Bootstrapping syntactic resources from isiZulu to Siswati

Laurette Marais[1], **Laurette Pretorius**[2], **Lionel Posthumus**[3]
[1]Voice Computing Research group, CSIR
[2]Division of Computer Science, Department of Mathematical Sciences, Stellenbosch University
[3]CALT@UJ, University of Johannesburg
lmarais@csir.co.za, lpretorius@sun.ac.za, lionelp@uj.ac.za

Abstract

IsiZulu and Siswati are mutually intelligible languages that are considered under-resourced despite their status as official languages. Even so, the available digital and computational language resources for isiZulu significantly outstrip those for Siswati, such that it is worth investigating to what degree bootstrapping approaches can be leveraged to develop resources for Siswati. In this paper, we present the development of a computational grammar and parallel treebank, based on parallel linguistic descriptions of the two languages.

Keywords: Grammatical Framework, parallel treebanks, computational grammar

1. Introduction

IsiZulu and Siswati[1] are Southern Bantu languages that belong to the Nguni group, and as such are morphologically rich languages that have a noun class system which in turn generates concordial agreement. The Nguni languages have a conjunctive orthography and also exhibit significant morphophonological affixing, leading to long tokens for which morphological analysis is non-trivial.

The Nguni languages are mutually intelligible (Ndhlovu, 2022), and this characteristic allows for exploitation in an under-resourced context. While isiZulu is an official language of South Africa[2], and Siswati an official language of South Africa and the Kingdom of Eswatini[3], they are both under-resourced, Siswati significantly more so than isiZulu (Moors et al., 2018).

Previous work by Bosch et al. (2008) showed the feasibility of bootstrapping finite state morphological analysers following a systematic approach. In this case, isiZulu served as the starting point from which resources for other Nguni languages could be developed. Some of the key findings of this work was that bootstrapping between the Nguni languages drastically reduces development time, which can be significant in the context of under-resourced languages. A bootstrapping approach also results in special focus being given to the differences between the languages: "By exploiting correspondences and linguistic relatedness, more effort may be spent on those aspects in which the languages differ, ensuring end products of superior quality, both linguistically and computationally." (Bosch et al., 2008, p. 85)

A natural next step would be to explore application of the bootstrapping approach beyond morphology to syntax. Our point of departure for this work is the Grammatical Framework (GF) isiZulu resource grammar, with the primary goal of bootstrapping a Siswati resource grammar. In the process, we develop a parallel treebank by hand, which we then augment using the parallel resource grammars to achieve a larger semi-synthetic treebank - a first for Siswati. We evaluate the resource grammars by manually evaluating a subset of the augmented data to ensure that the functions of the grammars behave as expected when combined in new ways.

We based our bootstrapping methodology on a set of two textbooks, on isiZulu and Siswati respectively, in order to ensure a systematic and linguistically aware approach. Even here, the Siswati textbook (Taljaard et al., 1991) is "largely based on" the isiZulu textbook (Taljaard and Bosch, 1988) and features the two authors of the isiZulu book alongside a specialist Siswati linguist. In a certain sense, we rely on the "bootstrapping" of high quality linguistic descriptions of the language by linguists in order to guide a systematic and reliable bootstrapping approach to computational resources.

2. Background

Bootstrapping of resource grammars, specifically GF resource grammars, has been done for various related languages, with the most relevant being the work on Runyankore and Rukiga by Nabende et al. (2020), as well as the work on the Kenyan Bantu Languages (Ekegusii, Kikamba and Swahili) by Kituku et al. (2021). Due to the under-resourced status of these languages, suitable evaluation corpora do not exist and require special development. Con-

[1]The three letter language codes for isiZulu and Siswati are zul and ssw respectively.

[2]IsiZulu has the largest number of L1 speakers of all the (Nguni) languages, namely around 15 million, while Siswati has around 3 million.

[3]Also known by its former official name Swaziland.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 77–85
25 May, 2024.

sequently, a full evaluation of the resource grammars for Runyankore and Rukiga has not been reported on. Evaluation for the Kenyan Bantu languages was focused on software engineering aspects of bootstrapping, with no specific mention of the final correctness of the grammars. The language fragments used in iterative testing during development were translated from English examples illustrating the purpose of each function in the grammar. In terms of coverage, then Kenyan Bantu language resource grammars are not as mature as the isiZulu resource grammar. For example, in the GF Github repository, the Kenyan Bantu language functor only contains one function for constructing verb phrases, namely `UseV`, which is used for intransitive verbs. The isiZulu (and now Siswati) resource grammars, by contrast, include 21 functions for constructing verb phrases, covering also transitive verbs, the reflexive construction, the copulative constructions, adverbial modification and verbs with verb and sentence complements.

Therefore, although previous GF work exists for other Bantu languages, it is difficult to provide a direct comparison of our work to these other efforts.

Our aim is to exploit existing linguistic resources for isiZulu and Siswati in order to base our bootstrapping of the Siswati resource grammar on a systematic and parallel exposition of the linguistic characteristics of the two languages.

3. Comparison of isiZulu and Siswati

As in all Bantu languages, the structure of isiZulu and Siswati is based on two principles, viz. nominal classification (the system of noun classes) and concordial agreement across various word categories (the system of concords). (These are but 2 outstanding characteristics of the Bantu languages.)

Generally speaking, the noun consists of two main parts, viz. a noun class prefix and a noun root/stem. Furthermore, every noun belongs to a so-called noun class by virtue of the form of its prefix, also referred to as its class gender. This notion of class gender is significant since it generates grammatical agreement by means of these class prefixes, also termed gender number prefixes. These noun classes are numbered, with the noun class system of isiZulu and Siswati being very similar.

A concord is a structural element (agreement marker/morpheme) which formally marks the relationship between a noun and all other words in a sentence that have a direct semantic-syntactic relationship with the noun. The above-mentioned gender agreement must be observed in all parts of the utterance which are directly linked to the noun. Therefore, we say that word categories such as verbs, pronouns, adjectives, relatives, possessives etc. are brought into concordial (i.e. grammatical) agreement by means of these concords. Examples (1) and (2) show an isiZulu and a Siswati sentence, respectively.

(1) *Leli bhubesi li-zo-yi-luma*
Dem5 NStem5 SC5-Fut-OC9-VStem
in-komo ya-mi
NStem9 PC9-PPron1PSg
'This lion will bite my cow.'

(2) *Leli-bhubesi li-to-yi-luma*
Dem5-NStem5 SC5-Fut-OC9-VStem
in-khomo ya-mi
NStem9 PC9-PPronP1Sg
'This lion will bite my cow.'

Before listing a number of systematic differences between isiZulu and Siswati that we exploit in our bootstrapping process, we take a closer look at examples (1) and (2). The one noun root *-bhubesi*, the verb stem, the class 5 demonstrative, the class 5 subject concord, the class 9 object concord, the class 9 possessive concord and the possessive pronoun, first person singular, are identical. Moreover, in both languages the noun root for 'cow' is *-khomo*. However, in isiZulu the class 9 surface form is subject to a morphophonological alternation rule and is realised as *-komo*. Finally, the future morpheme is *-zo* in isiZulu and *-to* in Siswati.

As a point of departure, important regular morphophonological differences between the two languages may be systematised as follows (Mordaunt et al., 2023; Bosch et al., 2008; Taljaard and Bosch, 1988; Taljaard et al., 1991):

1. The alphabet and click omission: While both languages use the Latin alphabet (A-Z), Siswati omits Q and X, while in isiZulu /q/ and /x/ represent click consonants. In isiZulu the click sounds /c/, /q/ and /x/ are represented by the click sound /c/ in Siswati. for example, *-qina* (zul) and *-cina* (ssw) both mean 'be hard'.

2. Consonant substitution or addition: The /z/ that often occurs in isiZulu roots/stems and in the class 8 and 10 prefixes and concords, is usually substituted with /t/ in Siswati. for example, *-zama* (zul) and *-tama* (ssw) both mean 'try'.

   The /th/ and /t/ in isiZulu is usually realised as /tf/ when followed by /o/, /u/ and /w/, and as /ts/ when followed by /a/, /e/ and /i/ in Siswati. Examples are *-thola* (zul) and *-tfola* (ssw), which mean 'find', and *-thatha* (zul) and *-tsatsa* (ssw), which mean 'take'.

   The /d/ in isiZulu converts to /dv/ when followed by /o/, /u/ and /w/, and to /dz/ when followed by /a/, /e/ and /i/ in Siswati, for example *-dubula* (zul) and *-dvubula* (ssw), meaning

'shoot', and *-dabula* (zul) and *-dzabula* (ssw), meaning 'tear'.

Other differences are the consonant clusters /mp/ and /nk/ in isiZulu that become /mph/ and /nkh/ in Siswati, for example *impendulo* (zul) and *imphendulo* (ssw), meaning 'reply'.

3. Pre-prefix vowel deletion, addition and substitution: The isiZulu noun class prefix consists of a consonant-vowel sequence (also referred to as the basic prefix), preceded by a so-called augment (also referred to as class pre-prefix), a preceding copy vowel that fulfils different grammatical functions, e.g. definiteness and specificity, and is subject to morphophonological processes such as vowel deletion and coalescence. In Siswati this augment is only present in classes 1, 3, 4 and 6 and in class 9 (where it precedes a nasal consonant). Moreover, in class 6 this pre-prefix is /e/ and not /a/. An example is ***a****makati* (zul) and ***e****makati* (ssw) for 'cats'.

4. The relative construction and concords: Whereas the relative construction in isiZulu has *a-* as so-called relative morpheme, the relative morpheme in Siswati is *la-*. In both languages the *a-* and *la-* respectively assimilates with the vowel of the basic prefix and vowel coalescence takes place across the consonant to form the relative concord. An example is *umfana omunye* (zul) and *umfana lomunye* (ssw), from *a+munye* and *la+munye*, meaning 'another boy'.

5. Lexical items: While the two languages share many noun and verb roots/stems, lexically there are differences, for example *-phuza* (zul) and *-natsa* (ssw), meaning 'drink'.

6. Orthography: In isiZulu, demonstratives are written disjunctively from the noun that follows, while in Siswati the first position demonstrative ('this/these') is written conjunctively with the following noun, as in example (1): *leli bhubesi* (zul) versus *lelibhubesi* (ssw), meaning 'this lion'.

7. The imperative: In isiZulu monosyllabic verb stems, *yi-* or *i-* are prefixed or *-na* suffixed to the stem for the imperative directed at one person. In Siswati *-ni* is suffixed to the verb stem, for example ***Yi****dla/****I****dla/Dla****na****!* (zul) and *Dla****ni****!* (ssw), meaning 'Eat!', directed to one person.

In summary, the differences 1-3 above apply to the two languages across all constructions and lexical items. Complementary to this general exposition, are the word category and grammatical construction based parallel expositions of Taljaard and Bosch (1988) and Taljaard et al. (1991), the latter two providing practical grammar orientated perspectives, ideally suitable for direct application to and implementation in the bootstrapping of the Siswati grammar from the isiZulu RG.

4. GF isiZulu resource grammar

The isiZulu resource grammar (isiZulu RG) used in this work is implemented in Grammatical Framework (GF), a computational grammar framework for the development of multilingual grammars. The framework utilises an interlingua architecture, such that a GF grammar consists of an abstract syntax and one or more concrete syntaxes, one for each language. Abstract categories and functions are defined in the abstract syntax, which are implemented in the concrete syntaxes as linearisation categories and linearisation functions. The GF runtime enables linearisation of abstract syntax trees into natural language strings, as well as parsing of natural language strings into abstract syntax trees (Ranta, 2011).

GF resource grammars typically form part of the Resource Grammar Library (RGL), which shares a common abstract syntax and custom extensions between over 40 languages (Ranta et al., 2020). The categories and functions are syntactic in nature, with categories for nouns, noun phrases, verbs, verb phases, adverbial phrases, clauses, sentences, etc., along with functions for combining these categories into tree structures.

Originally, the intent of the RGL was to serve as a linguistic software library to enable rapid development of application specific grammars (Ranta, 2009). The implementation of the syntactic categories and functions would capture the general morphology and syntax of the language, which could then be reused by application grammars for specific use cases. More recently, however, attempts have been made to employ the general use grammars of the RGL towards wide-coverage parsing as well as for bootstrapping Universal Dependencies treebanks (Ranta et al., 2020).

The isiZulu RG models the morphology and syntax of isiZulu via the implementation of some functions from the RGL common abstract syntax, in addition to a set of extra language specific abstract functions (Marais and Pretorius, 2023b).[4]

Following an approach typical for the implementation of Bantu languages, the isiZulu RG models the language at the subword level. In short, this means that the base tokens of the grammar do

[4]See the README at https://github.com/GrammaticalFramework/gf-rgl/blob/master/src/zulu/README.md

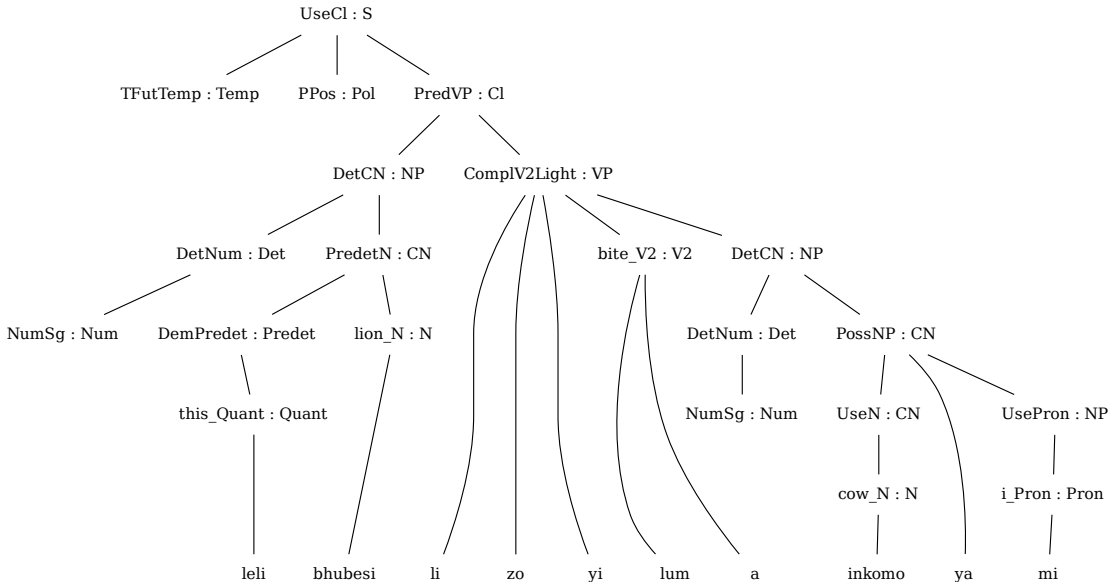

Figure 1: GF parse tree for example (1)

not correspond to orthographic words but to subword segments, which are glued together at runtime using built-in orthography engineering support in the GF C-runtime (Angelov, 2015). An example of this is given in Figure 1, showing how the surface segments of the isiZulu sentence in example (1) in Section 3 are produced by different functions in the isiZulu RG. We will say more about how morphophonological alternation is modelled in Section 6.3.

5. Methodology

Our methodology is depicted in Figure 2. We started with two resources (shown in blue) and from them developed three new resources (shown in orange). The isiZulu RG forms the computational basis for the work, with the set of parallel textbooks providing the linguistic information required to develop and evaluate a new Siswati resource grammar.

The isiZulu RG has so far been used to expand morphosyntatically complex entries in the isiZulu Wordnet (Marais and Pretorius, 2023a), as the general purpose syntactic parser for isiZulu (Marais and Pretorius, 2023b) and as a mechanism for generating annotated data for training morphological segmentation models for isiZulu (Mkhwanazi and Marais, 2024). We therefore consider it to be a mature model of isiZulu and a suitable basis upon which to develop similar models for related languages.

The parallel texts provide us with two kinds of information, namely a parallel linguistic exposition of the two languages, as well as high quality parallel example sentences exhibiting the linguistic features described in the books. The parallel linguistic exposition served as the basis for the development of the Siswati RG, while the parallel examples were used to create a parallel development treebank. Here, the isiZulu RG was used to parse the examples to speed up the process of obtaining a tree representation for each parallel sentence pair.

The treebank itself served as a regression test during development to ensure that adaptations for the Siswati rendered the correct linearisations (natural language strings) from the trees, and it also served to ensure that no errors were introduced in the process of some superficial refactoring of the isiZulu RG in order to minimise code divergence. We give more detail about this process in Section 6.

The final evaluation involved the creation of an augmented treebank based on the one used in development. It was created using a few basic rules defining tree modifications and applied to the development treebank. From the newly created trees, linearisations in both isiZulu and Siswati were generated, and these were manually evaluated. This would ensure that the adaptations that were made to the Siswati on the basis of the linguistic exposition and evaluated during development on the parallel treebank, would generalise to new trees.

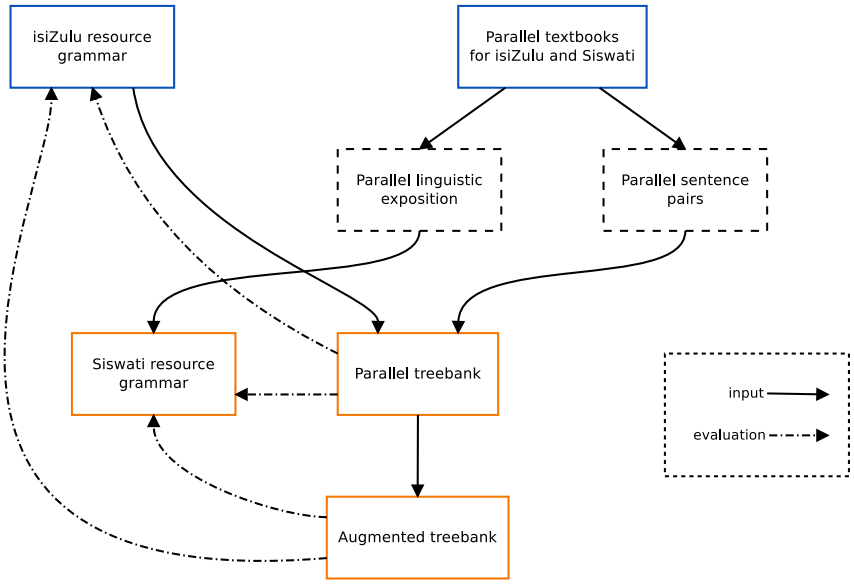

Figure 2: Methodology

6. Adapting the isiZulu grammar to Siswati

6.1. Software logistics

The most naïve way to bootstrap a new GF RG from an existing one is to make a copy of all the relevant files and to change them in specific ways. The GF requirement is that concrete module names have the form 'XY.gf', where X is the name of the abstract module being implement and Y is a code indicating the specific language, ideally based on the language's 3-letter ISO code. The headers of concrete modules also contain this code, and hence a first step would be to systematically change the file and header names. The new resource grammar therefore starts out as an exact copy of the original, which can be changed in precisely those places where the two languages differ.

Of course, the original resource grammar may also be added to or changed while work is ongoing on the new grammar, which could soon cause unnecessary code divergence. GF encourages modular design of grammars, along with the use of functors to model closely related languages according to sound software engineering principles (Ranta, 2009). However, the right moment to functorise a parallel implementation depends on having a good understanding of how the similarities and differences between two or more languages should be modelled. Our intent is to extend this bootstrapping approach to the other Nguni languages and beyond, and hence we have opted not to implement a functor yet, since a prematurely implemented one could turn out to be more of a hindrance than a help.

In order to minimise code divergence following a purely parallel implementation approach, the module system of GF was exploited so that all strings in the grammar (apart from roots and stems included in lexicon modules) are contained in the two main resource modules, namely ResZul.gf and ResSsw.gf. These strings could then be accessed by other modules exclusively via operations defined in the resource modules. This would ensure that differences at the orthographic level would be defined entirely in the respective resource modules, while morphosyntactic differences would be defined in the relevant linearisation functions of the concrete modules. The differences can therefore be monitored at a glance using software that indicates line differences between files, such as `diff`.

The isiZulu RG included a number of custom abstract modules, modelling aspects of isiZulu not found in the common abstract syntax. These files were moved to a folder named 'nguni', so that they could be utilised by both resource grammars.

6.2. Linguistically-driven adaptation

The parallel linguistic exposition of the two textbooks provided a practical and systematic basis for adapting the Siswati RG from the isiZulu. In contrast to the description of similarities and differences as summarised in Section 3, the parallel texts provided a map of where these differences manifest in the respective languages, which simplified the process of identifying which functions and operations in the resource grammars would be different. Not all constructions in the textbook are implemented in the isiZulu RG: we limited the scope of the adaptation to what is currently stable in the isiZulu RG, having established that it has already been used in a number of applications. As such, we excluded from this adaptation the situative

mood, certain interrogative constructions, auxiliary verbs and indirect relatives.

From Section 3 it is clear that morphophonology would play a central role in any bootstrapping effort. We next provide a short description of how morphophonological alternation is modelled in the isiZulu and subsequently the Siswati resource grammars.

6.3. Morphophonological alternation in GF

In GF, morphophonological alternation can be modelled by defining alternative forms of certain morphemes and selecting the correct form to use in a specific context based on one or more parameters supplied by the context. This is necessitated by the fact that the strings of a GF grammar cannot be inspected at runtime, only at compile time.

For example, due to morpheme fusion, the form of the possessive concord depends on the initial sound of the noun or pronoun to which it is prefixed. A parameter called `RInit` is used to keep track of this at runtime, defined to distinguish between the different vowels (with values `RA` to `RU`, as shown in Figure 3) and consonants as a whole (with value `RC`). The table containing the possessive concord is essentially 2-dimensional, with the first dimension representing the agreement information of the possessee, while the second dimension represents the initial sound of the possessor noun or pronoun. Figure 3 shows how this in encoded in a GF table.

Agreement is encoded as a compound parameter in which the first value is a constructor dealing with grammatical person, and the subsequent values deal with grammatical number and class gender where applicable. For example, `First Sg` refers to agreement with the first person singular pronoun, while `Third C3_4 Pl` refers to agreement with plural nouns of classes 3 and 4.

A significant number of adaptations to the Siswati resource grammar consisted of systematically altering the strings contained in tables such as these.

6.4. Changes to the Siswati resource module

Recall that the respective resource modules of the resource grammars were designed to contain the majority of differences between the two languages by containing all strings used in the grammar (apart from a lexicon). In this section we discuss changes made to the ResSsw.gf module, unless otherwise indicated.

The centrality of the noun class system makes nouns an obvious place to start, which is most likely why the two textbooks also devote the first few chapters to nouns, their classes and the associated prefixes. This is dealt with in the resource modules of the RGs in two main operations, `nomNoun` and `locNoun`, for nouns and locativised nouns. Supporting operations deal with the morphophonological alternation which occurs when noun roots/stems are joined with the relevant prefixes and suffixes. These were the first adaptations to be made to the Siswati RG.

The focus then shifted to verbs, starting with alternation that occurs within the verb root/stem, especially as it relates to the verb-final morpheme. After that, the various pre-root verbal morphemes were adapted by making changes to the subject and object concord tables, as well as to the operations for producing the appropriate forms of the tense markers and relative prefix. The forms of the reflexive prefix and relative suffix were also changed.

These changes were sufficient to also cover most of the changes necessary for correctly modelling the copulative constructions, although additional changes to the identifying copulative marker and the adjectival concord were also required. In fact, the identifying copulative prefix is not required in the Siswati grammar, which amounted to a syntactic change that was made in the `VerbExt` module. For example, in isiZulu the sentence 'The lion is an animal' is expressed as *Ibhubesi **yi**silwane*, while in Siswati it is expressed as *Libhubesi silwane* (Taljaard and Bosch, 1988; Taljaard et al., 1991).

The tables containing the absolute, possessive and all three sets of demonstrative pronouns were also changed, along with the possessive and quantitative concords.

Finally, the various adverbial prefixes were changed. This was, perhaps surprisingly, one of the more substantial changes required. In isiZulu, the morphophonological alternation of adverbial prefixes like *nga-* and *njenga-* is based on the class prefix of the noun to which it is prefixed, whereas in Siswati, the alternation is based directly on the class to which the noun belongs, regardless of the form of its prefix. The sound changes also follow a different pattern with regards to the classes compared to isiZulu. Hence, instead of altering strings in a table, the structure of the tables in which the adverbial prefixes were housed was changed, accurately reflecting this difference between the languages.

The other syntactically significant changes that were implemented relate to the imperative, since the morphosyntactic structure of imperatives differ between the two languages when it comes to monosyllabic verb stems and the copulative constructions. These changes were implemented in all modules containing functions for constructing `VP`s (verb phrases).

The most important insight gained during the process of bootstrapping from one Nguni language to another is the centrality of a transparent and sys-

```
param RInit = RA | RE | RI | RO | RU | RC ;

oper poss_concord_agr : Agr => RInit => Str = table {
    First Sg => table {(RA|RC) => "wa" ; (RE|RI) => "we" ; (RO|RU) => "wo" } ;
    First Pl => table {(RA|RC) => "ba" ; (RE|RI) => "be" ; (RO|RU) => "bo" } ;
    ...
    Third C3_4 Sg => table {(RA|RC) => "wa" ; (RE|RI) => "we" ; (RO|RU) => "wo" } ;
    Third C3_4 Pl => table {(RA|RC) => "ya" ; (RE|RI) => "ye" ; (RO|RU) => "yo" } ;
    Third C5_6 Sg => table {(RA|RC) => "la" ; (RE|RI) => "le" ; (RO|RU) => "lo" } ;
    Third C5_6 Pl => table {(RA|RC) => "a" ; (RE|RI) => "e" ; (RO|RU) => "o" } ;
    ...
  } ;
```

Figure 3: Table for the possessive concord, parameterised to contain alternative forms based on the initial sound of the possessor

tematic model of morphophonology. This ensured that the majority of changes required related to the strings in the resource modules that represent morphemes alongside their morphophonological alternatives, with very few changes requiring a more substantial structural change.

7. Developing a parallel treebank

Manually capturing parallel sentences from textbooks and obtaining trees to represent them is a time consuming and therefore expensive task. Consequently, we opted to select about four to five structurally dissimilar sentences from each relevant chapter of the parallel textbooks, although the capturing of all the sentences in the textbooks is continuing. In some places, the same linguistic construction was illustrated in the textbooks using multiple sentences with alternative word orders, some of which have not been included in the isiZulu RG. In such cases, we included the sentences whose word order is already implemented in the resource grammar. While it is in principle possible to implement functions for alternative word orders, the decision to do so must also weigh the computational cost associated with a larger grammar and will be considered in future, as well as the expected frequency in which the alternative word order appears in isiZulu and Siswati corpora. Moreover, the purpose of this work was to bootstrap the existing isiZulu grammar, which we consider to be mature. Inclusion of the additional sentences in the treebank, along with the implementation of functions to support them, is considered future work.

7.1. Obtaining trees

The process of finding trees to represent the sentences was somewhat expedited by employing the GF runtime as a parser. IsiZulu sentences were parsed using the isiZulu resource grammar, along with a large isiZulu lexicon. In almost all cases, the correct tree was selected from among those provided by the runtime. In cases where the syntactic ambiguities of the sentence made selecting from a large number of possible parses difficult, the correct tree was developed by hand on the basis of the context within which it is provided in the textbooks, as well as its English gloss. It was then linearised to isiZulu in order to confirm its correctness.

In this way, a tree was found for 125 pairs of sentences, covering the chapters on nouns, concordial agreement in verbs, adverbial forms, the various tenses of the verb, absolute and demonstrative pronouns, copulative forms, direct relatives, the enumerative, numerals, and the subjunctive form. While this would constitute, to our knowledge, the first treebank for Siswati, it is admittedly quite small. However, in stark contrast to one that would be based on a corpus, the treebank was designed specifically to test a wide variety of linguistic constructions and can therefore be said to be highly representative of the languages. For that reason, it is ideal as the basis for continuous evaluation of a computational grammar during development.

7.2. Lexicon support

The trees as they were developed via parsing using the isiZulu RG, which was paired with a large isiZulu lexicon, included lexical functions based on isiZulu roots and stems. For instance, the tree would use the function `theng_V2` for trees in which the verb *-thenga* (to buy) appeared. Since no computational lexicon currently exists for Siswati, the required lexical functions for modelling the Siswati sentences had yet to be developed.

Consequently, a bilingual isiZulu-Siswati lexical database was manually developed from the sentences in the treebank. To improve future interoperability with multilingual systems, entries were given English-based function names. The information necessary to derive parallel concrete GF lexicon modules was added for isiZulu and Siswati, such as the relevant root or stem and class information

for nouns. The lexicon size is 190 functions, of which 76 are for nouns (eg. `student_N`) and 89 are for verbs (eg. `come_V`).

Our focus was to develop and evaluate the morphosyntactic functions for a Siswati resource grammar from the existing isiZulu, for which a limited yet representative lexicon is sufficient. An essential resource that must still be developed is a large Siswati computational lexicon.

8. Evaluation

During development, continuous evaluation relied on the manually developed treebank and hence continued until regression tests for both isiZulu and Siswati succeeded. Now, it was time to evaluate the ability of the grammar to generalise to unseen combinations of functions.

While it is possible to use random generation of trees for evaluation, trees generated in such a way are often nonsensical. This limits the value of having them evaluated, and also confronts the evaluator with a difficult task, especially in case of failure: did the grammar render a meaningful tree incorrectly or did it "correctly" render a nonsensical one? Even the task of determining whether a tree represents a meaningful sentence can be difficult, with different kinds and degrees of problematic combinations of functions possibly occurring. False positives may also undermine the evaluation process.

Instead, in order to test the Siswati RG, an augmentation strategy was defined according to which each tree in the manually developed treebank was modified by randomly selecting from a list of possible modifications. These included swapping tense, polarity, number, subject nouns and pronouns. In this way, the same basic linguistic structures were retained in the new test set, but the syntactic context in which they occurred was changed in a guided yet randomised way.

This led to a new set of 125 trees, each with their isiZulu and Siswati linearisations produced by the respective resource grammars. The linearisations were then manually evaluated and errors categorised. Table 1 gives the outcome of the evaluation. Note that in all cases, errors either occurred in both languages or in none, indicating that the bootstrapping itself was entirely successful, i.e. the small percentage of grammatical errors was carried over from the isiZulu RG.

The first thing to note about the results is that inaccurate augmentation occurred for 14 trees (about 11%), often due to unidiomatic or ungrammatical use of lexical items. Making small changes to trees could place words in a syntactic context that was in some way problematic. This highlights that although this kind of augmentation can be very powerful, care has to be taken when designing tree modification rules to limit their application to appropriate contexts.

In three cases, small inaccuracies in the original parallel treebank, originating from the textbooks, were discovered, which we named seed errors. For both the augmentation and seed errors, the grammar still succeeded in producing reasonable, and in most cases morphologically acceptable, linearisations for problematic trees. The number of true grammar errors amounts to less than 2% of the treebank. This is a very encouraging result.

9. Conclusion

We have presented a bootstrapping process to develop a Siswati GF RG from the existing isiZulu RG. To aid in development and evaluation, a set of parallel textbooks was employed, which had themselves been "bootstrapped" due to the similarity of the languages. The parallel texts provided a practical and systemic basis for implementing known differences between the languages, as well as a set of high quality parallel sentences. These were used to develop manual and augmented parallel treebanks, which were utilised during development and evaluation[5].

Our work confirms the feasibility of such bootstrapping approaches for closely related languages. The isiZulu GF resource grammar was developed over a three-year period[6], while the Siswati resource grammar could be developed and evaluated in less than a year[7]. Such reductions in effort and cost are especially important in resource development for under-resourced languages, since their under-resourced status often relates as much to human and financial resources as to language resources.

We intend to explore a number of avenues for continued work. A refined set of tree modification rules could be utilised to further augment the manually developed parallel treebank, which in turn could be converted to a parallel Universal Dependencies treebank (Kolachina and Ranta, 2019) and used to bootstrap UD parsers for both isiZulu and Siswati. This would require the development of improved lexical resources, especially for Siswati. We may look to exploring the possibility of exploiting known orthographic and phonological differences, as discussed in Section 3, to enable this development from existing isiZulu lexical resources, taking care to deal with lexical differences accurately.

[5] https://github.com/LauretteM/gf-bantu-resources

[6] https://shorturl.at/pyUX3

[7] https://github.com/GrammaticalFramework/gf-rgl

Result	Description	Number
Tree error	The new tree is syntactically problematic	6
Lexical error	The new tree uses a word in the wrong syntactic context	8
Seed error	There was a problem with the original sentence	3
Grammar error	The grammar produced an incorrect linearisation	2
Correct	No problem with the new tree or its linearisations	106

Table 1: Summary of evaluation result on the augmented treebank

We also intend to repeat the bootstrapping process for isiXhosa (a relatively large Nguni language with around 8 million L1 speakers) and isiNdebele (a relatively small Nguni language with around 1 million L1 speakers), incorporating the insights gained from developing the Siswati RG. From there, resource grammars for other Southern Bantu languages beyond the Nguni group could be targeted.

We hope in this way to continue to build upon comparative linguistic research to develop digital language resources for the under-resourced languages of South Africa.

10. Bibliographical References

Krasimir Angelov. 2015. Orthography engineering in Grammatical Framework. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 33–40.

Sonja Bosch, Laurette Pretorius, and Axel Fleisch. 2008. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies*, 17(2):23.

Benson Kituku, Wanjiku Nganga, and Lawrence Muchemi. 2021. 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)*, pages 83–88. IEEE.

Prasanth Kolachina and Aarne Ranta. 2019. Bootstrapping ud treebanks for delexicalized parsing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 15–24.

Laurette Marais and Laurette Pretorius. 2023a. Extending the usage of adjectives in the Zulu AfWN. In *Proceedings of the 12th Global Wordnet Conference*, pages 303–314, Donostia, Spain.

Laurette Marais and Laurette Pretorius. 2023b. Parsing IsiZulu Text Using Grammatical Framework. In *Distributed Computing and Artificial Intelligence, Special Sessions I, 20th International Conference*, pages 167–177, Cham. Springer Nature Switzerland.

Sthembiso Mkhwanazi and Laurette Marais. 2024. Generation of segmented isiZulu text. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).

Carmen Moors, Ilana Wilken, Karen Calteaux, and Tebogo Gumede. 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*, SAICSIT '18, page 296–304.

Owen G. Mordaunt, Paul A. Williams, and Z.T. Motsa Madikane. 2023. What sets Siswati apart from isiZulu? *American International Journal of Humanities and Social Science*, 8(1):47–55.

Peter Nabende, David Bamutura, and Peter Ljunglöf. 2020. Towards Computational Resource Grammars for Runyankore and Rukiga. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*.

Finex Ndhlovu. 2022. Pan-African identities and literacies: The orthographic harmonisation debate revisited. *South African Journal of African Languages*, 42(2):207–215.

Aarne Ranta. 2009. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2.

Aarne Ranta. 2011. *Grammatical framework: Programming with multilingual grammars*, volume 173. CSLI Publications.

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract Syntax as Interlingua: Scaling Up the Grammatical Framework from Controlled Languages to Robust Pipelines. *Computational Linguistics*, 46(2):425–486.

P.C. Taljaard and S.E. Bosch. 1988. *Handbook of IsiZulu*. J.L. Van Schaik.

P.C. Taljaard, J.N. Khumalo, and S.E. Bosch. 1991. *Handbook of SiSwati*. J.L. Van Schaik.