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 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. However, support for Bantu languages lags far behind languages like English. Labels and descriptions, which make the data accessible to humans, are lacking in many items:
E.g.

- Wikidata contains English labels for 269 countries (Q6256 and subclasses), and descriptions for 206.
- In contrast, Wikidata contains isiZulu labels for 135 countries (Q6256 and subclasses), and descriptions for only 6.

The problem we aim to solve is the general lack of Bantu language labels and descriptions in Wikidata via natural language generation (NLG). The *severely resource-scarce* status of the Bantu languages constrains how the problem should be approached, since efficient use of resources is critical.

- Any linguistic data created in such a project must be of a *high quality*[1], since this data may form a significant part of what is available for these languages.
- 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.
- 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. 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.

Main research question: How can Grammatical Framework (GF) be used to address the lack of

[1] 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.

Bantu language labels and descriptions in Wikidata?

Research sub-questions:

1. What terminology must be collected/developed to describe countries (Q6256 and subclasses) 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? A related question is to what extent the cross-lingual API of the GF common abstract syntax is useful for isiZulu, a Bantu language.
3. How can data extracted from Wikidata be used to generate GF abstract syntax trees that correspond to correct isiZulu descriptions of countries?

We sketch the envisioned solution: Figure 1 shows the language independent information available about Botswana in Wikidata, and Figure 2 depicts a GF syntax tree expressing a useful description based on this information using typical Bantu language constructions. Figure 3 shows the tree linearised as an accurate isiZulu description of Botswana.

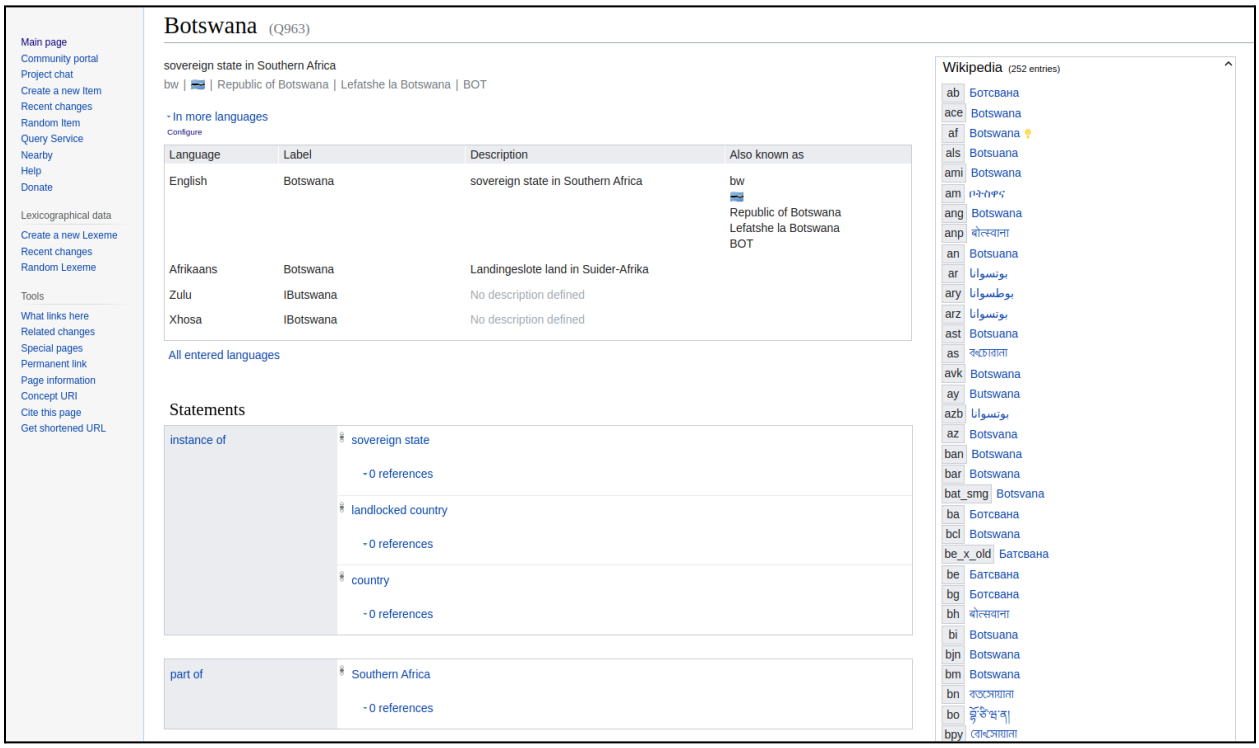

Figure 1: Wikidata page for Botswana (Q963): it is an *instance of* a **landlocked country (Q123480)** which is *part of* **Southern Africa (Q27394)**

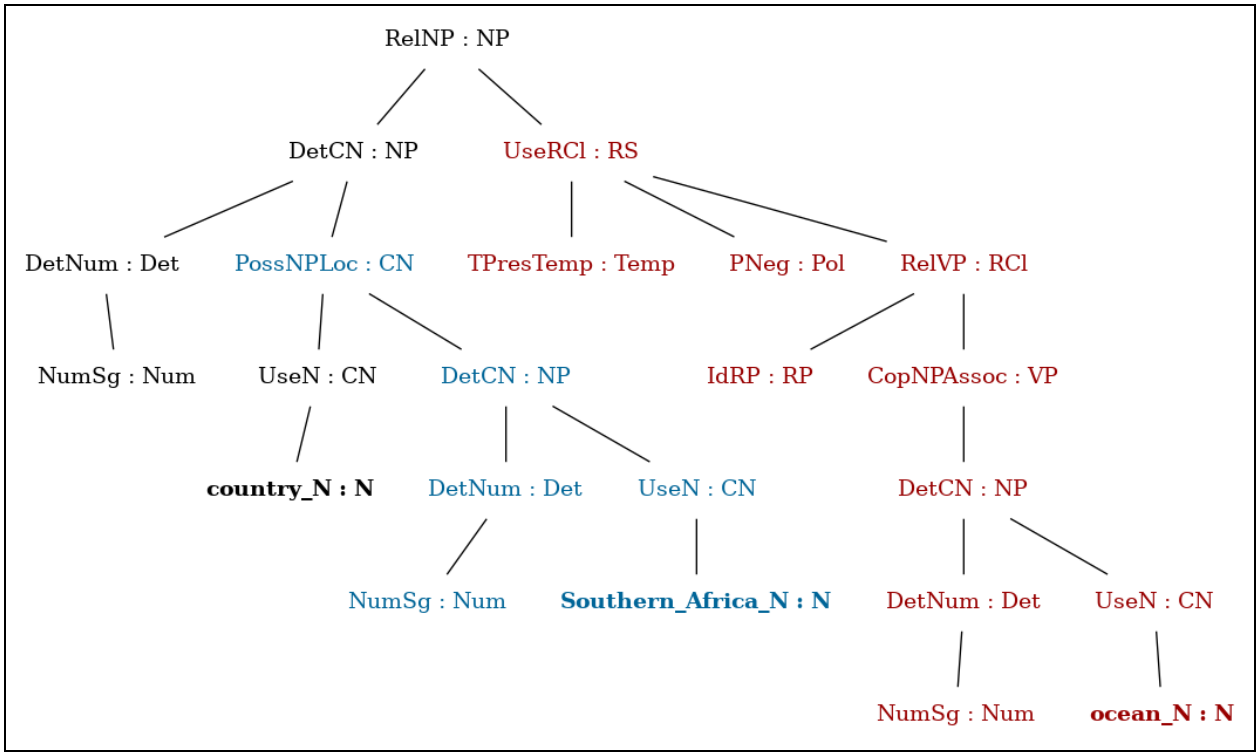

Figure 2: A GF abstract syntax tree expressing "a **country** of Southern Africa **which is not with the ocean**" using functions common to many Bantu languages.

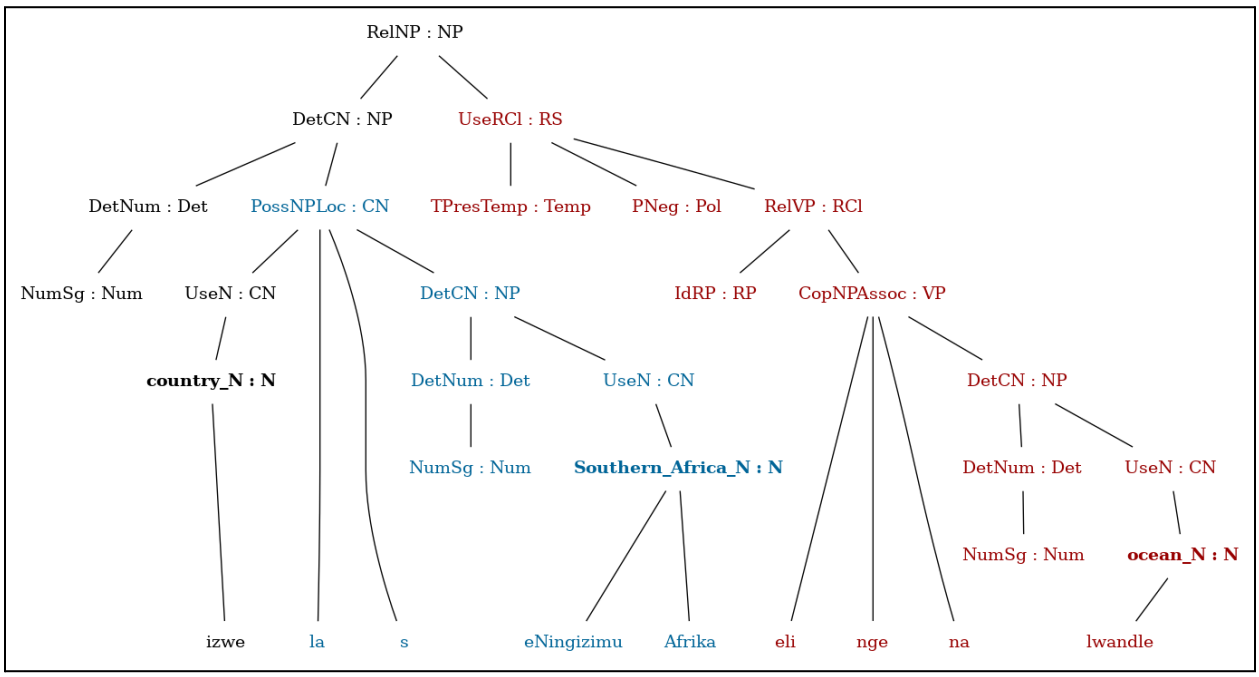

Figure 3: The GF tree linearised as isiZulu text: **izwe laseNingizimu Afrika elingenalwandle**

By answering these questions, we will expand the ability of Wikidata to be read and edited by

humans, especially those in Sub-Saharan Africa, with its 400-600 Bantu languages. This will benefit the multilingual support of Wikidata to other Wikimedia projects.

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

Related work

The gf-wikidata[2] project aims to generate Wikipedia articles from Wikidata (Ranta 2023). However, the current prototype has been tested for European languages that are resource rich to varying degrees.
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). A current project will contribute RGs[3] for Siswati and isiXhosa via bootstrapping. RGs for other Bantu languages are also in development (Kituku et al. 2021, Bamutura et al. 2020).

Methods

1. Collect and employ existing terminology resources and develop new terminology in consultation with expert isiZulu linguists/lexicographers (See footnote 1).
2. Follow a standard methodology for domain-specific GF grammar development:
    - Elicit a representative sample of text from the domain
    - Analyze the sample to design a GF-based model of the domain
    - Develop GF lexicon required by the domain
3. NLG component: develop a module that queries Wikidata and constructs suitable GF trees.

Expected output

1. isiZulu labels and descriptions for the countries in Wikidata. Audience: isiZulu users of Wikidata. We hope to encourage the growth of this audience via this project.
2. Open-source baseline GF-based NLG system for isiZulu descriptions. Audience: researchers interested in extending/adapting the system.
3. Scientific publication detailing process and findings. Audience: scientific community interested in Wikidata and Bantu languages.

Risks

The greatest risk is the resource-scarceness of isiZulu in that the appropriate terminology may be nonexistent or unavailable.

Community impact plan

- Expansion of the human readable isiZulu content in Wikidata, a project aimed at a wide audience.
- Engagement with terminology stakeholders, including PanSALB[4], the South African NLUs[5], USAf[6], SADiLaR[7] etc.

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

[4] Pan South Africa Language Board (https://www.pansalb.org/)
[5] National Lexicography Units (https://sanlu.africa/)
[6] Universities South Africa (https://usaf.ac.za/)
[7] South African Centre for Digital Language Resources (https://sadilar.org/)

- Liaising with the SADiLaR-Wikipedia-PanSALB[8] project.

Evaluation

We will evaluate the *accuracy* and *coverage* of:

1. Terminology collected/developed and contributed to Wikidata as isiZulu labels
2. GF grammar for the domain
3. isiZulu descriptions contributed to Wikidata

Expert linguists will be consulted to evaluate accuracy.

Budget

Task	**Budget ($)**
Terminology - Training linguist to use computational tools - Collection, organization and formalization of terminology into computational artifact - Stakeholder engagement	20 000
GF grammar - Domain analysis - Abstract and concrete models of domain	10000
Wikidata-based NLG - Generation of isiZulu descriptions from Wikidata	6000
Contribution of labels and descriptions to Wikidata	2000
Scientific publication	2000
Total	**40 000**

8 https://sadilar.org/index.php/en/2-general/416-swip

Prior contributions

- **Scientific journal articles:** Marais & Pretorius (2023a and 2023b), Pretorius (2016), Pretorius & Wolff (2020), Ranta (2023).
- **Workshops:** Pretorius (2015, 2016, 2017, 2020).
- **Talks and interviews:** Pretorius (2015, 2016, 2017, 2018, 2020, 2021, 2023).
- **Other:** Kotzé & Pretorius (2020).

References

Bamutura, D., Ljunglöf, P., Nabende, P.: Towards Computational Resource Grammars for Runyankore and Rukiga. In: Language Resources and Evaluation (LREC) 2020, pp. 2846–2854. Marseille, France (2020)

Kituku, B., Nganga, W., Muchemi, L.: 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 (2021). DOI 10.1109/ICT4DA53266.2021.9672222

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, 2020.

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: Mehmood, R., et al. Distributed Computing and Artificial Intelligence, Special Sessions I, 20th International Conference. DCAI 2023. Lecture Notes in Networks and Systems, vol 741. Springer, Cham. https://doi.org/10.1007/978-3-031-38318-2_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 (AALS), 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 2020. ISBN:9781108498821.

Ranta, A. (2023). Multilingual Text Generation for Abstract Wikipedia in Grammatical Framework: Prospects and Challenges. In: Loukanova, R., Lumsdaine, P.L., Muskens, R. (eds) Logic and Algorithms in Computational Linguistics 2021 (LACompLing2021). Studies in Computational Intelligence, vol 1081. Springer, Cham. https://doi.org/10.1007/978-3-031-21780-7_6