Bringing children's dictionaries to digital life

Wilken, Ilana
CSIR
iwilken@csir.co.za

Marais, Laurette
CSIR
lmarais@csir.co.za

Abstract

South Africa is facing a literacy crisis, with the latest PIRLS results showing that 8 out of 10 learners cannot read for basic comprehension by the time they leave the foundation phase. In this climate, the development of strategies to assist educators in harnessing the available resources to maximum effect is needed. However, most teaching resources are not digitally available, and even fewer are available in formats that make them readily available for use in natural language applications.

The *Ngiyaqonda!* project aims to provide an interactive, multimodal digital environment within which learners can practise their reading and writing skills. Computational grammars and speech technology are combined in a mobile application to facilitate the transition from oral competency in a language to written competency. In this paper, we show how words from a multilingual dictionary for foundation phase learners can be brought to digital life within the *Ngiyaqonda!* application to enhance the learning experience of core concepts and vocabulary.

We use the official foundation phase CAPS English-isiZulu dictionary (Mbatha et al. 2018) to ensure that the content of the computational grammars is aligned with relevant learning outcomes. The result is a fully parallel, multilingual computational grammar that is aligned at the semantic level, ready to be included in the *Ngiyaqonda!* application.

Keywords: literacy, Grammatical Framework, text-to-speech

1 Introduction

South Africa is a multilingual country and has 12 official languages. Children are taught in their home language during the foundation phase (Grades 1-3), but low literacy levels remain of concern. In South Africa, the foundation phase is seen as the *learning to read* phase and after that, the *reading to learn* phase starts (Spaull & Hoadley 2018). But even though many literacy initiatives have been put in place across South Africa and by various departments (Pretorius & Klapwijk 2016), the latest Progress in International Reading Literacy Study (PIRLS) results state that South Africa was the lowest-performing country of the 50 countries that participated in the study. In summary, "around 78% of South African Grade 4 learners do not reach the international benchmarks and therefore do not have basic reading skills by the end of the Grade 4 school year, in contrast to only 4% of learners internationally" (Howie et al. 2017).

In an effort to address the literacy crisis that South Africa is facing, a multimodal and multilingual application named *Ngiyaqonda!* (Marais et al. 2023) is being developed. The *Ngiyaqonda!* application aims to facilitate literacy development as well as language learning by using the content of existing children's stories to create lessons and tasks that children can complete in their home language. Grammar-based natural language generation is used to create word-building or sentence-building tasks and the sentence-building tasks are guaranteed to be both grammatically and semantically correct. Children are thus not only learning words in their home language but also what well-structured sentences in their home language typically look like. In addition, the children are learning English as well, because English is introduced later on in the application as they progress through lessons. English is added since the language of teaching and learning (LOLT) changes to English in Grade 4 for a large majority of South African learners.

A lack of suitable resources, especially in digital format, hampers the ability to democratise knowledge effectively. Having access to digital dictionaries has

become important in today's digital age and many works have been published that report on such efforts using either existing dictionaries or creating digital dictionaries from scratch (Nied Curcio 2022, Marye 2022, Makarov et al. 2022, Shyrokov et al. 2022). However, to date, nothing similar to what is described in this paper has been done. This paper describes how existing paper-based resources like multilingual dictionaries for the foundation phase can be brought to digital life.

Table 1: Extract from the digital dictionary

	SENTENCE	NOUN	VERB
1.	*Unekhala elilodwa.*	*ikhala*	
	You have one nose.	nose	
2.	*Nginemilenze emibili.*	*umlenze*	
	I have two legs.	leg	
3.	*Ngibona ngamehlo ami.*	*iso*	*bona*
	I see with my eyes.	eye	see
4.	*Ngihogela ngekhala lami.*	*ikhala*	*hogela*
	I smell with my nose.	nose	smell

2 Creating digital dictionaries

2.1 Domain and vocabulary analysis

Since the *Ngiyaqonda!* application is aimed at foundation phase learners, the official foundation phase CAPS English-isiZulu dictionary (Mbatha et al. 2018) was used to start the process of creating digital dictionaries. The particular dictionary that was used was developed to assist learners with either learning English or isiZulu. The dictionary is divided into categories, which are in turn divided into different themes. One such category is *All about me*, including themes such as *My body*, *My senses*, *My feelings*, *Things I can do*, etc.

Each theme lists words from different parts of speech, namely nouns, verbs, adjectives and adverbs. All the words are listed in English and the isiZulu translation is given below it. The words are also supported with images of individual items, but discussion images are also included. The discussion images are usually accompanied by a sentence, which is given to prompt a learner to generate their own sentences using the words given on a particular theme.

An aspect of the morphology of an agglutinating language like isiZulu is that it is often the case that the words as they appear in the dictionary cannot be re-used in a slightly different context without some alteration to the words themselves. We provide some examples of this phenomenon in Section 2.3. While home language speakers of the language may know that this morphophonological change occurs and be able to use the word in this form when speaking, it nevertheless presents a new decoding challenge for a learner who is still acquiring literacy. However, it is infeasible to include all or even a small subset of the isiZulu words in which the root *-limi* is the base lemma in a paper dictionary. A computational mechanism for making such forms available in a suitable syntactic context could provide learners with a more practical way of engaging and experimenting with the vocabulary in the dictionary.

Towards this end, the words and sentences for each category and theme were manually captured on a spreadsheet. The words were organised according to their part of speech. If the vocabulary did not have a sentence in which it appears, then sentences were also manually developed. The developed sentences followed the same structure as the existing sentences: they are short and appropriate for the foundation phase. Table 1 provides an extract from the spreadsheet, a digitised fragment of the words found in the dictionary.

The way the sentences were developed was done to specifically allow for learning by repetition. Most of the sentences can be changed by just changing the nouns. For example: *You have two hands.* can easily be changed to *You have two feet.* or *You have two eyes.*, etc. This method will repeat the first part of the sentence and teach the learner the structure of the sentence. For verbs, a similar method can be followed. *I see with my eyes.* can become *I taste with my tongue.*, etc. The sentences can also be changed into the negative and the process can be repeated. For

example: *I hear with my ears.* becomes *I do not hear with my ears.* or *I touch with my hands.* can become *I do not touch with my hands.* by simply changing the nouns and verbs.

2.2 Grammar development

Grammatical Framework (GF) is a formalism, programming language and runtime system for grammar engineering (Ranta 2011). It distinguishes between abstract syntax, which models the compositional relations between concepts, and concrete syntax, which defines how concepts are expressed in different natural languages. Multilingual domain-specific grammars for the South African languages have been used to facilitate speech-to-speech translation in the health domain (Marais et al. 2020), and they are currently employed in the *Ngiyaqonda!* project to support literacy development.

2.3 The role of resource grammars

Within the GF ecosystem, resource grammars are typically used as linguistic software libraries to enable rapid application grammar development. A resource grammar represents a comprehensive implementation of the morphology and syntax of a language, enabling application grammar engineering to focus on domain modelling instead of the linguistic details of natural language. This is particularly useful for morphologically complex languages such as isiZulu, where it would be especially tedious to repeatedly reimplement nominal classification and concordial agreement (the characteristic features of the language). The isiZulu GF Resource Grammar is the first comprehensive implementation of isiZulu morphosyntax of its kind to enable rapid application grammar development.

The example sentences in Table 1, which have been captured directly from the dictionary, each show clearly how isiZulu words cannot simply be used in sentences without knowledge of the morphosyntax of the language. In the first sentence, the noun *ikhala* is used with the associative copulative prefix *na-*, which causes the adjacent morphemes to fuse, resulting in the word *unekhala*. In the second example, the noun *umlenze* is used in the copulative construction, but in its plural form, where similar morphophonological alternation takes place as in the previous example. The result is the word *nginemilenze*, where only the stem of umlenze has been retained.

The third example involves an irregular noun: the singular form is given as *iso* or *ihlo* (eye), but its plural form is *amehlo*, which is used with the instrumental prefix *nga-* in the word *ngamehlo*. In this case, the noun stem itself begins with *i*, which is elided when the singular prefix is used, and which causes the plural prefix to undergo a sound change when prefixed to the stem. Finally, in the fourth example, similar alternation occurs between *ikhala* and *nga-* as in the first example.

The resource grammar is the key to effectively dealing with this kind of morphosyntactic complexity: pre-existing functions take care of the relevant details, so that the application grammar engineer is only tasked with capturing the concepts in the domain in terms of their linguistic categories.

2.4 From domain analysis to application grammar

The domain analysis presented in the previous section can be used as the basis for a dictionary-based, domain-specific application grammar. The analysis of the lexical items form the basis of the computational lexicon, while the example sentences can be used to inform template sentences that can exhibit differing levels of compositional variability.

To demonstrate how the contents of a themed dictionary can be made available in an interactive, enhanced way, we have chosen to implement the *My Body* domain with two sentence templates, five verbs, four numerals, two (implicit) pronouns and 25 nouns. We have also allowed for the sentences to be expressed positively and negatively. Given these few basic building blocks, a grammar supporting 460 multilingual sentences was developed.

```
1 lincat
2     BodyStatement = S ;                          -- utterance
3     BodyClause = Cl ;                            -- clause
4     Person = { np : NP ; agr : BodyAgr } ;       -- noun phrase and agreement
5     BodyPart = N ;                               -- noun
6     SenseBodyPart = { n : N ; num : Num } ;      -- noun and number
7     BodyPartNumber = { rs : RS ; num : Num } ; -- relative sentence and number
8     SenseAction = V ;                            -- verb
9     Polarity = Pol ;                             -- polarity
```

Figure 1: Domain concepts mapped to their syntactic components

An essential aspect of application grammar development when using a resource grammar, is the mapping of domain concepts to syntax categories. Figure 1 shows the mapping for the *My Body* domain. Most categories can be seen to map directly to an existing category in the resource grammar, such as `SenseAction` mapping directly the category for verbs (`V`) as seen in line 8, while some are composed of more than one, such as `SenseBodyPart` containing mappings to nouns (`N`) and number (`Num`, for singular and plural), as seen in line 6. The code snippet shown is the declaration of the so-called linearisation categories in a GF concrete syntax, introduced by the keyword `lincat`. Using this mapping, the functions of the resource grammar can be called upon to combine the syntactic components associated with the domain concepts in order to generate grammatically correct and semantically transparent natural language.

This is further illustrated in Figure 2. The tree on the left is made up of nodes representing syntactic functions from the resource grammar, while the tree on the right consists of nodes representing semantic functions from the application grammar. For example, the application function `UseSenseBodyPart` accepts three arguments of type `Person`, `SenseBodyPart` and `SenseAction` to produce a `BodyClause`. In contrast, the resource function `PredVP` accepts two arguments of type `NP` (representing a noun phrase) and `VP` (representing a verb phrase) to produce a `Cl` (a clause of which the tense and polarity have not yet been specified).

Color has been used to indicate which subtrees of the application grammar tree maps to the nodes of the resource grammar tree. Some mappings are relatively simple, such as the application function `Smell` mapping to the resource function `hogel_V` which represents the verb *hogela* (smell) (compare line 8 of Figure 1). On the other hand, the application function `Eye` maps to two parts of the resource grammar tree, namely the function `iso_5_6_N` which represents the (irregular) noun *iso, amehlo* (eye, eyes) as well as the information that the bodypart in question should be referred to in the plural, using the `NumPl` function (compare line 6 of Figure 1). Once these mappings are defined, the resource grammar can be relied upon to ensure that linguistic features such as concordial agreement are handled correctly.

3 Challenges and opportunities

Not only do GF application grammars provide the ability to deal effectively with the complex morphosyntax of isiZulu, but they are supported by a runtime system which enables them to be integrated into real solutions to societal challenges. In this section we illustrate this to show how our solution serves to bring a children's dictionary to digital life.

3.1 Dealing with qualificatives

One challenging aspect of bilingual dictionaries that have been developed from an English word list and translated to a language like isiZulu are the ad-

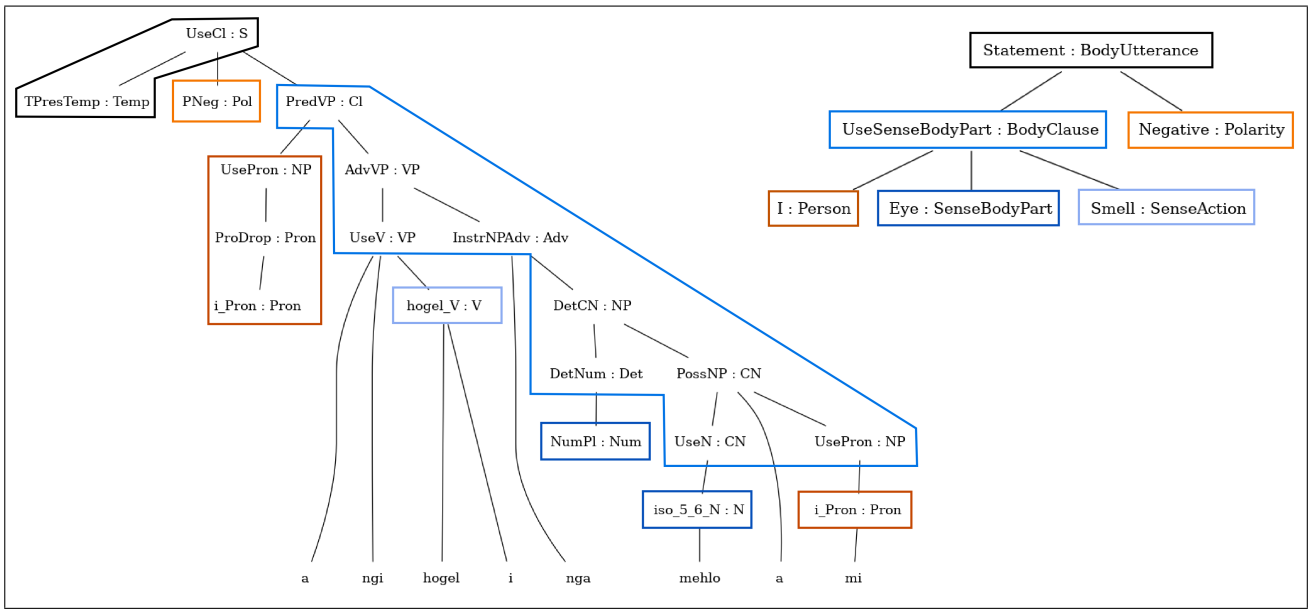

*Figure 2: Visual representation of mapping between resource grammar syntax tree and application grammar semantic tree to represent 'Angihogeli ngamehlo ami.' (*I don't smell with my eyes*).*

Table 2: Examples of qualificatives

NUMERICAL CONCEPT	ISIZULU SENTENCE	ENGLISH SENTENCE
One	Unomlomo *owodwa.*	You have one mouth.
One	Unenkaba *eyodwa.*	You have one bellybutton.
Two	Nginezingalo *ezimbili.*	I have two arms.
Two	Unamehlo *amabili.*	You have two eyes.
Ten	Ngineminwe *eyishumi.*	I have ten fingers.
Ten	Anginazinyawo *eziyishumi.*	I don't have ten feet.

jectives. Their equivalent in the Bantu languages are the qualificatives, which "straddle a number of morphosyntactic categories" (Mojapelo 2014). This means that their use in example sentences are even more often subject to the kind of morphosyntactic variation seen in Table 1. In the *My Body* grammar, the numbers (representing one, two, five and ten) are an example of qualificatives that are constructed using differing linguistic structures, and which must exhibit morphological agreement with the nouns that they modify. By modelling these qualificatives as relative sentences, the grammar generates the correct forms of these qualificatives, regardless of the noun class in question. Table 2 gives some examples of new sentences generated by the grammar in both isiZulu and English.

3.2 Integrating the grammar into a mobile application

The sentences are presented in an interactive, multimodal environment via the *Ngiyaqonda!* application, as shown in Figure 3. Users of the application can drag tokens from a selection area to an authoring area in order to construct grammatically correct sentences. Using text-to-speech, all completed sentences can be verbalised in both isiZulu and English. A gamified version of the activity randomly generates sentences as prompts, and users must recreate the sentence they hear. Even though the application in its current form requires a stable internet connection, it is possible to embed the grammars in an offline version as well. A more detailed discussion

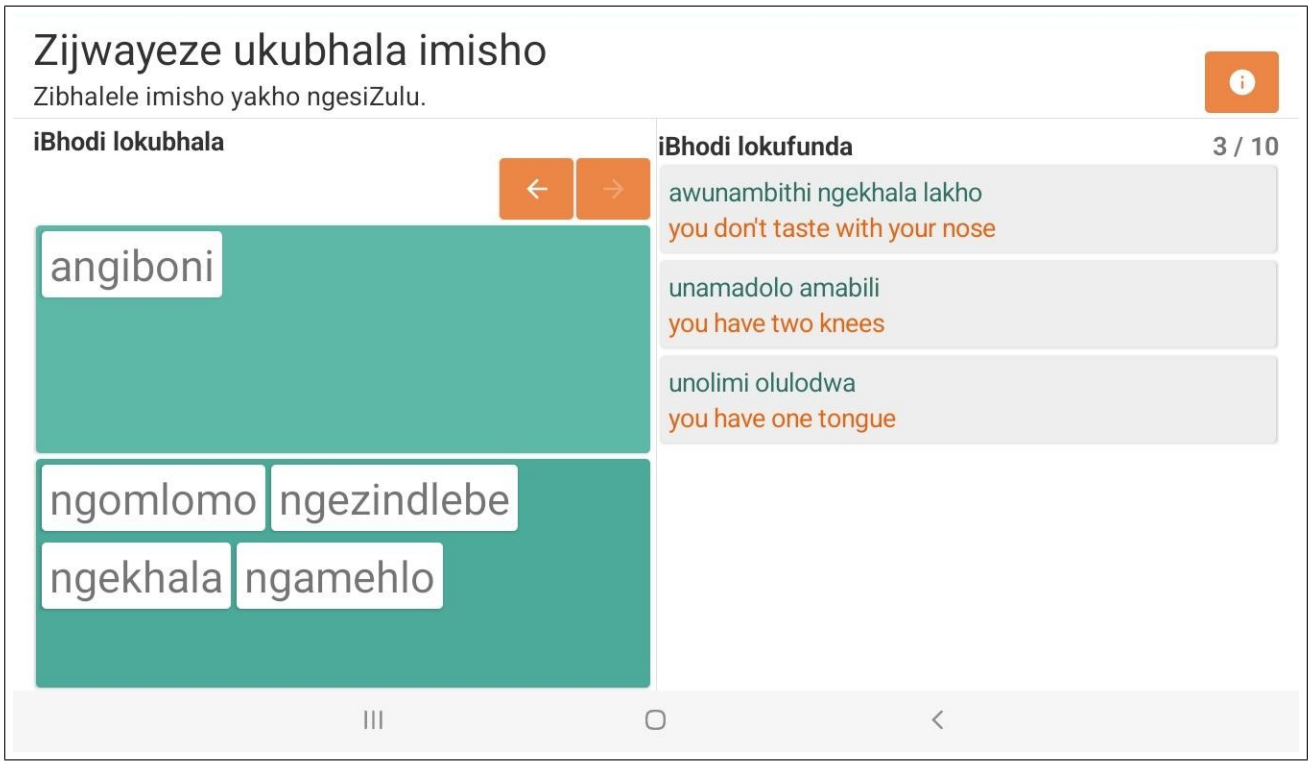

Figure 3: Screenshot of a writing activity using the My Body *multilingual grammar*

of the application is beyond the scope of this paper: the interested reader is referred to Marais et al. (2023).

The approach described here is scalable in a number of important ways. We have presented the development of one domain as an example, but in principle, all the domains of the CAPS dictionary could be included in a full-scale digitisation effort, which comprises 7 categories covering 47 themes. Furthermore, scalability with regard to supported languages is an essential aspect of supporting literacy development in South Africa. The dictionary-based application grammars rely on resource grammars for the modelling of morphosyntax: as more resource grammars are developed for the South African languages, the development of dictionary-based application grammars in these languages will become possible.

3.3 Piloting the grammars

While the *Ngiyaqonda!* application is currently being piloted at a school in Soweto, Gauteng, this pilot was conducted using a narrative-based grammar. Preliminary feedback indicates that teachers and learners benefit from using the application, but a need for dictionary-based grammars has already been voiced. Grammars such as those described in this paper will therefore be included in a subsequent pilot at the same school. Qualitative and quantitative analyses will be performed to determine the effectiveness of the dictionary-based grammars via the application.

4 Conclusion

We have presented an approach to harnessing the organised content of a children's dictionary in order to make it available in an enhanced and accessible way. This will enable the rapid development of content that can be used for the creation of digital teaching resources in the written official languages of South Africa.

In this way, the vital work of dictionary development for the low-resourced languages of South Africa can be enhanced by making the content and organisational structure of it available in an interactive, digital environment. In particular, the limitations of paper-based dictionaries to deal with the complex morphosyntax of languages like isiZulu can be overcome through the use of computational methods.

Acknowledgements

The *Ngiyaqonda!* project is sponsored by the South African Department of Sport, Arts and Culture. The development of the isiZulu GF Resource Grammar was sponsored by The South African Centre for Digital Language Resources (SADiLaR).

References

Howie, S., Combrinck, C., Roux, K., Mokoena, M. & Palane, M. (2017), South africa grade 4 pirls literacy 2016 highlights report: South africa, Technical report.

Makarov, Y., Melenchenko, M. & Novokshanov, D. (2022), Digital Resources for the Shughni Language, *in* 'Proceedings of The Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference', pp. 61–64.

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)', IEEE, pp. 1–8.

Marais, L., Wilken, I., Pretorius, L. & Posthumus, L. C. (2023), Multimodal, multilingual dynamic stories for literacy development and language learning, *in* 'Proceedings of the 5th International Conference on Conversational User Interfaces', CUI '23, Association for Computing Machinery, New York, NY, USA.
URL: *https://doi.org/10.1145/3571884.3604303*

Marye, H. S. (2022), 'Digitization of the Concise Sociopolitical and Mass-Media Dictionary (English—Amharic)', *Journal of the Text Encoding Initiative* .

Mbatha, M., South African National Lexicography Units & IsiZulu National Lexicography Unit (2018), *Official foundation phase CAPS English-isiZulu picture dictionary*, South African National Lexicography Units.

Mojapelo, M. L. (2014), Morphosyntactic discrepancies in representing the adjective equivalent in african wordnet with reference to northern sotho, *in* 'Proceedings of the Seventh Global Wordnet Conference', pp. 355–362.

Nied Curcio, M. (2022), Dictionaries, foreign language learners and teachers. new challenges in the digital era, *in* 'Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany', IDS-Verlag, pp. 71–84.

Pretorius, E. & Klapwijk, N. (2016), 'Reading comprehension in south african schools: Are teachers getting it, and getting it right?', *Per Linguam* **32**, 1–20.

Ranta, A. (2011), *Grammatical Framework: Programming with Multilingual Grammars*, Vol. 173, Center for the Study of Language and Information/SRI.

Shyrokov, V., Ostapova, I., Kupriianov, Y., Dorozhynska, A., Yablochkov, M. & Verbynenko, I. (2022), Terminology dictionary digitalization, *in* 'Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022). Volume I: Main Conference', pp. 3–15.

Spaull, N. & Hoadley, U. (2018), 'Getting reading right: Building firm foundations', *ChildGauge* p. 201777.