Wikidata for the People of Africa

Laurette Marais
Council for Scientific and Industrial Research, South Africa

Laurette Pretorius
Stellenbosch University, South Africa

Aarne Ranta
University of Gothenburg, Sweden

Krasimir Angelov
University of Gothenburg, Sweden

Abstract

We aim to employ natural language generation for expanding Wikidata entries with high quality labels and descriptions in the Bantu languages using Grammatical Framework (GF), with an initial focus on isiZulu and Siswati, and the geopolitical domain. Relying on the distinction GF makes between abstract and concrete syntax will ensure that the effort of expanding the solution to other Bantu languages is significantly reduced.

Introduction

Wikidata is an international, multilingual project. It describes itself as “a free and open knowledge base that can be read and edited by both humans and machines”. Reading and editing by humans is facilitated through the use of natural language labels and descriptions of items. It is therefore necessary for Wikidata to be a multilingual project so as to empower humans speaking any language to contribute and benefit from it. The Wikidata introduction expands on this by saying the following about multilingual support:

- “Editing, consuming, browsing, and reusing the data is fully multilingual”
- “Editing in any language is possible and encouraged.”

One obvious advantage of a multilingual approach is that it broadens the perspectives from which contributions are made.

However, support for Bantu languages lags far behind languages like English. The Bantu language family comprises between 440 and 680 languages in Sub-Saharan Africa and more than 350 million people speak one or more of the Bantu languages as mother tongues (30% of the population of Africa). These languages are mostly severely under-resourced and marginally represented in Wikipedia and Wikidata. The purpose of this project is to show how the presence of two such languages, viz. isiZulu and Siswati, can be extended in the Wikidata knowledge graph that supports other Wikimedia projects, including Wikipedia.

IsiZulu (12 million mother tongue speakers) and Siswati (2 million) are two of the eleven official languages of South Africa. Siswati is also the official language of the Kingdom of Eswatini. These languages belong to the Nguni group of languages and are linguistically closely related.

While our focus is on isiZulu, the largest Southern African language, and Siswati, a rather small one, the approach that we propose readily generalises to any other Bantu language for which a resource grammar exists. We return to this aspect in the section on Related Work.

Labels and descriptions, i.e. that aspect of Wikidata which makes the data accessible to

humans, for the Bantu languages are lacking in many items:
E.g.

- Wikidata contains English labels for 254 countries (Q6256 and relevant[1] subclasses), and descriptions for 252.
- In contrast, Wikidata contains isiZulu labels for 135 countries (Q6256 and relevant subclasses), and descriptions for only 6.

As with all Wikimedia projects, Wikidata aims to involve both contributors and users in its goal, which is "to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world" by collecting structured data in "a free, collaborative, multilingual, secondary knowledge base".

The contribution of labels and descriptions in a new language has the effect of empowering both those who wish to contribute to Wikidata as well as those who use Wikidata, whether directly or via other Wikimedia projects. This contributes to the dissemination of information to under-resourced communities.

However, the task naturally extends to communication in the other direction: the people of Africa should be enabled to add new data items and relations between such items, such as geographical places in their regions, which leads to the dissemination of African knowledge to the rest of the world.

For example, consider the case where an isiZulu speaker wants to contribute the fact that the King Cetshwayo District Municipality (Q311668) is named after (P138) Cetshwayo kaMpande (Q380403). Without isiZulu labels and descriptions for the relevant items and property, an isiZulu speaker does not have access to these items and properties and therefore cannot contribute this knowledge in isiZulu. As it happens, there is currently no isiZulu description for Q311668 and no label or description for P138.

Additions to the Wikidata knowledge graph can take the form of contributing new items and properties, but also of contributing triples that relate items to each other via properties. Our example is aimed at showing that the availability of labels and descriptions for existing items and relations directly enables the second kind of contribution. Contribution of new items and properties is a natural next step, and it should be possible to contribute new knowledge to Wikidata in one's own language.

Problem statement

Therefore, the aim of this project is to tackle the general lack of Bantu language labels and descriptions in Wikidata. We will use natural language generation (NLG) to address the problem in a scalable way.

Not only is the *severely resource-scarce* status of the Bantu languages a key reason for the lack of such labels and descriptions, but it also constrains how the problem should be approached. In resource-scarce contexts, efficient use of resources is critical, whether it be digital language resources, human resources or financial resources.

[1] The Wikidata knowledge graph contains some errors wrt subclasses of the item *country* (Q6256). For example, *Nigerian country* (Q55608482) has the English description "overview of country music in Nigeria", but is included as a subclass of *country* (Q6256) instead of *country music* (Q83440).

- Any linguistic data created in such a project must be of a *high quality*[2], since this data may form a significant part of the natural language content that is available for these languages. This is especially true in the case of Wikidata, where the formal nature of the data requires that it be accurately and precisely represented in natural language.
- By initially focusing on a single language and a specific domain, *applying insights gained* during the project to other Bantu languages is made more efficient. We will demonstrate the feasibility of expanding the solution to other Bantu languages by including Siswati.
- The Bantu languages exhibit substantial *linguistic similarities*. To exploit this effectively, any solution for a single language must be readily extensible to other Bantu languages.

In this project we focus on isiZulu labels and descriptions within the geopolitical domain, before demonstrating the effectiveness of bootstrapping to new languages by extending our solution to Siswati. This includes Wikidata items referenced in the description of countries. E.g. describing *Namibia* (Q1030) as "a country in Southern Africa" references items *country* (Q6256) and *Southern Africa* (Q27394), both of which lack isiZulu labels. While developing the description of *Namibia*, new labels for *country* and *Southern Africa* will also be added. Table 1 gives a summary of what is currently available in English and what we hope to contribute for isiZulu and, subsequently, Siswati.

[2] Here it will be important to engage with the natural custodians of the languages, e.g. the Pan South African Language Board (https://www.pansalb.org/) and the National Lexicography Units (https://sanlu.africa/) of South Africa.

Research questions

Our main research question is: How can Grammatical Framework (GF) be used to address the lack of Bantu language labels and descriptions in Wikidata?

The research sub-questions are:

1. What terminology must be collected/developed to describe items in the geopolitical domain in Wikidata?
2. How can a GF grammar be used to model descriptions of countries in isiZulu so that it is readily extensible to other Bantu languages, with Siswati as an example? A related question is to what extent the cross-lingual API of the GF common abstract syntax is useful for the Bantu languages.
3. How can data extracted from Wikidata be used to generate GF abstract syntax trees that correspond to correct isiZulu and Siswati descriptions of countries?
4. How can a workflow be designed that minimises the effort that would be required to expand the solution to new languages and domains? A corollary is: What are the concrete steps required to expand the solution to new languages and domains?

We sketch the envisioned solution: Figures 1 and 2 show the language independent information available about Botswana in Wikidata. Figure 1 shows what is seen on the Wikidata webpage for Botswana, while Figure 2 represents pertinent structured data about Botswana as RDF triples. This knowledge is accessible to speakers of English, for example, via the English labels and descriptions indicating what is represented by items such as Q963 and Q123480 and properties such as P31 and P361.

A description of an item is "a short phrase designed to disambiguate items with the same

or similar labels", which can be done effectively by stating key facts about the item. Indeed, such facts are exactly what is already encoded in the Wikidata knowledge graph. The fact that Botswana is an "instance of" a landlocked country and is "part of" Southern Africa is exactly the kind of information that can be used to generate a good, unambiguous description of the item. If a mechanism can be developed for combining such structured data into natural language strings, such descriptions could be generated automatically.

GF is ideally suited to this task. Figure 3 depicts a GF abstract syntax tree expressing a useful description based on this information using typical Bantu language constructions. The construction of this abstract syntax tree from structured data can be done, either directly or via a GF application grammar, within a program that interacts with the GF C runtime. Figure 4 shows the tree linearised as an accurate isiZulu description of Botswana using the isiZulu resource grammar. The Siswati resource grammar would be used to linearise the same abstract syntax tree into Siswati.

By answering the stated research questions, we will expand the ability of Wikidata to be read and edited by humans, especially those in Sub-Saharan Africa, with its many Bantu languages. This will benefit the multilingual support of Wikidata to other Wikimedia projects.

We also intend to showcase the fact that the project enables the contribution of new knowledge to the Wikidata knowledge graph, by identifying appropriate data which is not yet in the Wikidata knowledge graph and for which isiZulu and Siswati labels and descriptions can be generated via a GF grammar. New items or new properties of existing items will then be contributed to Wikidata with isiZulu and Siswati labels and descriptions. We particularly hope to find data relating to geopolitical realities in Southern Africa that we can contribute, although we will consider other domains if necessary.

Date: June 1, 2024 - June 30, 2025.

Related work

The gf-wikidata[3] project aims to generate entire Wikipedia articles from Wikidata. However, the current prototype has been tested for European languages that are resource richer.

GF resource grammars (RGs) are key to linearising abstract syntax trees into natural language. An RG exists for isiZulu, and has been used to develop isiZulu language resources (Marais & Pretorius 2023a, 2023b). In a current project[4], an RG for Siswati has recently been completed via bootstrapping, with one for isiXhosa planned to be completed within 18 months. RGs for other Bantu languages are also in development (Kituku et al. 2021, Bamutura et al. 2020).

We intend to use a GF application grammar in this project. The next section includes examples of the use of GF application grammars in previous work.

Methods

Grammatical Framework (GF) is a computational grammar framework for the development of multilingual grammars and may be considered "the de-facto open source general framework for developing resources for engineering multilingual CNLs" (Safwat and Davis, 2017).

[3] https://github.com/krangelov/gf-wikidata
[4] https://github.com/LauretteM/gf-bantu-resources

GF grammars are characterised by an interlingua architecture, where an abstract syntax models a domain of utterances in a language independent way, and a set of concrete syntaxes defines how the utterances are expressed in different natural languages.

Besides distinguishing between *abstract* and *concrete* syntaxes, another important distinction is between GF *resource* grammars and GF *application* grammars.

GF resource grammars define syntactic categories and functions and serve as a linguistic software library for application grammar development. Application grammars define semantic categories and functions that model domains of application, such as utterances relating to geopolitical concepts.

Items	**# English**	**# isiZulu**	**# Siswati**	**Estimated minimum contribution**
All countries and relevant subclasses - labels	254	135	172	119 isiZulu labels 82 Siswati labels
All countries and relevant subclasses - labels and descriptions	252	6	0	246 isiZulu descriptions 252 Siswati descriptions
Capital cities of countries and relevant subclasses - labels	248	57	15	191 isiZulu labels 233 Siswati labels
Capital cities of countries and relevant subclasses - labels and descriptions	248	0	0	248 isiZulu descriptions 248 Siswati descriptions
Languages with more than a million speakers - labels	350	49	14	301 isiZulu labels 336 Siswati labels
Languages with more than a million speakers - labels and descriptions	350	6	1	344 isiZulu descriptions 349 Siswati descriptions
Currencies of countries and relevant subclasses - labels	167	4	1	163 isiZulu labels 166 Siswati labels
Currencies of countries and relevant subclasses - labels and descriptions	166	0	0	167 isiZulu descriptions 167 Siswati descriptions
Total labels				**774 isiZulu labels** **817 Siswati labels**
Total descriptions				**1005 isiZulu descriptions** **1016 Siswati descriptions**

Table 1: Estimated contribution based on existing labels and descriptions.

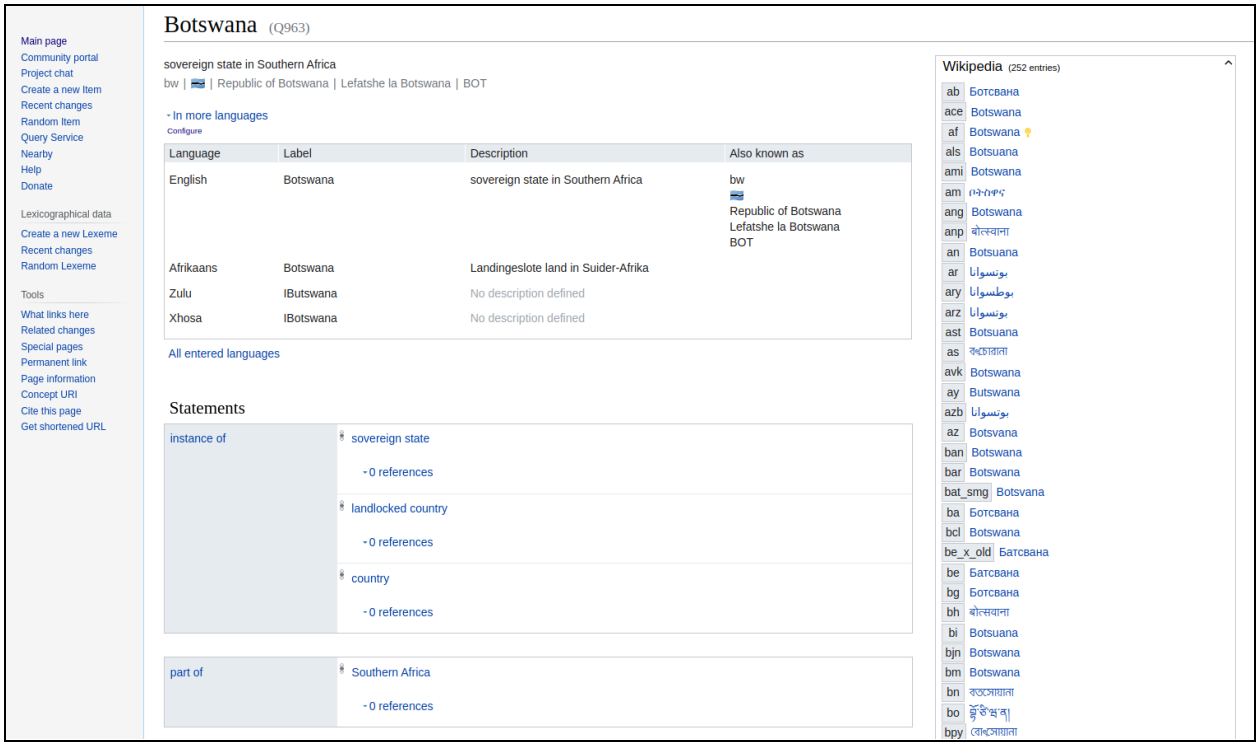

Figure 1: Wikidata page for Botswana (Q963): it is an *instance of* a **landlocked country (Q123480)** which is *part of* **Southern Africa (Q27394)**

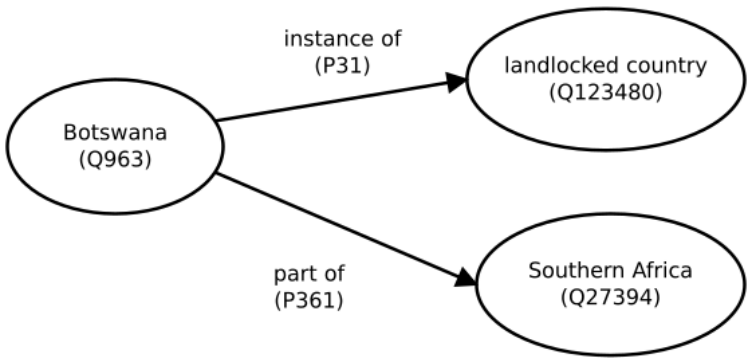

Figure 2: RDF triples from the Wikidata knowledge graph that can be utilised in a description of the item for Botswana (Q963).

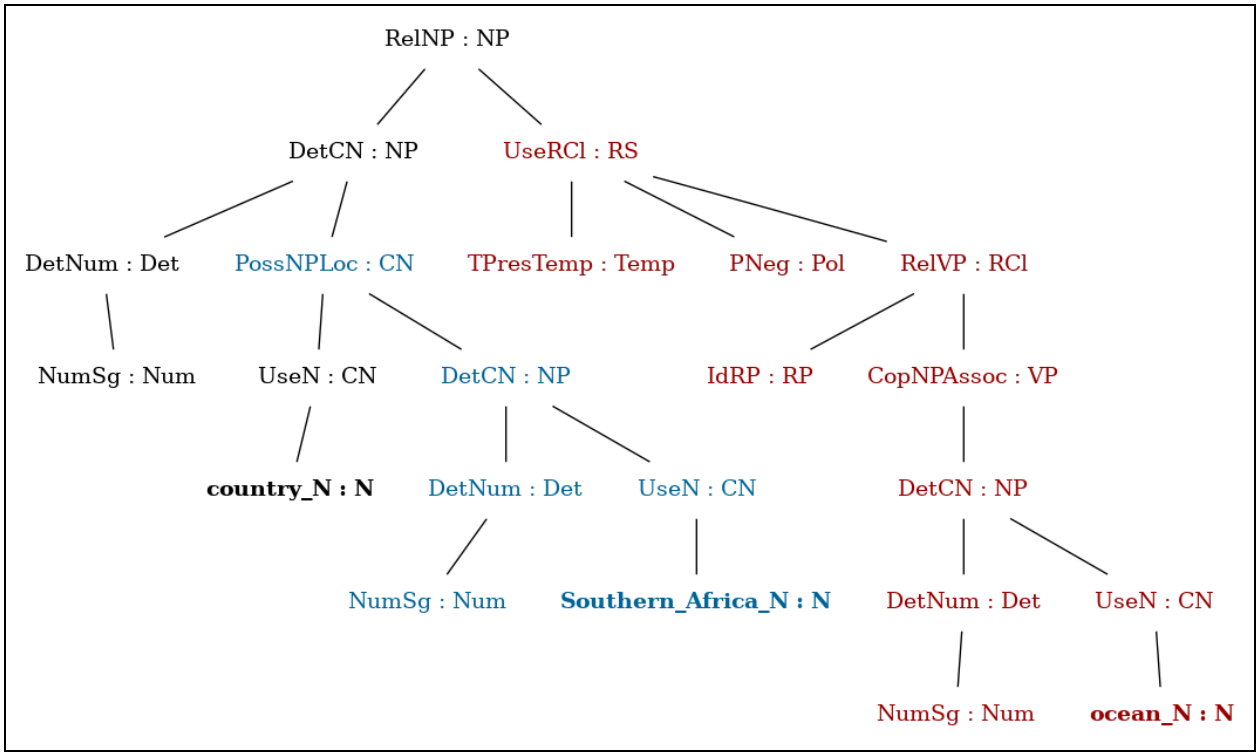

Figure 3: A GF abstract syntax tree expressing "a **country** of Southern Africa **which is not with the ocean**" using functions common to many Bantu languages, which can be derived from the knowledge graph.

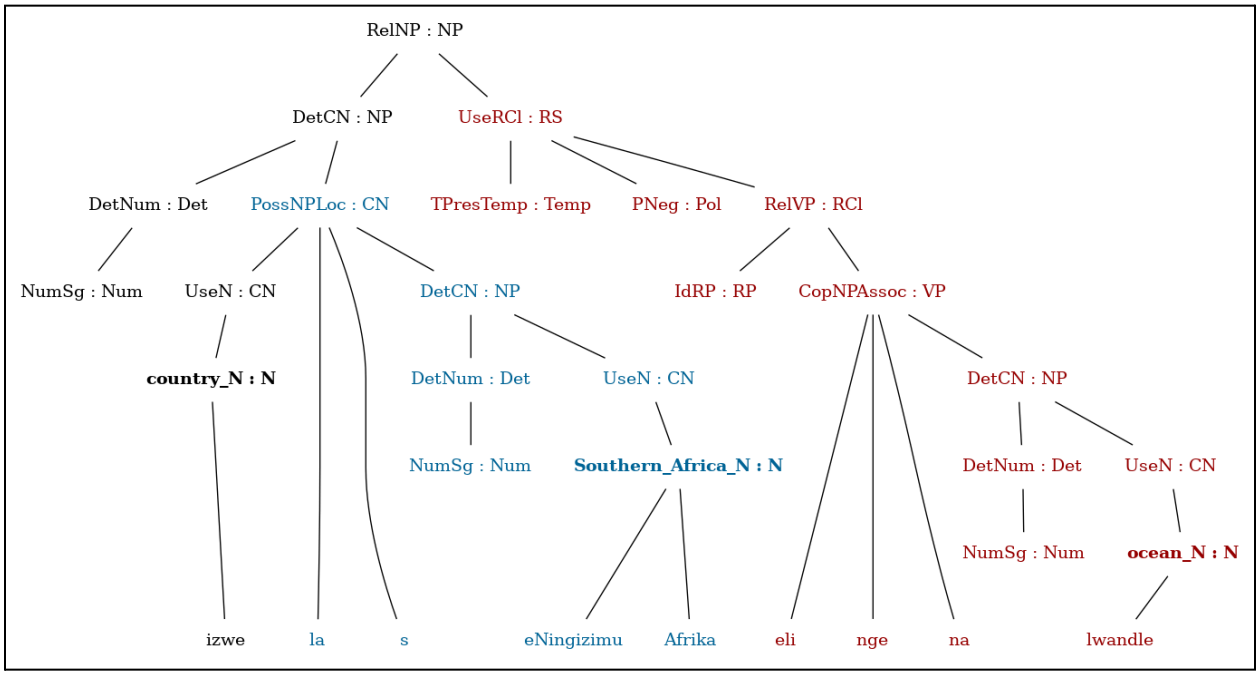

Figure 4: The GF tree linearised as isiZulu text: **izwe laseNingizimu Afrika elingenalwandle**

GF application grammars have been used to model utterances in a large number of diverse multilingual domains, including mathematics (Saludes 2011), transport (Bringert et al 2005), weather reports (Lobanov 2017), healthcare (Marais et al 2020, Ranta et al 2017), literacy development and language learning (Marais et al 2023c), technical texts describing properties of places and objects related to accessibility by disabled people (Ranta et al 2015), etc. Verbalisation of structured data, in particular, has been done for biomedical linked data (Marginean 2017), descriptions of museum objects (Dannélls 2011) and modular ontologies (Davis et al 2012).

The description of items that will be generated in this project is a form of ontology verbalisation, since the descriptions we intend to contribute to Wikidata will, in effect, be multilingual verbalisations of knowledge already present in the Wikidata knowledge graph. The example in Figures 1 to 4 shows how two triples from the Wikidata knowledge graph are combined via GF to produce an isiZulu utterance that serves as an apt description of the item. The distinction in GF between abstract and concrete syntax is the hinge for this kind of multilingual natural language generation, making GF ideally suited for this project.

Methodology

The main steps of the methodology are as follows:

1. **Terminology:** Collect and employ existing terminology resources and develop new terminology in consultation with expert isiZulu linguists/lexicographers (See footnote 1). We will use a data language such as YAML to represent and maintain a multilingual lexicon of the relevant terminology. This will be automatically convertible to GF lexicon modules. For this part of the project, we will engage with developers of terminology, as well as PanSALB, via SADiLaR (see Appendix A).
2. **GF application grammar:** Follow a standard methodology for domain-specific GF grammar development. A full description of the process is outside the scope of this proposal. It has been described in Ranta (2011) as well as various other scientific publications, but typically follows these steps:
    - Elicit a representative sample of text from the domain
    - Analyze the sample to design a model of the domain in the form of a GF application grammar
    - Create a GF lexicon as required by the domain
3. **NLG component:** develop a system that queries Wikidata and constructs suitable GF trees. This system will be written in Python, and will interact with the compiled GF grammar via the *pgf*[5] package, which provides Python bindings to the GF C runtime.
4. **Verification:** expert linguists will be consulted to evaluate the accuracy of the generated descriptions.
5. **Wikidata contribution:** we will contribute the newly developed labels and descriptions using the Wikidata REST API.

Expected output

The expected outputs of the project are listed below.

[5] See package details at: https://pypi.org/project/pgf/

1. isiZulu labels and descriptions for the countries in Wikidata. The audience is isiZulu contributors and users of Wikidata. Indeed, we hope to encourage the growth of this audience via this project. This will be released under a CC0 license.
2. Open-source baseline GF-based NLG system for isiZulu and Siswati descriptions. The audience is researchers interested in extending/adapting the system. This will be released under LGPL.
3. Scientific publication detailing process and findings. The audience is the scientific community interested in Wikidata and Bantu languages.
4. Project report discussing the above outputs, research findings and potential future work. The intended audience is the Research Fund chairs of the Wikimedia Foundation Research Fund 2024.

Seen together, these outputs represent a blueprint for how this kind of work could continue and be expanded in future. We regard this blueprint in itself as an essential and significant contribution to the expansion of the digital presence of the Bantu languages within the Wikimedia projects and beyond.

Risks

1. The greatest risk is the resource-scarceness of isiZulu and Siswati in that the appropriate terminology may be nonexistent or unavailable. Funding may have to be reallocated from the project team towards terminology development.
2. Another risk is the possibility of existing errors in Wikidata. Depending on the nature of the errors, mitigation strategies may vary.

Community impact plan

The community impact of this project can be summarised in the following points:

- Expansion of the human readable isiZulu and Siswati content in Wikidata, a project aimed at a wide audience. The expansion of the human readable isiZulu and Siswati content in Wikidata will not only facilitate contributions from speakers of these languages, but will also serve to make the content available in these languages for users of Wikidata as well as those projects that rely on Wikidata.
- Engagement with terminology stakeholders, including PanSALB[6], the South African NLUs[7], USAf[8], SADiLaR[9] etc. This will be essential to ensure that the labels and descriptions contributed in this project are of a high quality and deemed acceptable by key members of the speaker communities of the languages.
- We will be liaising with the SADiLaR-Wikipedia-PanSALB[10] project, as mentioned in the letter of support provided as Appendix A. In keeping with the ethos of Wikimedia, we do not consider the contributions of this project to be final, but rather to be useable and of a high quality. It will provide a basis upon which speakers of isiZulu and Siswati, and perhaps other related languages, can build via

[6] Pan South Africa Language Board (https://www.pansalb.org/)
[7] National Lexicography Units (https://sanlu.africa/)
[8] Universities South Africa (https://usaf.ac.za/)
[9] South African Centre for Digital Language Resources (https://sadilar.org/)
[10] https://sadilar.org/index.php/en/2-general/416-swip

improvements and additions. In their letter, SADiLaR indicates that our contributions could be used in workshops in which native speakers are trained to use and contribute to Wikidata.

Evaluation

Our evaluation will focus on *accuracy* and *coverage*.

Terminology

Terminology collected/developed and contributed to Wikidata as isiZulu labels. If existing terminology can be identified, there would be no need to evaluate its accuracy. If new terminology must be developed, it will not only be important to engage with appropriate lexicographers, but to determine a suitable evaluation strategy, including identifying the appropriate stakeholders to engage in this process.

GF application grammar/descriptions contributed to Wikidata

The evaluation of a GF application grammar typically involves generation of an appropriate sample of utterances from the grammar, and presenting them to an expert. In our case, we will generate a sample of true descriptions based on the Wikidata knowledge graph. Both linguists and lexicographers will be approached in the evaluation at this point.

This evaluation will be included in the project report.

Budget

Link to budget.

Response to reviewers and meta-reviewers

We structure our responses along three themes that emerged from the reviews.

Suitability of GF

We have so far expanded on GF as a formalism and platform for developing multilingual grammars in order to clarify its suitability to this project, specifically in the Introduction and Methods sections. GF as a choice of formalism does not pose a risk to this project. Here we want to address its suitability in comparison to the other suggested approach, namely the use of statistical or another kind of data driven machine translation.

The heart of this project is the combination of terminology and a GF multilingual grammar to reliably produce labels and descriptions for items in the Wikidata knowledge graph. The spirit in which this work is undertaken is to adopt the Bantu languages as our point of departure in the development of the labels and descriptions. In our view, a statistical or data-driven machine translation approach has significant drawbacks.

Most importantly, the labels and descriptions produced via translation will of necessity be derivative of English, instead of having been developed as isiZulu-first (or Bantu-language-first) content. The significant linguistic and lexical differences between English and the Bantu languages could lead to translations that are easily identified as such as opposed to content that was developed from a Bantu-language perspective. This is especially true if there is a lack of suitable domain data to train good models, which is certainly the case here. We view the contribution of English-derivative descriptions, when it is possible to generate idiomatic, domain

appropriate descriptions directly in the Bantu languages, as inappropriate.

Moreover, machine translation from English to the Bantu languages, not to mention cross-lingual transfer learning between Bantu languages, is still an active field of research, where reported BLEU scores (see eg Reid 2021 and Ngomane 2023) do not justify this approach for such a specialised domain for which no suitable training data is available. This is especially true in view of GF as an alternative, where linguistic accuracy and idiomatic, domain appropriate natural language generation based on structured data can be relied on, and where the path to expanding to additional Bantu languages is comparatively clear (Bosch et al 2008, Kituku 2021).

Native speaker involvement

Native speaker involvement in this project will not be direct, but engagement will certainly be required in the collection and/or development of suitable terminology. Moreover, native speakers will be approached in the evaluation phase to establish the acceptability of the generated outputs.

Furthermore, we are grateful for the support of SADiLaR for this project, given the promising collaboration opportunities with projects like SWiP, where direct community engagement is the focus. See Appendix A.

Scope and impact

We have significantly increased the scope of this project to demonstrate the feasibility of extending the solution to an additional Bantu language, and one, moreover, that is significantly more resource-scarce (Moors 2018) than isiZulu, namely Siswati. This is possible due to the successful completion of a bootstrapping effort from the isiZulu resource grammar to a new Siswati resource grammar.

We have also included detail about the number of labels and descriptions that will be targeted in Table 1.

A suggestion to translate the interface was made. We agree that this would be a good idea, but it falls outside the scope of this project.

This project has been specifically conceptualised to advance the stated aims of the Wikidata project, namely to be "a free, collaborative, *multilingual*, secondary knowledge base, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world."[11]

References

Bamutura, D., Ljunglöf, P., & Nebende, P. (2020). Towards Computational Resource Grammars for Runyankore and Rukiga. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2846-2854).

Bosch, S., Pretorius, L., & Fleisch, A. (2008). Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies*, 17(2), 23-23.

Bringert, B., Cooper, R., Ljunglöf, P., & Ranta, A. (2005). Multimodal dialogue system grammars. In *Proceedings of DIALOR'05, ninth workshop on the semantics and pragmatics of dialogue* (pp. 53-60).

Dannélls, D., Damova, M., Enache, R., & Chechev, M. (2011). A framework for improved access to museum databases in the semantic web. In *Proceedings of the Workshop on Language*

[11] See the Wikidata introduction at: https://www.wikidata.org/wiki/Wikidata:Introduction

Technologies for Digital Humanities and Cultural Heritage (pp. 3-10).

Davis, B., Enache, R., Van Grondelle, J., & Pretorius, L. (2012). Multilingual verbalisation of modular ontologies using GF and lemon. In *Controlled Natural Language: Third International Workshop, CNL 2012, Zurich, Switzerland, August 29-31, 2012. Proceedings 3* (pp. 167-184). Springer Berlin Heidelberg.

Kituku, B., Nganga, W., & Muchemi, L. (2021, November). Leveraging on cross linguistic similarities to reduce grammar development effort for the under-resourced languages: a case of Kenyan Bantu languages. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)* (pp. 83-88). IEEE.

Kotzé, G., Pretorius, L. (2020). Die Afrikaanse Wikipedia en hoe om by te dra: Handleiding. Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns, Pretoria, 69 pages.

Lobanov, G. (2017). *Grammatical Framework For Multilingual Natural Language Generation: The Weather Report Case*. [Unpublished master's thesis], Chalmers University of Technology and the University of Gothenburg.

Marais, L., Louw, J. A., Badenhorst, J., Calteaux, K., Wilken, I., Van Niekerk, N., & Stein, G. (2020). AwezaMed: A multilingual, multimodal speech-to-speech translation application for maternal health care. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)* (pp. 1-8). IEEE.

Marais, L., Pretorius, L. (2023a). Extending the usage of adjectives in the Zulu AfWN. In *Proceedings of the 12th Global Wordnet Conference*, pages 303–314, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Marais, L., & Pretorius, L. (2023b). Parsing IsiZulu text using Grammatical Framework. In *International Symposium on Distributed Computing and Artificial Intelligence* (pp. 167-177). Cham: Springer Nature Switzerland.

Marais, L., Wilken, I., Pretorius, L., & Posthumus, L. C. (2023c). Multimodal, multilingual dynamic stories for literacy development and language learning. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1-5).

Marginean, A. (2017). Question answering over biomedical linked data with Grammatical Framework. *Semantic Web*, 8(4), 565-580.

Moors, C., Wilken, I., Calteaux, K., & Gumede, T. (2018). Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists* (pp. 296-304).

Ngomane, D., Mabuya, R., Abbott, J., & Marivate, V. (2023, May). Unsupervised Cross-lingual Word Embedding Representation for English-isiZulu. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)* (pp. 11-17).

Pretorius, L. (2016). Die rol van die Afrikaanse Wikipedia in die uitbou van Afrikaans. *Tydskrif vir Geesteswetenskappe*, 56(2-1),371-390.

Pretorius, L. (2015). Workshop on the role of Wikipedia in the intellectualisation of the African Languages: Why, what and how?, *invited speaker Mr Amir Aharoni (Wikimedia Foundation),* hosted by the Academy of African Languages and Science, College of Graduate Studies, University of South Africa, Pretoria. 3 June.

Pretorius, L. (2015). Die Afrikaanse Wikipedia, Interview with National Radio RSG on the programme "Die tale wat ons praat" on 1 November.

Pretorius, L. (2016). Die Afrikaanse Wikipedia. Keynote address, Universiteit van Pretoria, Spring Seminar, Pretoria. 16 September.

Pretorius, L. (2016). Workshop on building the Afrikaans Wikipedia. Bloemfontein, University of the Free State. 29 September.

Pretorius, L. (2017). The importance of the Afrikaans Wikipedia for Digital Language Vitality. Waverley Leeskring, Pretoria. 7 Januarie.

Pretorius, L. (2017). Die Afrikaanse Wikipedia. Werkswinkel, Vrystaat Kunstefees, Bloemfontein, 20 Julie.

Pretorius, L. (2018). Hoekom is die Afrikaanse Wikipedia belangrik vir die toekoms van Afrikaans? Wikwinkel Leeskring, 12 June.

Pretorius, L. (2020). Die Afrikaanse Wikipedia. Interview with National Radio Station RSG on the programme "Taaldinge", 26 January.

Pretorius, L. (2020). Hoe brei ons die Afrikaanse Wikipedia uit? Workshop on the Afrikaans Wikipedia, Suid-Afrikaanse Akademie vir Wetenskap en Kuns (SAAWK), Pretoria. 30 January.

Pretorius, L. (2021). Hoekom is die Afrikaanse Wikipedia belangrik vir die toekoms van Afrikaans? Wikipedia 20 jaar, Suid-Afrikaanse Akademie vir Wetenskap en Kuns, Pretoria. 16 November.

Pretorius, L. (2023). Facilitator, Panel Discussion on Preserving Languages & Scientific Information: Accessible Knowledge for All. SWiP Event (SADiLaR-Wikipedia-PanSALB), CSIR, Pretoria, South Africa, 6 December.

Pretorius, L., Wolff, F. (2020). Wikipedia as a transformative multilingual knowledge resource. In book *The transformative Power of Language: From Postcolonial to Knowledge Societies in Africa*, Russell H Kaschula & H Ekkehard Wolff (eds), Cambridge University Press, September.

Ranta, A. (2011). *Grammatical framework: Programming with multilingual grammars* (Vol. 173). Stanford: CSLI Publications, Center for the Study of Language and Information.

Ranta, A. (2023). Multilingual Text Generation for Abstract Wikipedia in Grammatical Framework: Prospects and Challenges. *Logic and Algorithms in Computational Linguistics 2021* (LACompLing2021), 125-149.

Ranta, A., Unger, C., & Hussey, D. V. (2015). Grammar engineering for a customer: A case study with five languages. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 workshop* (pp. 1-8).

Ranta, A., Angelov, K., Höglind, R., Axelsson, C., & Sandsjö, L. (2017). A mobile language interpreter app for prehospital/emergency care. In *Medicinteknikdagarna*, Västerås Sweden, October 10-11, 2017.

Reid, M., Hu, J., Neubig, G., & Matsuo, Y. (2021). AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1306-1320).

Safwat, H., & Davis, B. (2017). CNLs for the semantic web: a state of the art. *Language Resources and Evaluation*, 51(1), 191-220.

Saludes, Jordi, and Sebastian Xambó. (2011). The GF Mathematics Library. CTP *Components for Educational Software,* 46.

Appendix A

South African Centre for Digital Language Resources
Private Bag X1290, Potchefstroom
South Africa 2520

Tel: +2718 285 2750
Email: info@sadilar.org
Web: www.sadilar.org

RE: SUPPORT TOWARDS WIKIDATA FOR THE PEOPLE OF AFRICA PROJECT

To whom it may concern,

The Wikidata for the People of Africa (W4PA) project aligns well with our vision to ensure a digital future for all the official languages in South Africa. It also links well with current projects, in particular our SADiLaR-Wikipedia-PanSALB(SWiP) - Project, which is bringing together communities of indigenous language users to gain experience on how to create and review content on Wikipedia, thereby promoting the online presence of all South Africa's indigenous languages.

The South African Centre for Digital Language Resources (SADiLaR) as national research infrastructure pledges to support the W4PA project by linking its generated outputs with community members who can use these outputs in African Languages, isiZulu and Siswati in particular, as material to refine during SWiP workshops. Furthermore, SADiLaR can facilitate contact between W4PA and key role stakeholders such as the Pan South African Language Board through our existing network.

SADiLaR also welcomes a partnership towards the continuation of SWiP training workshops.

Kind regards

Juan Steyn
Operations Director

Nomsa Skosana
Lebogang Boemo
SWiP Project Leaders