in the context of each user's performance. Most users had good tutorial scores (most users scored 8+/10, 6/6 and 4+/6 for translation, check translation and noun class questions respectively in the tutorial) and contributed data of high quality most of the time for translation and check translation questions.

High-risk users with low tutorial scores were more likely to input incorrect data, specifically for noun class classification questions. One particular user with a low tutorial score, playing the advanced gamified interface, had inputted many noun class answers that were incorrect. It should be noted this user stayed in the top 3 spots on the leaderboard throughout the experiment. This can also be linked to them having high extrinsic motivation to get the R200 prize associated with the leaderboard and some intrinsic motivation to stay at the top of the leaderboard. Another user in this same game also displayed the same behaviours but did not answer as many questions as the original user mentioned.

Another finding specific to the advanced gamified interface is that on the leaderboard, users who were in the top 5 spots had scores in a close range of each other while users in the bottom 5 spots had scores in a larger range of each other indicating that the top 5 players were consistently playing the game throughout the study as they were motivated to maintain their spots while the lower ranked users were playing the game a lot less.

Informal interactions with participants during the study showed that most participants signed up due to the monetary incentives offered by the study, especially for the advanced gamification interface. This shows that the main motivation to play the games was not due to the gamification elements selected but by the monetary incentives included.

4.3.7 Results. Both of the hypotheses hold true.

For the first hypothesis, the more gamified interface did result in higher levels of user motivation leading to more lexicographical data being collected on this interface but there were quality issues surrounding the quality of data for noun class classification questions. There is a difference in the numbers of contributions between the two interfaces as seen above, with the more gamified interface having a larger number of contributions.

For the second hypothesis, both of the gamified interfaces individually collected more lexicographical data for isiZulu and isiXhosa compared to the current Wikidata interface. As mentioned earlier, Wikidata only has 92 and 5 lexemes for isiZulu and isiXhosa respectively and both games collected many more lexemes than this. The basic gamification interface collected 474 lexemes for isiZulu and 671 lexemes for isiXhosa while the advanced gamification interface collected 1623 lexemes for isiZulu and 480 lexemes for isiXhosa.

5 DISCUSSION

The results of the above experiments fall in line with related literature.

Gamification did increase user motivation and consequently, the number of contributions made by users and the more gamified interface with the leaderboard did collect more words and this is consistent with the gamification theme in related literature [11]. The leaderboard encouraged users to submit more contributions for its associated interface which relates to literature saying leaderboards are effective [11]. It's important to note that the leaderboard mainly motivated users at the top to contribute more and it seemed to decrease motivation in users in the lower ranks. This links to the study conducted by Packham and Suleman [18] which found similar user behaviours in one of the experiments they carried out with a leaderboard.

Monetary incentives were users' primary motivation to participate and this links to the findings of the study conducted by Packham and Suleman [18]. They found in low-resource settings like South Africa, users only participate when monetary incentives are offered [18]. This was the same case here as it is unlikely most of the participants would've participated without the monetary incentives offered.

Related literature discusses two types of validation techniques - gold standard and inter-annotator agreement validation [7]. Both validation techniques picked up when there was a risk of data being invalid in the games. In this study, gold standard validation flagged users who gave inaccurate answers (in the case of noun class classification questions) while inter-annotator validation flagged submissions that were mostly incorrect.

Related literature also discusses how extrinsic motivation related to monetary incentives can comprise data quality [4, 5] as demonstrated by the instance of the one user inputting many incorrect noun class answers on the advanced gamification interface.

5.0.1 Limitations of Study. This study has several limitations such as its short duration, small number of participants and the snowball sampling method used. The short duration of just five days does not provide insights into the long-term behaviours of users playing the games. This is an important consideration as most GWAPs struggle with long-term motivation as most users lose interest over extended periods of time.

The limited pool of 20 participants (who were mostly university students) does not represent the larger population of isiZulu and isiXhosa speakers especially since snowball sampling was used to recruit participants and this method is known for introducing bias. Considering all of the above, the results of this research cannot be generalised.

5.0.2 Future Work. Future work could focus on three areas: designing more effective ways to ask for noun class classifications and other similar lexicographical information, optimising incentives and targeting specific user groups such as students learning the language.

Many participants struggled with answering the noun class questions accurately because even though they subconsciously use it when speaking the language, they don't know how to explicitly identify noun classes when asked to do so. With this in mind, implementing an in-depth tutorial or a learning lesson about noun classes in the game that refreshes their memory about noun classes could be a possible solution. Questions could also be specifically designed to reduce errors in user input - for example, using the prefix of a word, users can be given a small list of possible noun classes that also start with the same prefix. Another example would be implementing a checking question for noun class classification questions.

Monetary rewards seem the most effective in low-resource settings like South Africa, and alternative incentives like data packages or vouchers could also be explored and ways of lowering the costs of incentives while maintaining user motivation can also be investigated.

Targeting specific groups, such as high school students learning isiZulu and other languages, could provide users who would play such lexicographical games for reasons not associated with extrinsic motivation. For example, a learning app (similar to the translation game but with heavier validation) could be created for this group to collect lexicographical data from these users without having to implement monetary incentives and teachers can be involved in validating data submitted by students.

6 CONCLUSIONS

In this study, both interfaces collected more lexicographical data compared to Wikidata and the advanced gamification interface collected more data than the basic gamification interface. Although gamification did increase user motivation to contribute more data, this was most likely a result of gamification being directly tied to monetary incentives especially, for the advanced gamification interface with the leaderboard. The leaderboard had a positive effect on user motivation but also introduced quality issues (especially in the noun class classification answers) as some users prioritised the number of answers inputted to have a high ranking on the leaderboard and this can be mitigated by implementing validation techniques.

Overall, gamified interfaces do offer a good solution for collecting more lexicographical data as compared to the current Wikidata interface as it means more non-expert users can contribute and more gamification elements in an interface (such as using a leaderboard and feedback screens) results in more user motivation leading to more contributions.

While the limited number of participants and the short duration of the experiments mean definitive conclusions should not be drawn from this study, its findings do line up with existing literature and this can drive future research in the field.

References

[1] 2023. Abstract Wikipedia. https://meta.wikimedia.org/wiki/Abstract_Wikipedia

[2] 2023. MachtSinn – Das macht doch alles keinen Sinn! https://machtsinn.toolforge.org/

[3] 2023. Wikidata - The Distributed Game. https://wikidata-game.toolforge.org/distributed/#mode=stats

[4] Martin Benjamin. 2016. Crowdsourcing microdata for cost-effective and reliable lexicography. In *Proceedings of the 9th ASIALEX Conference*. Hong Kong.

[5] Martin Benjamin and Paula Radetzky. 2014. Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining. In *Expert Input, Crowdsourcing, and Gamification Acquiring Lexical Data for LRLs. 9th edition of the Language Resources and Evaluation Conference*. https://infoscience.epfl.ch/record/200375

[6] Jon Chamberlain, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade, and Massimo Poesio. 2013. *Using Games to Create Language Resources: Successes and Limitations of the Approach*. 3–44. https://doi.org/10.1007/978-3-642-35085-6_1

[7] Jaka Čibej, Darja Fišer, and Iztok Kosem. 2015. The role of crowdsourcing in lexicography. *Electronic lexicography in the 21st Century: linking lexical data in the digital age. Proceedings of the eLex* 2015 (2015), 70–83. https://elex.link/elex2015/proceedings/eLex_2015_05_Cibej+Fiser+Kosem.pdf

[8] English Corpora. 2014. *Wikipedia Corpus*. https://www.english-corpora.org/wiki/ 1 September.

[9] Darina Dicheva, Christo Dichev, Gennady Agre, and Galia Angelova. 2015. Gamification in education: A systematic mapping study. *Journal of educational technology & society* 18, 3 (2015), 75–88.

[10] Darja Fišer and Jaka Čibej. 2017. The potential of crowdsourcing in modern lexicography. *Dictionary of modern Slovene: Problems and solutions* (2017), 212–228.

[11] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[12] C Maria Keet, Langa Khumalo, and Zola Mahlaza. 2022. Considerations for a model for NCB noun classes in Wikidata. (2022). https://wikiworkshop.org/2022/papers/WikiWorkshop2022_paper_31.pdf

[13] Maximilian Kristen. 2019. COMO: A LEXICOGRAPHICAL DATA STRUCTURING GAME WITH A PURPOSE. (2019). https://www.en.pms.ifi.lmu.de/publications/projektarbeiten/Maximilian.Kristen/PA_Maximilian.Kristen.pdf

[14] Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward More Meaningful Resources for Lower-resourced Languages. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 523–532. https://doi.org/10.18653/v1/2022.findings-acl.44

[15] Jakob Nielsen. 1994. Guerrilla HCI: Using discount usability engineering to penetrate the intimidation barrier. *Cost-justifying usability* (1994), 245–272.

[16] Jakob Nielsen. 2005. Ten usability heuristics. (2005). http://www.nngroup.com/articles/ten-usability-heuristics/(acc-essed~âĂę

[17] Jakob Nielsen. 2012. Thinking aloud: The 1 usability tool. (2012). https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

[18] Sean Packham and Hussein Suleman. 2015. Crowdsourcing a Text Corpus is not a Game. In *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, December 9-12, Proceedings 16*. Springer, Seoul, Korea, 225–234. https://doi.org/10.1007/978-3-319-27974-9_23

[19] Wikidata SPARQL Query Service. 2023. *Lexeme Count for English*. https://query.wikidata.org/#SELECT%20%28COUNT%28%3Flexeme%29%20AS%20%3Fcount%29%0AWHERE%20%7B%0A%20%20%3Flexeme%20dct%3Alanguage%20wd%3AQ1860%3B%20%23%20English%20lexeme%20count%0A%20%20%7D Accessed: August 17, 2023.

[20] Wikidata SPARQL Query Service. 2023. *Lexeme Count for isiXhosa*. https://query.wikidata.org/#SELECT%20%28COUNT%28%3Flexeme%29%20AS%20%3Fcount%29%0AWHERE%20%7B%0A%20%20%3Flexeme%20dct%3Alanguage%20wd%3AQ34004%3B%20%23%20Shona%20lexeme%20count%0A%20%20%7D Accessed: August 17, 2023.

[21] Wikidata SPARQL Query Service. 2023. *Lexeme Count for isiZulu*. https://query.wikidata.org/#SELECT%20%28COUNT%28%3Flexeme%29%20AS%20%3Fcount%29%0AWHERE%20%7B%0A%20%20%3Flexeme%20dct%3Alanguage%20wd%3AQ10179%3B%20%23%20Zulu%20lexeme%20count%0A%20%20%7D Accessed: August 17, 2023.

[22] Wikidata SPARQL Query Service. 2023. *Lexeme Count for Shona*. https://query.wikidata.org/#SELECT%20%28COUNT%28%3Flexeme%29%20AS%20%3Fcount%29%0AWHERE%20%7B%0A%20%20%3Flexeme%20dct%3Alanguage%20wd%3AQ13218%3B%20%23%20Shona%20lexeme%20count%0A%20%20%7D Accessed: August 17, 2023.

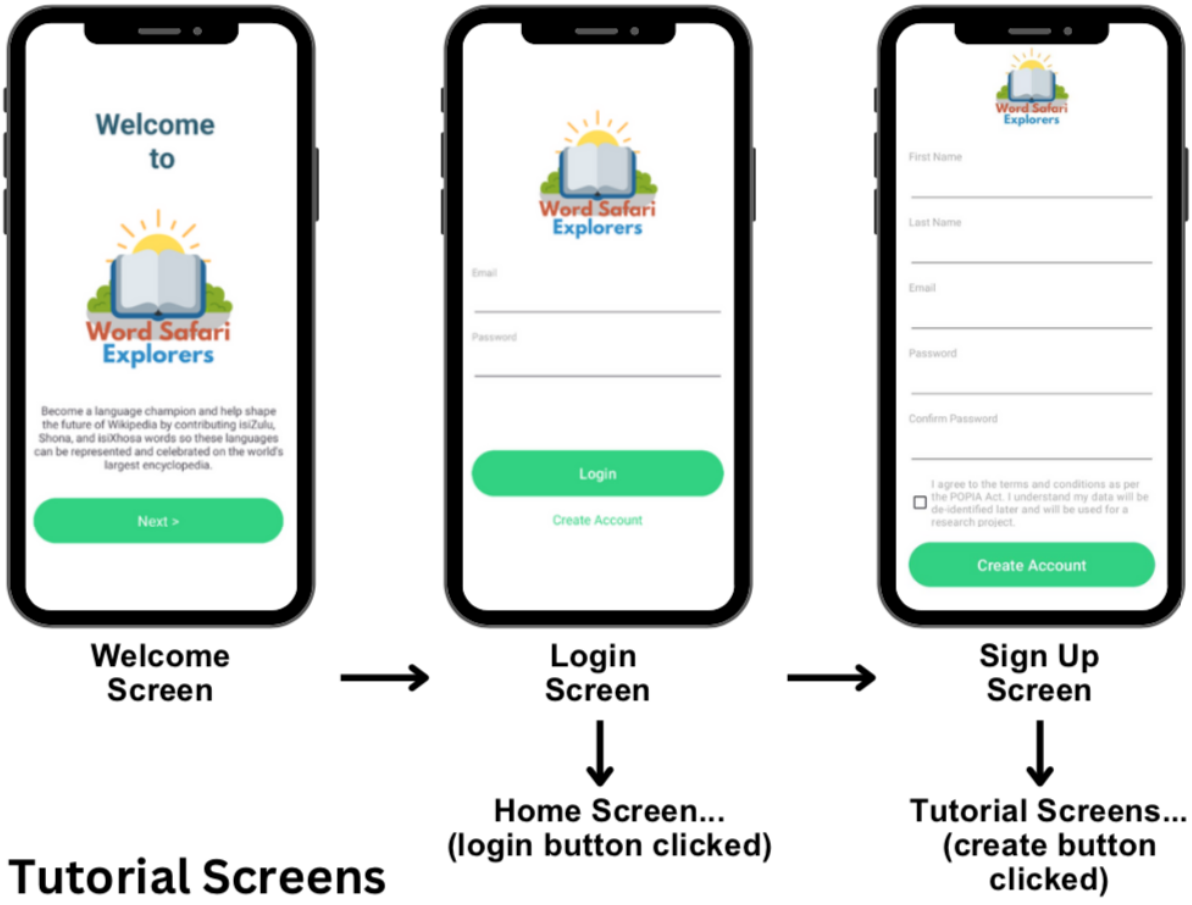
[23] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, and Niklas Elmqvist. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (sixth ed.). Pearson. http://www.cs.umd.edu/hcil/DTUI6

[24] Denny Vrandečić. 2020. Architecture for a multilingual Wikipedia. *arXiv preprint arXiv:2004.04733* (2020).

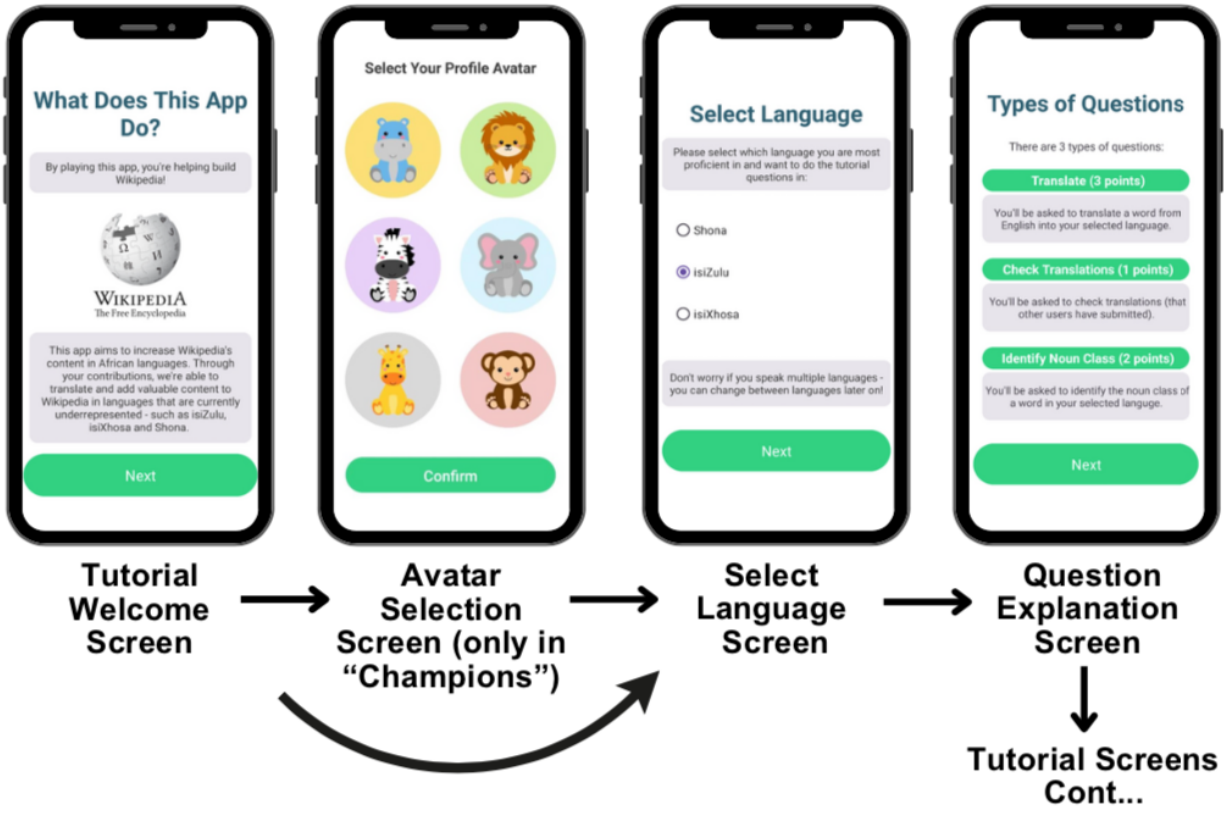
A Appendix

Please look at the next page to see the interface.

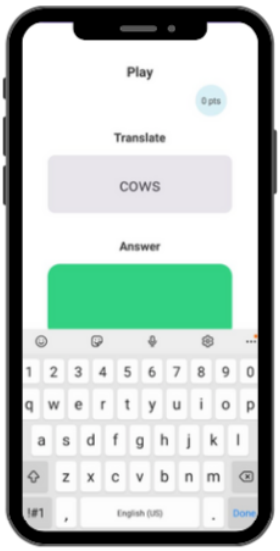
Welcome, Login and Sign Up Screens



Tutorial Screens

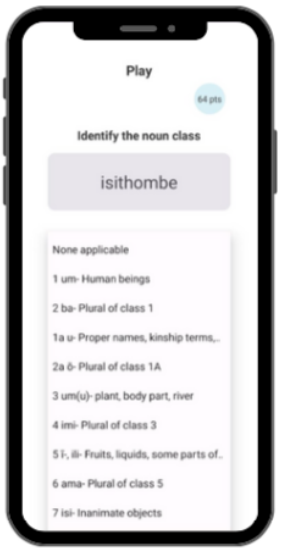


Tutorial Screens Continued

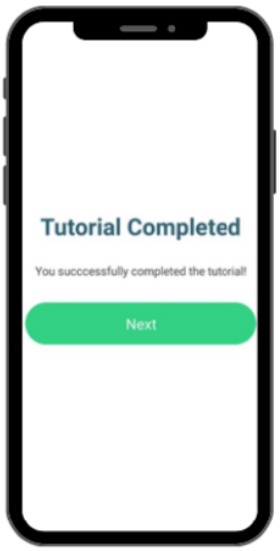


Translation Screen (10 questions asked)

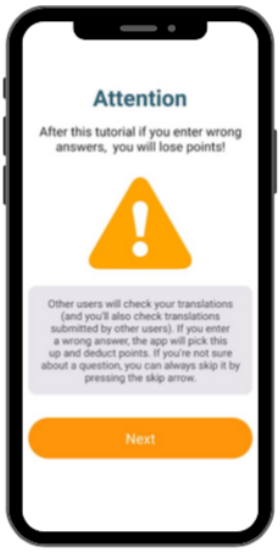
Check Translation Screen (6 questions asked)



Noun Class Screen (6 questions asked)

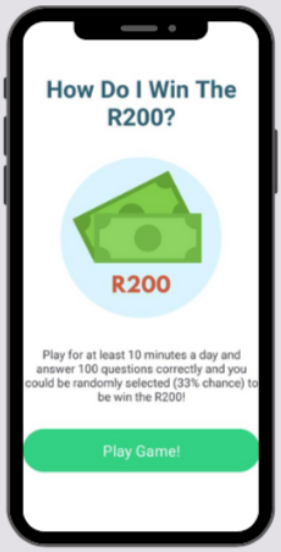


Tutorial Completed Screen

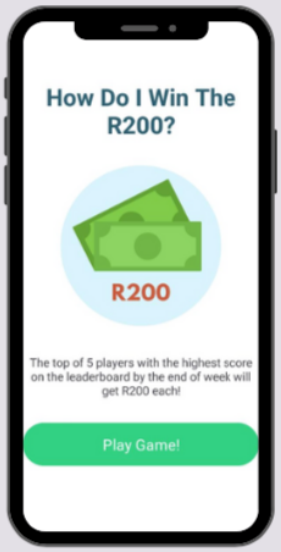


Tutorial Warning Screen

Skip Explanation Screen

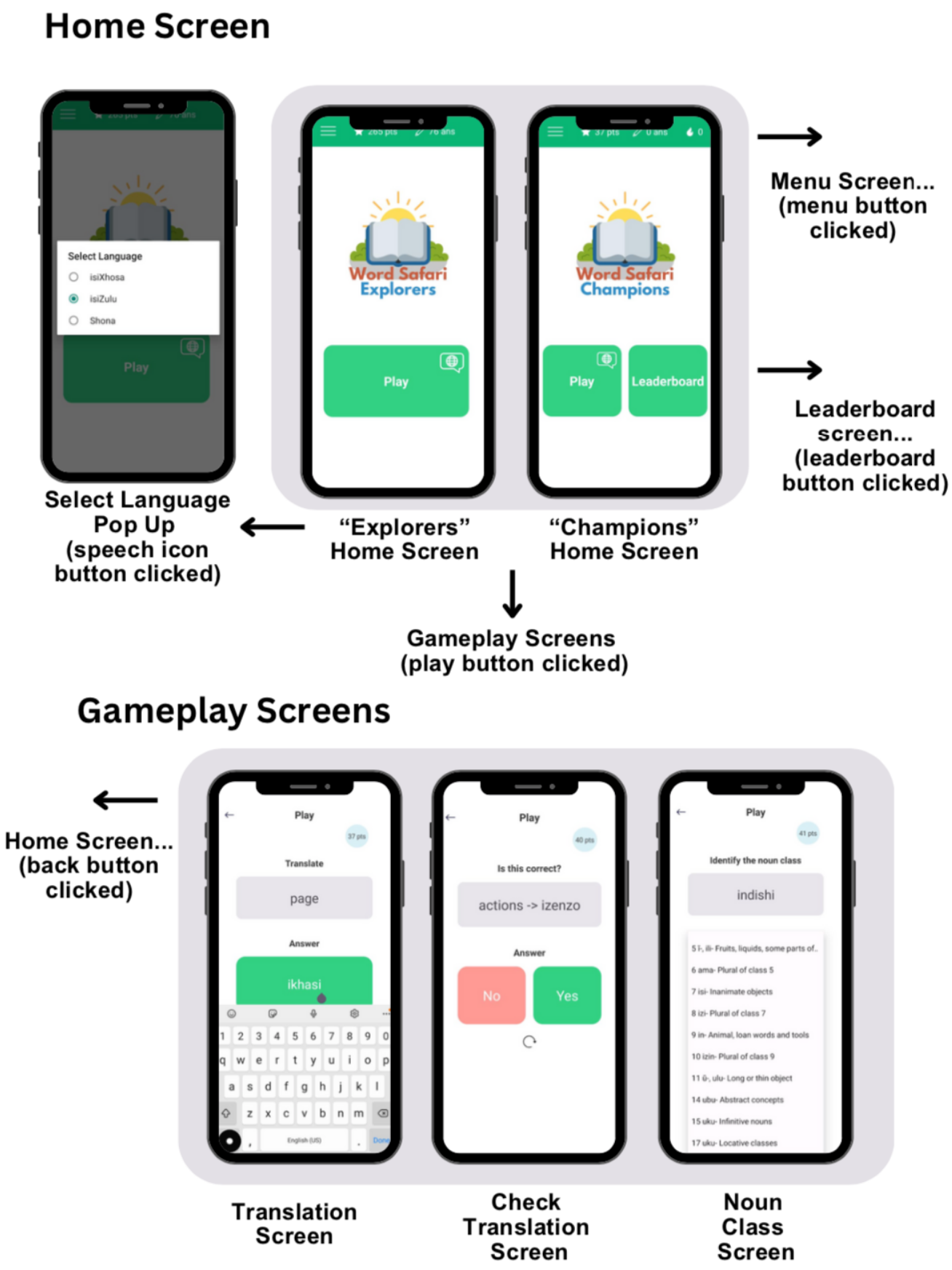


"Explorers" Prize Screen

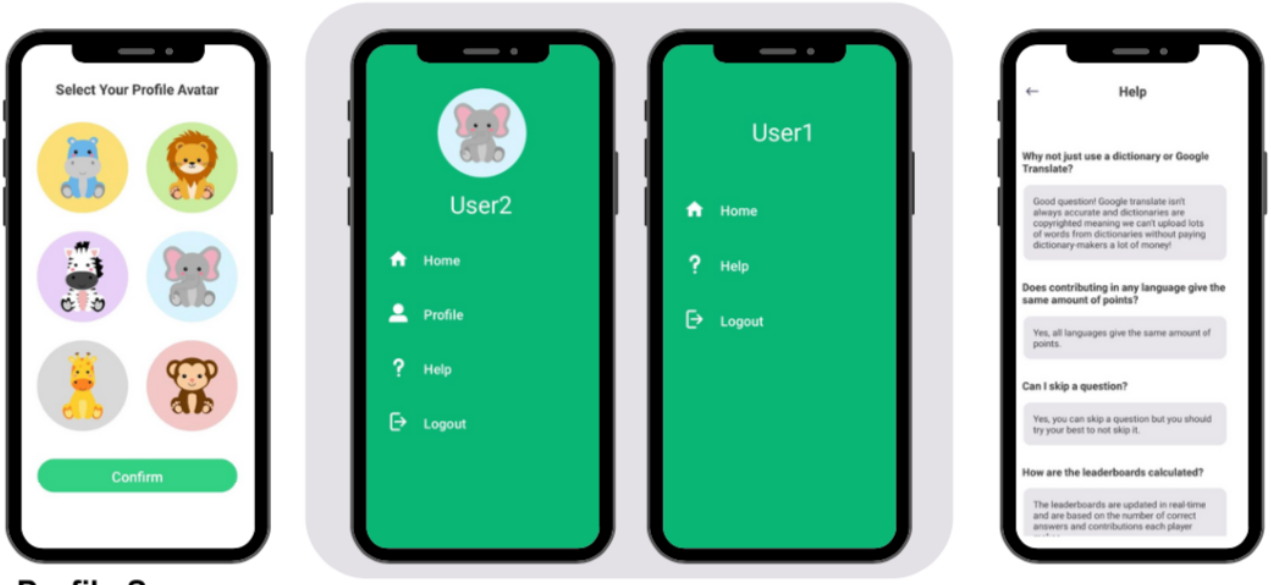


"Champions" Prize Screen

Home Screen...

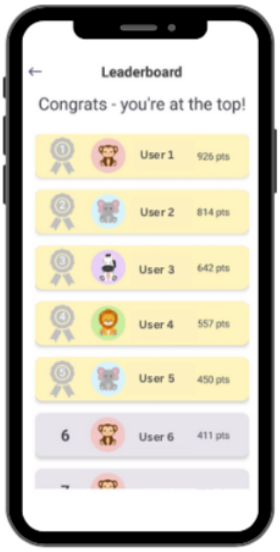

Home Screen
Select Language
isiXhosa
isiZulu
Shona
Play
Word Safari Explorers
Play
Word Safari Champions
Play
Leaderboard
Menu Screen... (menu button clicked)
Leaderboard screen... (leaderboard button clicked)
Select Language Pop Up (speech icon button clicked)
"Explorers" Home Screen
"Champions" Home Screen
Gameplay Screens (play button clicked)
Gameplay Screens
Home Screen... (back button clicked)
Play
Translate
page
Answer
ikhasi
Play
Is this correct?
actions -> izenzo
Answer
No
Yes
Play
Identify the noun class
indishi
Translation Screen
Check Translation Screen
Noun Class Screen
*Appears in a randomized sequence with approximate probabilities of 1/11, 5/11, and 5/11 for each screen respectively

Menu, Profile and Help Screens



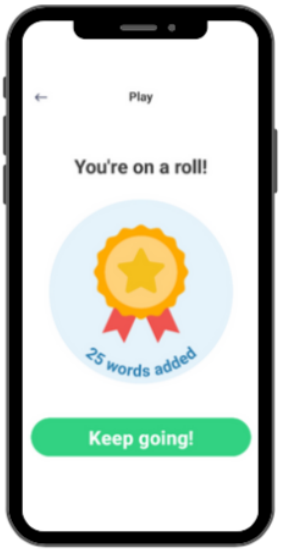
Profile Screen (profile button clicked - only in "Champions") ← **"Champions" Menu Screen** **"Explorer" Menu Screen** → **Help Screen (help button clicked)**

Leaderboard and Feedback Screens



Leaderboard Screen

***Only on home screen for "Champions"**



Twenty Five Words Added Screen

Daily Goal Met Screen

***Appears everytime 25 questions are answered and when 100 questions are answered for daily streak - only on "champions"**