Creating an Interface for Crowdsourcing Lexicographical Data for Low-Resourced African Languages for Wikidata

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

Abstract

Wikipedia has the goal of anyone in the world being able to share knowledge and Abstract Wikipedia aims to help fulfil this ambitious goal by allowing for the automatic generation of articles. This is more scalable than the current method of manually written articles and it would allow for the same article to be available in multiple languages. To do so requires extensive lexicographical data for these languages which Wikidata currently lacks, especially for African languages as they are low resourced. This data can be collected by using crowdsourcing. A literature review was conducted into using crowdsourcing for gathering lexicographical data. It found that crowdsourcing is well suited to do this for low-resourced languages as other methods require resources that are not available for such languages. The literature review highlights how quality issues in crowdsourcing can be overcome by validating collected data, different forms of motivation in crowdsourcing and how gamification can be used in crowdsourcing to keep contributors engaged. It then explores various crowdsourcing projects that have collected lexicographical data, both for Wikidata and other lexicographical databases, that have had promising results. Lastly, an analysis of the literature is given.

CCS Concepts: • **Information systems** → *Crowdsourcing*.

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

1 Introduction

The goal of Wikipedia is so everyone in the world can share knowledge [22], regardless of what language they speak. This goal is currently not being achieved as seen by the uneven distribution of languages on Wikipedia [16, 22]. The most prevalent languages such as English have very comprehensive Wikis while less common languages have much smaller Wikis and not all languages have a Wiki. Although there are 7000 languages spoken globally [7], Wikipedia only has 332 language editions. From these 332 language editions of Wikipedia, the top 18 language editions have over 1 million articles each while the bottom 208 language editions have less than 20 000 articles each [3]. For example, there are under 11 000 Wikipedia articles in isiZulu [3]. This uneven distribution means speakers of most languages cannot read most of the content on Wikipedia.

Wikipedia also inadvertently limits who can contribute to it as its current format only allows contributions to reflect in the language edition it is submitted to [22]. For example, a French speaker can write a comprehensive article for the French Wikipedia but their contribution will not be reflected on other Wikis, which limits who can view this information. African languages are heavily underrepresented as their Wikis are quite small (as seen by the isiZulu example above) and their language editions could benefit from having many more articles. Abstract Wikipedia aims to help with this (for all languages, not just African ones) and fulfil Wikipedia's ambitious goal of knowledge sharing for all by automating articles. This will reduce the effort required to generate short articles in different languages as they do not need to be manually written, which will allow Wikis for African languages to be expanded. Abstract Wikipedia will be utilising a language-independent format in order to do this.

Abstract Wikipedia's language-independent format will make use of Wikidata and Wikifunctions [1]. Wikidata is a database that stores information as structured data using Q-items, P-items and L-items and Wikifunctions helps render data from Wikidata into natural language. Q-items refer to entities while P-items refer to properties and P-items are mapped to Q-items. For example, the Q-item Q8023 refers to Nelson Mandela and has the P21 (the gender property) set to Q6581097 (the Q-item for male). This allows for a language independent format as the actual names of the entities and properties in particular languages are not used to reference them. L-items are data structures that store lexical elements of languages called lexemes (which are defined in the next section) [17]. Some examples of lexemes are the lexeme L42 being 'answer' and the lexeme L4041 being 'everything'.

In order for Abstract Wikipedia to generate articles in a given language, it must have the sufficient lexemes stored as L-items to do so but many African languages do not have a sufficient amount of L-items stored in Wikidata. For articles to be generated in African languages requires more L-items for these languages to be entered into Wikidata.

Currently, there are three major projects for Abstract Wikipedia under development. The first one is figuring out how a constructor (a novel data structure that will allow users to mix and match content for articles) would exactly work. The second one is template creation - templates are pre-constructed sentences that can be filled with relevant data to complete them (languages differ in grammar and sentence structure so different templates are needed per a language). The third one is collecting lexicographical data for all languages. Since African languages are low-resourced, there is a lack of lexicographical data available for them.

In order for African languages to be better represented on Wikipedia, lexicographical data collection needs to be done and this can be achieved through two methods. The first one is creating a lexicographic database where batch uploads can be done and this is the focus of my partner's literature review. The second one is creating a more usable interface that is Wikidata integrated and this is the focus of this literature review.

This literature review will focus on using crowdsourcing for collecting lexicographical data and will start with a discussion on background terminology. It will then go on to discuss different methods of constructing lexicons and why crowdsourcing has advantages for collaborative lexicography. It will then go over microtasks, gamification and crowd motivation in crowdsourcing. It will then look at various crowdsourcing projects for collecting lexicographical data and it will end with an analysis of the literature.

2 Using crowdsourcing for lexicographical data collection

Most interfaces for lexicographical data collection make use of an in-depth editing system that is more aimed at a language expert than a typical person [9] and this applies to the current Wikidata interface. Since low-resourced languages tend to have few language experts, more input is required from native speakers [9]. Crowdsourcing approaches that are quick, engaging and fun to use can encourage more native speakers to upload lexical data [9]. Crowdsourcing can be an effective method of collecting lexicographical data for African languages as it accounts for the barriers of low resourced languages (i.e. lack of experts as mentioned).

2.1 Background

This section contains relevant terminology for understanding lexicography. A lexeme is a 'unit of lexical meaning underlying a set of words' [17]. For example, run, runs, ran and running all refer to the same lexeme of run [17]. A lemma is the base of a lexeme - for the above example, run would be the lemma [17]. Forms are different ways of representing a lemma (the root word) [17]. For the above example, the forms would be runs, ran and running [17]. Senses are different meanings of a given lexeme [17]. For example 'king' can refer to a male monarch, a playing card or chess piece and all three variations would be senses [17].

2.2 Methods for constructing lexicons for languages

When building a lexicon for a language, there are different approaches that can be used such as expert-driven, corpus-based, crowdsourcing, machine learning and hybrid approaches. Expert-driven approaches involve language experts (such as lexicographers), curating a lexicon using texts, dictionaries and other sources [9]. This approach faces challenges such as the lack of experts available and limited funding for lexicographical projects [9]. This is especially true for low-resourced languages, which have fewer resources in terms of data, funding and experts compared to other languages [9]. This means they present a unique set of challenges in terms of collecting data to build lexicons [15].

Corpus-based approaches involve using large collections of text to build lexicons. However, copyright issues and other permissions present a challenge in terms of data collection for this approach [9]. For low-resourced languages, most large collections of text tend to have bad, inconsistent data [9], which makes using this approach difficult.

Machine learning approaches use natural language processing, supervised learning and other machine learning methods to build lexicons. However, these approaches require large amounts of data to be effective which can be a challenge for low-resourced languages as they do not have such large amounts of data available.

Crowdsourcing breaks the complex workload of collecting lexicographical data into smaller tasks that are distributed over a large number of people [12] and this approach presents many advantages over the above approaches. There are no copyright issues or experts required and the latter is a significant advantage for low-resourced languages as they have less experts compared to other languages as mentioned earlier. Crowdsourcing uses native speakers when developing lexicographical resources and this can prevent negative consequences, such as data not being validated [20]. For many low-resourced languages, Wikidata has had data quality issues due to bots entering false entries and data validation is needed to prevent this [17] and using native speakers for quality checks and assurance can counteract this [11].

Hybrid approaches combine two or more of the above approaches [9]. For example, crowdsourcing can be used to collect data initially and then experts will review everything in order to validate the data [9].

2.3 Crowdsourcing for collaborative lexicography

Modern lexicography utilises collaborative lexicography methods like crowdsourcing [13]. As described above, this method divides the large workload of lexicographical data collection into smaller tasks [13]. For example, instead of having one person enter all the forms of several lexemes, it can be split into individual tasks where several people enter one form each for one lexeme. Other examples would be having contributors identify parts of speech for specific words or provide senses for lexemes. The simplicity of these tasks make it easy for non-experts to complete them, reducing the need for experts that more traditional methods rely on [13].

The advantages of crowdsourcing include scale, speed, cost-effectiveness and diversity of input. Crowdsourcing allows for large-scale lexicographical data collection that traditional expert-driven methods cannot achieve as they require extensive resources in terms of expertise [8]. It can take decades to build a complete lexicon using experts only and there are too few language experts to document everything [6, 9]. Distributing the intense workload of building a lexicon among many contributors speeds up the process while also lowering the cost as lexicographical experts are more expensive than crowdsourcing [8]. This allows for data collection in crowdsourcing to be cost-effective and quick [8]. Another advantage is that it allows for diversity of input - experts might miss regional terms that native speakers will be able to contribute [8], but these regional terms might not be of value for the lexicon.

Another important consideration is that there are very few experts for low-resourced languages meaning often these languages do not receive any attention in terms of constructing lexicons for them [8]. Crowdsourcing gives such languages a chance at getting a lexicon they otherwise would not have [8].

However, crowdsourcing is not without problems with the most evident one being quality [9]. Crowdsourcing relies on the quality of contributors' contributions and sometimes they will enter incorrect data due to misreading or misunderstanding instructions, making spelling and grammar errors and adding malicious data on purpose [8, 9]. The incentives for inputting quality data input tend to be low even when there are monetary incentives as this often leads to maximising income with quantity of input being prioritised over quality input [8, 9]. Thus, it is important to validate data in order to counteract potential quality issues. Many crowdsourcing tools have various methods to do this such as gold standard, inter-annotator agreement, refereeing and intra-annotator agreement validation[13]. Gold standard validation checks data inputted by a user against test data from experts to check the reliability of a user and if the user is found to be unreliable, their input is not accepted [12]. Inter-annotator agreement is when different answers are given to the same question by different contributors and the most inputted answer is taken [12]. Refereeing [12] is when there's an ambiguous case (e.g. maybe inter-annotator agreement was used but there are two answers with five entries each) that an expert makes the final decision on. Intra-annotator agreement is when a contributor is given the same task throughout them using the interface to check for their consistency [12].

It is possible to have reliable and cost-effective data collection using crowdsourcing by validating the data. All tools described later on use some form of validation on their data. Sometimes crowdsourcing projects that validate their data can be both more accurate and quicker than other approaches [12] and can even rival expert methods [13]. Another minor problem to be aware of with crowdsourcing is dogmatic contributors who overwrite contributions due to feeling like an authority figure [8]. Crowdsourcing shouldn't be used in isolation, but in conjunction with other methods [16].

2.4 Microtasks and gamification for crowdsourcing

Microtasks are simple, short tasks utilised in crowdsourcing that are easy to understand and to complete [8, 12]. In crowdsourcing, a large and complex problem is split into smaller tasks [13, 20] and these tasks are called microtasks. Microtask design is important in order to engage the contributor base and keep them consistently contributing. Questions in a well designed microtask should be well-formulated, objective, one-dimensional and simple [13]. Other aspects to having well designed microtasks are having short instructions; no skill, knowledge or training requirements; providing feedback to crowdsourcers; having a user friendly interface; and offering entertainment through using challenge, randomness and time constrictions [13]. Microtasks should not be time consuming and should allow the user to use the interface in short bursts [9].

Gamification is a popular crowdsourcing approach. A popular type of gamification approach is game with a purpose (GWAP) [10] - where the main objective is entertainment for the users and the secondary objective is to collect data (in this case lexicographical data) [13]. Most tools described below are GWAPs and make use of various gamification elements. There are a few GWAPs that have been created for the Wikidata project to encourage more contributions [17] which are discussed later on.

2.5 Different types of motivation for crowdsourcing

Crowdsourcing heavily relies on contributors and maintaining their motivation to contribute is needed for the success of the overall project [12]. Contributors can be motivated through different ways such as psychological, social, educational and economic motivation. Psychological motivation is when contributors participate for altruistic or entertainment reasons [12]. As mentioned earlier, the gamification approach uses entertainment as the main motivating factor for contributors [10]. Social motivation is when contributors participate to interact with others, seek validation or climb leaderboards [9] and this type of motivation uses scoring systems, acknowledgements and titles as incentives.

Educational motivation is when a contributor participates to fulfil academic obligations [12]. Finally, economic motivation compensates contributors through micropayments [12]. As mentioned earlier, an issue with this type of motivation is that it can negatively impact quality of contributions as quantity is prioritised to maximise income [9].

Considering contributor motivation in the design of a crowdsourcing project is crucial, especially over the long term as their enthusiasm must be maintained for years so they keep contributing [9]. This is particularly true for low-resourced languages, as there are often a limited number of contributors and keeping this limited pool engaged is essential for the project's success [13].

It is also important to evaluate what is motivating participants as original hypotheses for crowd motivation from project design might not hold true. In a study about a crowdsourcing game for collecting isiXhosa lexicographical data - money was found to be a driving factor in participants' motivation as few were willing to participate without financial compensation which went against the original hypothesis for motivation being enjoyment [21]. However, in this study financial incentives alone were not sufficient to maintain participants' engagement in the long term as enjoyment was needed to keep participants engaged [21]. Also some game features thought to be motivational can have the opposite effect in some scenarios. In the above study, using a leaderboard drove participants away instead of motivating them [21], and the game 'Phrase Detectives' found timing features can put unwanted pressure on participants which causes them to disengage (this timing feature was removed for later versions of the game) [10].

2.6 Crowdsourcing projects for lexicographical data collection

There are many projects that collect lexicographical data that show promising results. Many of these projects make use of a GWAP format to engage the wider volunteer community to contribute and some projects focus on collecting data for Wikidata specifically.

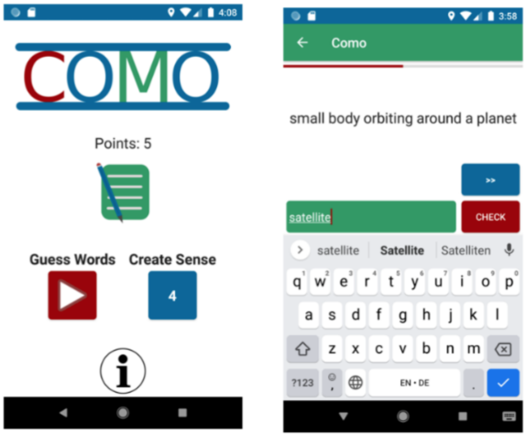
2.6.1 Features of GWAPs. The GWAPs described below have a few common features that will be touched on briefly here. For a few of the GWAPS, when a user starts contributing, gold standard validation is used to validate the user to see if their contributions are reliable. In all of the GWAPs, inter-annotator agreement is used to crosscheck answers from different contributers to validate the data. Most GWAPs use buttons (such as yes, no and skip buttons or similar variations) to speed up input [10] but a few GWAPs do require typed input.

The GWAPs below are mainly one player games but a two player format is also offered in two projects.

What differs the most between the GWAPs is the crowd motivation used - some focus on entertainment elements and user enjoyment as motivation while others rely on more altruistic reasons such as users knowing their contributions will improve Wikipedia.

Scoring system quality match another player [[10]]

2.6.2 GWAPs for Wikidata. Three GWAPS for Wikidata were found - the the Como app, the Distributed Game and Macht Sinn game. **Como** [17], as shown in figure 1, is an Android app that gamifies the collection of lexeme senses for Wikidata. It takes inspiration from crosswords in its gameplay design. Users are asked to input senses for lexemes and Como validates these user entries by having other users guess the lexeme from the inputted senses (allowing for inter-annotator agreement) [17]. Cross-checking user entries like this prevents incorrect data from being put on Wikidata [17]. Unlike other games described below, it does not make use of a basic yes/no question format as actual typed input is required from users [17].



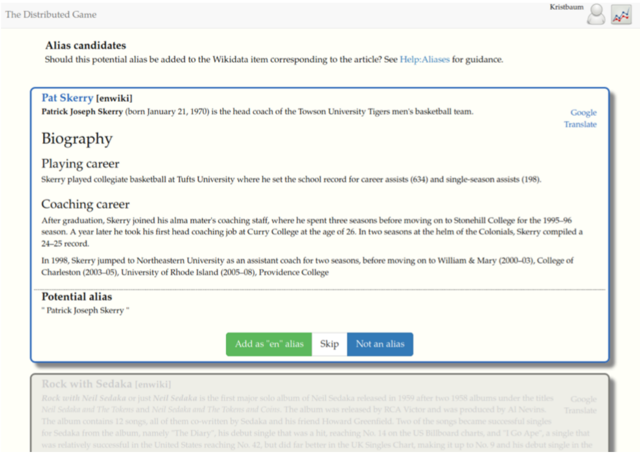
small body orbiting around a planet

satellite

Figure 1. Como App home page and guessing mode where users can match senses to lexemes [17]

A one month long study with English and German speakers found the app allowed for high quality contributions. Future work could include of the Como app being used to replace bad lexeme senses on Wikidata with better lexeme senses generated from the app and the app can be expanded to allow for entry of other types of data for lexemes such as antonyms, synonyms, word types and so on [17]. A consideration for the future is that currently there is no login required (for ease of use) but this increases the chance of malicious behaviour so a strategy should be implemented to deal with this [17].

The Distributed Game [5], as shown in figure 2, is a web platform that lets users play games to input data for Wikidata and it mainly focuses on suggested edits for Q-items. Although this data is not necessarily lexicographical data, it is worth mentioning due to its overall success as there have been over 1.3 million user interactions with the various games on it [17]. This platform allows for game creation meaning less experienced community members can get involved in the game creation process [17]. It is only playable in a browser and has three buttons for players: confirm, deny and skip [17]. This allows it to be executed very quickly compared to typing [17]. This platform does not utilise gamification elements like rewards or different tasks as it relies on motivating contributors on the fact that they are improving data statements for Wikidata [17].



Alias candidates

Should this potential alias be added to the Wikidata item corresponding to the article? See Help:Aliases for guidance.

Pat Skerry [enwiki]

Patrick Joseph Skerry (born January 21, 1970) is the head coach of the Towson University Tigers men's basketball team.

Biography

Playing career

Skerry played collegiate basketball at Tufts University where he set the school record for career assists (634) and single-season assists (198).

Coaching career

After graduation, Skerry joined his alma mater's coaching staff, where he spent three seasons before moving on to Stonehill College for the 1995–96 season. A year later he took his first head coaching job at Curry College at the age of 26. In two seasons at the helm of the Colonials, Skerry compiled a 24–25 record.

In 1998, Skerry jumped to Northeastern University as an assistant coach for two seasons, before moving on to William & Mary (2000–03), College of Charleston (2003–05), University of Rhode Island (2005–08), Providence College

Potential alias

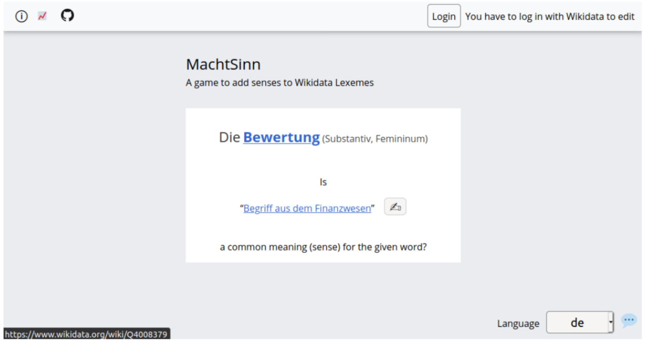
" Patrick Joseph Skerry "

Rock with Sedaka [enwiki]

Rock with Neil Sedaka or just *Neil Sedaka* is the first major solo album of Neil Sedaka released in 1959 after two 1958 albums under the titles *Neil Sedaka and The Tokens* and *Neil Sedaka and The Tokens and Coins*. The album was released by RCA Victor and was produced by Al Nevins. The album contains 12 songs, all of them co-written by Sedaka and his friend Howard Greenfield. Two of the songs became successful singles for Sedaka from the album, namely "The Diary", his debut single that was a hit, reaching No. 14 on the US Billboard charts, and "I Go Ape", a single that was relatively successful in the United States reaching No. 42, but did far better in the UK Singles Chart, making it up to No. 9 and his debut single in the

Figure 2. T
he Alias game on 'The Distributed Game' [5] that lets users add more aliases to entities [17].

Macht Sinn [4], as shown in figure 3, focuses on collecting lexicographical data, specifically lexemes and senses, for Wikidata [17]. It focuses on adding senses to lexemes. Unlike Como, Macht Sinn takes senses from Wikidata while Como creates senses from scratch [17]. It is similar to The Distributed Game as it gives users three options of yes, no and skip for them to respond to suggested edits [17].



Login
You have to log in with Wikidata to edit

MachtSinn

A game to add senses to Wikidata Lexemes

Die **Bewertung** (Substantiv, Femininum)

Is

"Begriff aus dem Finanzwesen"

a common meaning (sense) for the given word?

Language
de
https://www.wikidata.org/wiki/Q4008379

Figure 3. Macht Sinn Game [4] for adding senses for lexemes [17]

2.6.3 GWAPs for other lexicographical sources.

GWAPs for collecting data for other lexicographical resources were found and these GWAPs are the Kamusi Project, JeuxDeMots and Game of Words.

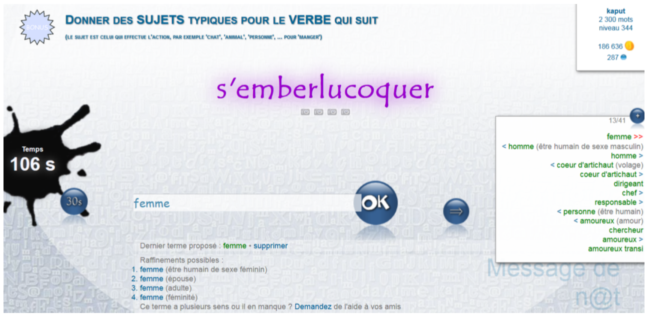
The **Kamusi project** [8, 9] is a large online dictionary project that aims to provide comprehensive dictionaries for all languages, especially low-resourced, African ones. The project has a game that can be played on Facebook, mobile and on the web [8]. It makes use of 'targeted microtasks' in a game environment and has a point system. It collects different types of data and uses well-thought-out questions to take into account the risks of user input [8]. An example would be not directly asking a user what a lexeme is in their language but rather asking if the senses for a given lexeme actually align (for example instead of asking 'what is a pen in your language', a user would be asked if the statement 'pen: a writing instrument' is correct or they could also be asked if 'pen: small enclosure for animals' is correct) [9]. Other types of questions are asked to get information on lexemes, lemmas and plurals.

The game validates users' inputs by having vote for definitions [8]. Users are given points for good definitions and also for voting for good definitions [9]. Before lexical data is added to the database, a threshold of a certain number of votes must be met (allowing for inter-agreement validation) [9]. This validation method of having multiple agreements leads to better quality results than a solo lexicographer doing the work and is also faster on top of being more accurate [13]. The game motivates users with a point system (which recognises their efforts), altruism and having fun [9].

The game can be configured based on the needs of a specific language when it comes to lexical input [9]. The project has some similarities to the Wikidata lexicographical games described earlier as it also collects high quality lexicographical data for a structured database.

JeuxDeMots [2], as shown in figure 4, is a GWAP for collecting French lexicographical data where two players play against each other [19]. It makes use of different game modes, in order to avoid monotony and keep players engaged (to reduce contributor turn over). Players can play games based on preferred themes (a player is more likely to be excited to play a game about themes they're interested in), difficulty levels, can retry games if needed and can chat to other players [19]. It also includes features such as rankings, gift parties and duels [19]. At both interface and content levels the game is easy to understand, attractive, fun and interesting in order to encourage addiction in the players so they keep contributing [19]. Players contribute collaboratively together as the games happen anonymously and asynchronously [19]. The anonymity helps with validation as it reduces the possibility of malicious behaviour since users do not know each other [19]. Some users who played the game a lot, started getting interested in the purpose aspect of the game (collecting French lexicographic data) and wanted to become direct contributors [19]. This led to a platform called Diko being developed where users can contribute lexical terms directly [19]. Another interesting aspect of JeuxDeMots, is its use of taboo words, words that have been entered many times in the game. When a word becomes taboo after reaching a certain threshold, it is not awarded any more points and this it to encourage other words to be inputted by users [18].

This helps collect and validate lexical data not in the dataset yet.



DONNER DES SUJETS TYPIQUES POUR LE VERBE QUI SUIT

(LE SUJET EST CELUI QUI EFFECTUE L'ACTION, PAR EXEMPLE 'CHAT', 'ANIMAL', 'PERSONNE', ... POUR 'MANGER')

kaput
2 300 mots
niveau 344
186 636
287

s'emberlucoquer

femme

13/41
femme >>
< homme (être humain de sexe masculin)
homme >
< coeur d'artichaut (volage)
coeur d'artichaut >
dirigeant
chef >
responsable >
< personne (être humain)
< amoureux (amour)
chercheur
amoureux >
amoureux transi

Dernier terme proposé : femme • supprimer

Raffinements possibles :
1. femme (être humain de sexe féminin)
2. femme (épouse)
3. femme (adulte)
4. femme (féminité)

Ce terme a plusieurs sens ou il en manque ? Demandez de l'aide à vos amis

Figure 4. A game play of JeuxDeMots [2] from this paper [19]

Game of Words is a mobile app designed for collecting Slovenian collocations (words that are commonly paired together e.g. heavy rain) and synonyms. The app offers different modes, scoring systems and both a single and two player mode to add variety to the game [6]. The creators of the app found that when creating such an app, the need for balancing lexicographical data collection intentions with user enjoyment and motivation for the game is an important task [6]. The creators also found that for such games the interface should be clear and easy-to-follow with content being the centre of attention (rather than being overshadowed by design elements such as colour and shapes) [6]. The results of the paper were not definite as more evaluations were needed to confirm them [6]. The app could be potentially integrated into the school curriculum for Slovenian to help students improve their language skills [6].

sloWCrowd is a tool for lexicographical data collection that was originally made to correct mistakes in databases but it was expanded to collect data [12]. This tool was used to help build a Slovene dictionary for Wordnet [12] and made use of simple yes/no and multiple choice questions to do so [14]. It obtains multiple answers to the same question, takes the answer with the majority vote and discards the other answers [14]. To check if a user is reliable, it makes use of gold standard validation where a user's answers are compared to expert data and if it does not match, that user is determined unreliable [12]. The tool also prioritises reliable users who give good reference answers by giving them more questions for data to validate (so they have more chances to contribute) compared to less reliable users who are given more questions from a set of reference questions [12].

2.6.4 Other notable GWAPs. Phrase Detectives is a GWAP that focuses on annotating corpora for English [13] and found a study on it found that in terms of validation, disagreement of user input against expert input does not mean either entered term is wrong but that the terms might be ambiguous [11]. A study on a crowdsourcing game for collecting isiXhosa lexicographical data was also found but the study [21] did not discuss elements of the game as the focus of it was evaluating if the game was effective or not.

2.6.5 Crowdsourcing platforms. Two notable crowdsourcing platforms are CrowdFlower and Amazon Mechanical Turk. They are different from GWAPs as they are platforms where crowdsourcers are paid to do microtasks (financial incentives are the main driver of motivation) [10]. A project done on CrowdFlower showed that annotating semantic roles in English texts in small steps with crowdsourcing was more accurate and faster than traditional annotation [13]. Amazon Mechanical Turk is a popular crowdsourcing website that makes use of micropayments to get a variety of tasks done [13]. It is dominated by American and Indian crowdsourcers as they make up 85% of crowdsourcers [21], so it is well suited for English and non-linguistic tasks. This platform along with CrowdFlower are not suitable for collecting lexicographical data for African languages due to the dominance of English speakers on them. It also has questionable ethics with regards to its micropayments system as there's a lack of legislation around payments to workers which can lead to workers not being fairly compensated [13].

2.6.6 Other similiar platforms to Wikidata. To the best of our knowledge, platforms similar to Wikidata that store lexicographical data - such as African Wordnet, Babelnet and Wordnet - have not been subject to research in terms of how they collect lexicographical data.

3 Analysis of literature

The above literature makes for a strong argument for using crowdsourcing and gamification for gathering lexicographical data as almost all projects were successful (only the isiXhosa one was not). The literature shows how gamification can be used for lexicographical data collection and what types of gamification elements work. The literature showed that it is possible to build accurate lexicons with crowdsourcing as long as the data is validated.

There are limitations in this literature that need to be taken into consideration. Although most papers discussed motivation, they did not differentiate between motivation of new users playing the game (i.e. what attracts them) and what keeps users playing the game over a long period of time. An expanded discussion on this would be useful along with how recruitment methods are used to attract users to see what works and does not work.

Two of the projects described above (the Como app [17] and the project to source isiXhosa lexicographical data [21]) were only studied over a short time period and not deployed into real world settings. This limits how useful their findings are compared to the other papers which have been implemented in real world settings. It would have been interesting to see what the results would have been like beyond the studies and how it would differ to what they found initially. The isiXhosa project [21], found that enjoyment and entertainment as primary motivation was not true for their project and stated that participants were only willing to contribute when there were financial incentives. The project conducted four experiments, with one not making use financial incentives while the other three did. The former experiment attracted very few contributors as the project was only shared to one of the author's Twitter pages while the latter experiments were marketed better. Their findings about financial and entertainment motivations might be because of how little the first experiment was marketed rather than the actual motivations of participants. This study also used university participants only which might skew the findings. The findings of the study might not be universally true due to these limitations discussed.

Some gaps in the literature are that there does not seem to be much literature on non-gamified interfaces for lexicographical data collection and literature specific to African crowdsourcing for lexicographic data that discusses contextual factors in-depth. For the former, this makes it difficult to see if a gamified interface is better and there are no papers on non-gamified interfaces to compare it to. For the latter most papers did not take user barriers into account like mobile access, education levels and bandwidth access (most papers used here are from more developed countries thus no need to do such in their contexts) and this would need to be taken into account as users speaking African languages would most likely be hindered by such barriers in terms of being able to contribute lexicographical data. The only paper that could have done this was the project for collecting isiXhosa lexicographical data [21] but it focused on a small selection of university students so it did not have to account for these factors.

No papers discussed how the morphology of languages affected data collection. Most African languages are Niger-Congo languages and these are agglutinative languages. Agglutinative languages have words that can have multiple prefixes and suffixes which can make data collection challenging. It would be interesting to see how the challenges agglutinative languages pose are taken into account when designing a crowdsourcing project for them.

The current literature could also give more comprehensive details around validation and crowdsourcing, as there are a few gaps in the literature around it. For validation methods for lexicography, although papers with projects could motivate why they chose their specific validation method, there were no explicit comparisons on which ones are the best (or which combination is the best to use or should all be used). It looks like it could be inter-annotator agreement (which was used in all of the projects) and then gold standard validation looking at how often used in the projects discussed.

Crowdsourcing seems to be portrayed as a panacea to lexicographical data collection and this could be the bias of the authors of the papers since they might feel invested in it as it is the method they used for data collection. More thorough criticism and challenges (beyond validity) around crowdsourcing could make the literature more comprehensive. The only paper that offered criticism was the isiXhosa one but as discussed earlier it had some limitations meaning its findings might have been specific for that paper.

A discussion around combining crowdsourcing with the other methods (especially machine learning and corpus-based methods) would also make the literature more thorough even if the discussion is on why the combination of methods would not work.

4 Conclusions

There are several successful crowdsourcing projects that collect lexicographical data for Wikidata and other lexicographical databases through a GWAP format. Two of these projects, the Kamusi project and the isiXhosa one, focus on African languages with the Kamusi one being relatively successful. These projects use well-designed microtasks that are simple and easy to answer to split the workload up of lexicographical data collection. These projects address the data quality issues of crowdsourcing by validating the data. These projects also make use of crowd motivation to encourage contributors to keep on contributing. The current gaps in literature should be accounted for but overall crowdsourcing is a promising approach for lexicographical data collection.

References

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

[2] 2023. JeuxDeMots : accueil. https://www.jeuxdemots.org/jdm-accueil.php

[3] 2023. List of Wikipedias. https://en.wikipedia.org/wiki/List_of_Wikipedias

[4] 2023. MachtSinn – Das macht doch alles keinen Sinn! https://machtsinn.toolforge.org/

[5] 2023. Wikidata - The Distributed Game. https://wikidata-game.toolforge.org/distributed/#mode=stats

[6] Špela Arhar Holdt, Nataša Logar, Eva Pori, and Iztok Kosem. 2020. "Game of Words": Play the Game, Clean the Database. In *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-9 September 2021, Alexandroupolis, Vol. 1*, Zoe Gavriilidou, Maria Mitsiaki, and Asimakis Fliatouras (Eds.). Democritus University of Thrace, Alexandroupolis, 41–49. https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol1-p041-049.pdf

[7] Peter K Austin and Julia Sallabank. 2011. *The Cambridge handbook of endangered languages.* Cambridge University Press.

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

[9] 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

[10] 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

[11] Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2009. Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*. 57–62.

[12] 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

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

[14] Darja Fišer, Aleš Tavčar, and Tomaž Erjavec. 2014. sloWCrowd: A crowdsourcing tool for lexicographic tasks. *Proceedings of LREC 2014* (2014), 4371–4375. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=bceafbfa0c9c64bab2af6d8d2d0dc5422458dc07

[15] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems* 30, 4 (2015), 81–85.

[16] Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Learning to Generate Wikipedia Summaries for Underserved Languages from Wikidata. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 640–645. https://doi.org/10.18653/v1/N18-2101

[17] 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

[18] Mathieu Lafourcade. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th international symposium on natural language processing*. Pattaya, Chonburi, Thailand, 7.

[19] Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2018. The JeuxDeMots Project is 10 Years Old: What We have Learned. In *Proceedings of the LREC Games4NLP workshop, May*. Miyazaki, Japan.

[20] 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

[21] 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

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