SACJ 31(2) December 2019
Research Article

A method for measuring verb similarity for two closely related languages with application to Zulu and Xhosa

Zola Mahlaza, C. Maria Keet

Department of Computer Science, University of Cape Town, South Africa

ABSTRACT
There are limited computational resources for Nguni languages and when improving availability for one of the languages, bootstrapping from a related language's resources may be a cost-saving approach. This requires the ability to quantify similarity between any two closely related languages so as to make informed decisions, of which it is unclear how to measure it. We devised a method for quantifying similarity by adapting four extant similar measures, and present a method of quantifying the ratio of verbs that would need phonological conditioning due to consecutive vowels. The verbs selected are those relevant for weather forecasts for Xhosa and Zulu and newly specified as computational grammar rules. The 52 Xhosa and 49 Zulu rules share 42 rules, supporting informal impressions of their similarity. The morphosyntactic similarity reached 59.5% overall on the adapted Driver-Kroeber metric, with past tense rules only at 99.5%. This similarity score is a result of the variation in terminals mainly for the prefix of the verb.

Keywords: Xhosa, Zulu, similarity measure, phonological conditioning, context free grammar, natural language generation

Categories: • **Natural language processing** ~ **Language resources**

Email:
Zola Mahlaza zmahlaza@cs.uct.ac.za (CORRESPONDING),
C. Maria Keet mkeet@cs.uct.ac.za

Article history:
Received: 4 Apr 2019
Accepted: 16 Oct 2019
Available online: 20 Dec 2019

1 INTRODUCTION

Of the eleven official languages in South Africa (SA), the constitution recognises the "diminished use and status of the indigenous languages" and tasks "the state [with the responsibility of taking] practical and positive measures to elevate the status and advance the use of these languages". Advancing the use of specific languages in the modern age requires computational support. However, the state of South African Human Language Technologies (HLTs), especially for South Africa's indigenous languages, is still very poor (Grover, Van Huyssteen, & Pretorius, 2011; Moors, Wilken, Calteaux, & Gumede, 2018). In particular, there are only a few Natural Language Generation (NLG) systems

Mahlaza, Z. and Keet, C. M. (2019). A method for measuring verb similarity for two closely related languages with application to Zulu and Xhosa. *South African Computer Journal 31*(2), 34–56. https://doi.org/10.18489/sacj.v31i2.698

SACJ is a publication of the South African Institute of Computer Scientists and Information Technologists. ISSN 1015-7999 (print) ISSN 2313-7835 (online).

that generate a South African language. NLG systems generate natural languauge text from various kinds input (e.g. ontology axioms, numerical data, etc.). For instance, there exists a Zulu ontology verbaliser (Keet, Xakaza, & Khumalo, 2017) and a grammar-based Zulu language learning exercise NLG system (Gilbert & Keet, 2018). The former system relies on an extension of traditional templates (namely patterns (Keet & Khumalo, 2017b)) and the former computational grammar rules for surface realisation. An alternative to the two approaches is the use of data-driven methods. However, data-driven methods are not viable due to the lack of large corpora in the languages in question. This work focuses on computational grammar rules as we also want to investigate grammatical similarity in addition to building a surface realisation solution. Computational grammars for Nguni languages are very few as building them is a time-consuming task. In particular, there are some rules that have been reported on but are not “freely available yet" (Griesel & Bosch, 2014) and the only available grammar rules for Bantu languages are limited (Byamugisha, Keet, & DeRenzi, 2016b; Keet & Khumalo, 2017a; Spiegler, van der Spuy, & Flach, 2010). For instance, Byamugisha et al. (2016b) built a small set of rules for a single Part-of-Speech (POS) category, the verb, within the context of generating prescriptions in Runyankore, a Ugandan Bantu language. On the other hand, the other rules (Keet & Khumalo, 2017a; Spiegler et al., 2010) are built with applicability in multiple domains in mind even though Keet and Khumalo (2017a) only consider limited features (i.e. only one tense) for the verb and Spiegler et al. (2010) covers a larger number of POSs at the cost of depth. Nonetheless, none of the existing rules have the features required to generate all verbs used in weather forecasts.

Zulu has more mature resources than all other Nguni languages (Moors et al., 2018). In particular, it has more publicly available computational grammar rule sets than other Nguni languages. So the verb’s grammatical features that are required for weather forecasts could easily added to the existing Zulu grammar rules, and the rules bootstrapped to other Nguni languages. This approach to resource development was used by Bosch, Pretorius, Podile, and Fleisch (2008) when creating morphological analysers for other Nguni languages by using their Zulu morphological analyser, ZulMorph[1]. In order to make informed predictions relating to bootstrapping resources, however, there is a need to be able measure grammatical similarity between any two related languages. More specifically, in this paper we seek to answer the following three questions:

1. *How grammatically similar are Xhosa and Zulu[2] verbs?*
2. *Can a single merged set of grammar rules be used to produce correct verbs for both languages?*
3. *What is the degree of improvement in verb correctness that can be observed after the introduction of conditioning rules that target consecutive vowels on the Context Free Grammars (CFGs)?*

We answer these questions by determining the grammatical features of verbs used in Xhosa weather forecasts, developing a set of computational grammar rules for each of the two languages, adapting binary similarity measures from other domains to measure grammatical similarity on grammar rules, and applying the four chosen measures. A corpus from the weather domain is used here since we are interesting in verb similarity within the larger context of generating well-formed verbs within a bilingual weather NLG system. We also analyse the two sets of grammar using parse trees and

[1]At the time of writing, the rules used the analyser are not publicly available

[2]More correctly isiZulu and isiXhosa but the prefix is omitted for ease of reading by the English speaker.

https://doi.org/10.18489/sacj.v31i2.698

develop a function for each language that takes a set of grammar rules and the number of verb roots with(out) trailing/leading vowels and then calculates the ratio of verbs that would would need phonological conditioning.

We have found that from the resulting 52 Xhosa rules and 49 Zulu rules, there are 42 rules in common. This supports existing intuition on the similarity of the two languages. The morphosyntactic similarity measured with the binary coefficients is 59.5% overall (adapted Driver-Kroeber), with 99.5% for the past tense only rules. The lower score cf. the structure of the CFG is due to the small differences in terminals mainly the prefix component of the verb. The CFGs generate a large number of verbs with consecutive vowels. For instance, when we generate all the possible verbs whose verb root does not have a leading or trailing vowel, we see that Xhosa has 45% of the total number of generated strings have consecutive vowels whereas Zulu has 70%. However, these high ratios should not be interpreted as necessarily meaning that documents whose verbs are generated by the CFGs will be greatly improved by the introduction of phonological conditioning rules. This is because the verbs generated by the CFGs do not have the same probability of being used in text that has some communicative goal. Nonetheless, the phonological conditioning quantification method presented may be a valuable tool in other domains with large corpora available where the probabilities can be estimated. The Xhosa weather corpus and developed rules are available at https://github.com/AdeebNqo/VerbRulesXhZu.

This paper is an extended version of work published in (Mahlaza & Keet, 2018). We have extended that previous work by adding a method and results for quantifying the impact of phonological conditioning that targets consecutive vowels in verbs and the entire set of Xhosa verbs extracted from the corpus. Other additions are the release of the entire Xhosa corpus and grammar rules for the two languages considered in the work. The remainder of this paper is structured as follows: Section 2 introduces the characteristics of Zulu and Xhosa, existing computational methods determining for 'Bantu' language similarity, and the adapted measures used in this study, Section 3 presents the methods and materials for weather corpus collection, CFG development and evaluation, comparison of Xhosa and Zulu, and phonological conditioning rules, Section 4 presents the results, Section 5 presents the discussion, and Section 6 concludes.

2 BACKGROUND

This section provides a brief overview of the characteristics of Xhosa and Zulu, the phonological conditioning process, methods for used to measure document similarity, and binary similarity measures that will be adapted for this investigation.

2.1 South African languages

Official South African indigenous languages can be classified into two groups: Nguni and Sotho-Tswana. All the languages are under-resourced, but to different degrees, and the most recent HLT audit in SA shows that Zulu has the most mature resources (Moors et al., 2018) hence when improving resource availability for the Nguni cluster, resources may be bootstrapped from the existing

Zulu resources. Nguni languages are characterised by complex morphology that is classified as agglutinating (Nurse, 2008) and are mainly spoken in Zone S in the the geographical classification of Bantu languages (Maho, 1999). These languages encode information in the verb that would ordinarily be conveyed via syntactical/lexical means in other languages (Nurse, 2008). To illustrate, consider the following two examples:

(1) *aba-ntu* *ba-za-ku-bhal-a*
2.people 2.SC-IFUT-INF-write$_{VR}$-FV
'people will write'

(2) *aba-bon-an-i*
2.NEGSC-see$_{VR}$-REC-FV
'(they) do not see each other'

The '2' in (1) denotes the *noun class* of the plural noun for 'person', which then requires the subject concord of that noun class to be present in the the verb (the '2.SC') to ensure agreement, and similarly for (2) the verb's subject concord ('2.NEGSC') that is in the negative form agrees with the implied subject that belongs in *noun class* 2. Each noun belongs to a single class in Bantu languages and each class is associated with different kinds of concords (e.g. quantitative, object, etc.). The ones of interest to this work are the subject and object concord morphemes that ensure the verb's agreement with its subject or object. The number of noun classes varies for each language in the Nguni cluster. For instance, Zulu has 17 noun classes and Xhosa 15 (based on Meinhof's 1948 classification).

Other morphemes include the immediate future tense *-za-* and the infinitive *-ku-* in example (1), and the reciprocative *-an-* and negative final vowel *-i* in example (2). The verb largely follows a fixed slot system (Khumalo, 2007; Maho, 1999) and a simplified example is given in Figure 1. The verb varies in form in to capture agreement with other parts of speech such as the noun and to indicate its aspect, mood, and tense. Its suffix is made up of verbal 'extensions' (Keet & Khumalo, 2017a) and the final vowel. Verb extensions are used to denote that an action is passive, reciprocative, etc.

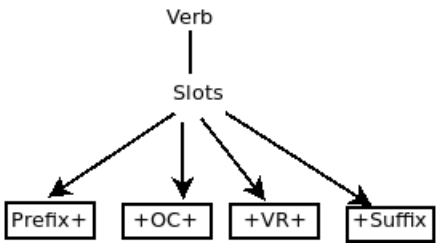

Figure 1: A (simplified) example of the verb slot system. Prefix is a slot that contains a number of elements such as the subject concord; OC is the object concord, VR is the verb root, and Suffix is slot that contains the verb extensions (e.g. the passive) and the final vowel. A plus symbol (+) on a slot indicates that another slot can be attached on that side

When words are being formed by appending morphemes together in the two languages, certain processes take effect and change the form of certain morphemes. These rules are either of morphological, phonological, or lexical extract and they ensure that the correct version of an allomorph is used in the correct contexts. These rules are important because Zulu and Xhosa do not permit consecutive

vowels in verbs, among other things. The rules responsible for eliminating consecutive vowels are needed especially for vowel-commencing verb roots. For instance, consonantalisation, a process whereby the vowels **i** and **u** become **y** and **w** respectively, can occur in some Nguni languages as follows *u* + *enza* > *wenza* (s/he does). The rules that govern the combination of vowels vary for each of the processes. Some processes require morphological features, some require phonological features, and other operate only on the values of the vowels. More information about consonantalisation process and other 9 other processes can be found in Msimang (Msimang, 1989). In this work, we deem the rules for conditioning orthogonal to the rules for the grammatical constituents of the verb.

2.2 Language similarity

There are multiple ways one can measure similarity for any two documents of a specific language. In particular, one can measure syntactic, lexical (or phrasal), and semantic similarity through some function whose range is between zero and one where zero means completely dissimilar and one completely the same. These functions need not make use the text directly as they can also use the typographical features as evidenced by Georgi, Xia, and Lewis (2010)'s work. Semantic and lexical similarity can be calculated through a variety of techniques such as Spearman's correlation coefficient, Resnik (Huffaker, Jorgensen, Iacobelli, Tepper, & Cassell, 2006; Warin & Volk, 2004), Leacock-Chodorow, latent semantic analysis, or custom metrics (e.g. (Indukuri, Ambekar, & Sureka, 2007)). Syntactic similarity has been compared at the sentence level through the use of w-shinglings (Broder, 1997; Pereira Jr. & Ziviani, 2003). however, comparisons of lexical, syntactic, or semantic similarity yield limited information pertaining to how to bootstrap resources.

The only work that compares Bantu langauges using computational means , to the best of our knowledge, makes use of lexical and orthographical features (Chavula & Suleman, 2017; Keet, 2016). In particular, the work is interested in similarity regarding the stem of words (Chavula & Suleman, 2017) or the distribution of word lengths (Keet, 2016). Both works applies to all words in a language and not only to verbs. Chavula and Suleman (2017)'s goal is cluster induction hence their work can only produce a list of similar words and it does not provide a singular metric for measuring similarity for any two languages. Keet (2016)'s work provides word-level insights by using the word length distributions on the Universal Declaration of Human Rights (UDHR) document, however, they also do not produce a single metric for comparing any two languages. For informing bootstrapping, grammatical similarity is key and both works (Chavula & Suleman, 2017; Keet, 2016) do not take into account the grammar of the measured languages in determining similarity.

The limitation of word length and stem-based similarity comparisons and the need for a method that also incorporates grammar is clear when one considers that Sesotho and Setswana are not similar with respect to word length (Keet, 2016) despite belonging to the Sotho-Tswana language group and being mutually intelligible. Moreover, since Chavula and Suleman (2017)'s metric disregards grammar and assumes a normal distribution when calculating the weights, it is likely to miss the similarity that exists between words that do not have roots of similar length. The next section, therefore, takes a step back in order to consider several similarity measures that have been used in other domains.

https://doi.org/10.18489/sacj.v31i2.698

2.3 Similarity measures

Binary similarity measures are functions that measure containtment or resemblence of two items (Broder, 1997). Technically, they can be used used in any subject area where the individuals being compared can be represented as some structured set whose features are binary (that is, absent/present). This representation is similar to how Componential Analysis (CA) represents the meaning of natural language words. For instance, according to Goddard (2011), in CA the following words can be represented with two features: *man* [-female, +adult], *boy* [-female, -adult], *woman* [+female, +adult], *girl* [+female, -adult] where symbol (-) represents absence and (+) represents presence. These word representations can be compared with some binary measure $f(w_i, w_j) = d$ where $0 \leq d \leq 1$. To date, binary measures have been used in a variety of areas, from detecting image manipulation (Bayram, Avcibas, Sankur, & Memon, 2005) to botany (Rogers, Tanimoto, et al., 1960).

There are numerous measures; Cheetham and Hazel (1969) describe 20, Todeschini et al. (2012) list 51 measures for chemoinformatics data, and Choi, Cha, and Tappert (2010) collect and analyse 76 measures. We have selected four measures that we will adapt and make suitable for comparing language morphosyntax. These are the Driver-Kroeber 1932, Sorenson 1945, Jaccard 1912, and Sorgenfrei 1959[3]. They are all asymetric and are chosen because they are well documented. The task of determining the nature of the other metrics discussed by Choi et al. (2010), the largest collection of the three, was not fruitful. In order to understand these metrics, we first introduce the association index (Dice, 1945). It was created and used in ecology to study the association of one species (Θ_1) to another (Θ_2) in a region. In particular, the association of Θ_1 to Θ_2 is the ratio of the size of the region in which both species are found to the size of the region in which Θ_1 only is found. The association index is not commutative as the size of the area in which Θ_1 is found is not necessarily the same as that of Θ_2. Therefore, the four selected measures can be understood as different methods why the index is transformed into a commutative measure. Sorenson-Dice is the ratio of the sizes of the shared spaces[4] to the total number of species in both sets ($SD(\Theta_1, \Theta_2) = \frac{2|X \cap Y|}{|X|+|Y|}$). Jaccard's metric is the ratio of shared items to the total number of items that exist in two sets ($J(\Theta_1, \Theta_2) = \frac{|X \cap Y|}{|X|+|Y|}$). The Driver-Kroeber measure the weight of the shared space similar to Dice (1945) but merges two indices by determing the geometrical mean ($DK(\Theta_1, \Theta_2) = \frac{|X \cap Y|}{\sqrt{|X||Y|}}$). The Sorgenfrei metric is the multiplication of association indices ($Sorg(\Theta_1, \Theta_2) = \frac{|X \cap Y|^2}{|X||Y|}$). Sorenson, Sorgenfrei, and Jaccard were developed to study associations of between species in a region. In particular, Jaccard's measure was for studying the distribution of flora in the Alps. The Driver-Krober was used in ethnology to measure the similarities of cultural traits between two groups of peoples.

[3] as cited by Todeschini et al. (2012)

[4] Technically, there is one shared space but it is counted twice.

https://doi.org/10.18489/sacj.v31i2.698

3 SIMILARITY ASSESSMENT AND PHONOLOGICAL CONDITIONING

In order to answer the posed questions, we determine the grammatical features of weather forecast verbs and develop computational grammar rule sets capable of generating those verbs. This section presents how the corpus was collected, analysed, and used to develop the grammar rules. The section also presents the methods and materials used to compare the developed rules and quantify the ratio of verbs with consecutive vowels.

3.1 Materials and Methods

Section 3.1.1 details the method used to create the corpus and analyse its verbs, Section 3.1.2 presents the evaluation of the developed CFGs, the method by which the adapted measure are applied to the rules is detailed in Section 3.1.3, Section 3.1.4 presents the method for comparing the two languages, and Section 3.1.5 presents the method used to quantify phonological conditioning that targets consecutive vowels.

3.1.1 Corpus collection and CFG development

We decided to formalise a subset of the verb that will be of practical use since formalising the complete verb is difficult due to the complexity of the verb and the grammar is neither well documented nor fully studied. In particular, we have chosen to focus on verb features that are suitable for generating verbs for weather forecasts.

These features were extracted from an Xhosa weather forecast corpus that was attained by translating an English corpus obtained from the South African Weather Service (SAWS)[5]. The corpus was translated into Xhosa by an individual from the African languages and literature section of the School of languages and literatures at the University of Cape Town. The verbs from the Xhosa corpus were identified manually and their features studied. These features were used together with grammar literature (C. M. Doke, 1931, 1992; Grout, 1859; McLaren, 1944, 1955; Peters, 1966; Taljaard & Bosch, 1988) to design CFGs. In particular, the literature was used to gather specific information pertaining to tense, aspect, and mood.

The verb's components we consider are the prefix, object concord (OC), verb root (VR), verb extensions (VE) and final vowel (FV). The rules, for both languages, are developed incrementally and the steps per incremental stage are given below. The verb root and pre-final vowel suffix are relatively easy and are therefore not separate incremental steps.

- Increment 0: Prefix
  1. Gathering preliminary rules
  2. Verb generation, correctness classification, and elimination of incorrect verbs.
- Increment 1: Prefix + OC + VR + VE
  1. Suffix addition, verb generation and correctness classification
  2. Elimination of incorrect verbs, verb generation and correctness classification

[5] http://www.weathersa.co.za/

- Increment 2: Complete verbs
  1. Investigate missing features, add missing features (where necessary), add final vowel, correctness classification. This step is added since the removal of incorrect verbs may unintentionally remove some features from a set in the previous increment.
  2. Elimination of incorrect verbs, verb generation and correctness classification

3.1.2 Verb rule evaluation

The verb rule's quality between increments was done by one of the authors, who is an L1 speaker of Xhosa. The evaluation was conducted by generating strings from a fixed root using Python and Natural Language Toolkit (NLTK) and then counting the number of correct and incorrect strings. The generated verbs used the verb root (*-zol-*) as it was randomly selected from the list verbs extracted from the translated SAWS forecast text. As a starting point, the subject concord *li* and a empty object concord were chosen.

The quality of the rules after the development was also evaluated by linguists. Twenty-five verbs that exist in both Zulu and Xhosa were extracted from an English-Zulu dictionary (C. Doke, Malcolm, Sikakana, & Vilakazi, 1990) in alphabetical order from the a-commencing to e-commencing words sections (five verbs from each of the five sections) and five pairs of subject and object concords were randomly selected.

The 25 roots together with the *-zol-* root were paired with the five concords pairs. The root (*-zol-*) is re-used from the previous random verb root selection. The pairings of concords and root were inserted into the rules to generate strings in Zulu and Xhosa ($n = 49400$). Ninety nine[6] strings are randomly sampled from each language set using the Python random module, put into a spreadsheet, and sent to two linguists for evaluation. The linguists were instructed to annotate each word with True/False for syntactic correctness and for semantic correctness, and add a comment if they have one.

3.1.3 Similarity measures assessment

The measures, as presented in Section 2.3, are adapted to make them suitable for languages. In particular, we considered the subset of the languages under consideration as being represented by a vector whose binary features are natural languages words. For instance, using the three fixed order words *dog, cat, inja* we can represent the English as the vector [1,1,0] and Zulu as [0,0,1] where 0 at position n means that the *nth* word from the three is absent and 1 present. This kind of representation allows us to define, for any two languages A,B and set of words, the variables $a = |A \cap B|$ being the shared words, $b = |B - A|$ words only found in B, and $c = |A - B|$ being words found only in A. We then substituted these variables into the four functions from Section 2.3 in order to obtain the functions below. Effectively, any instance of $|X \cap Y|$ in the original measures listed

[6] 100 verbs were sampled and a single verb was mistakenly discarded from the Xhosa and Zulu lists.

https://doi.org/10.18489/sacj.v31i2.698

above is substituted with a, $|X|$ with $a+b$ and $|Y|$ with $a+c$.

$$J(A,B) = \frac{a}{a+b+c} \tag{1}$$

$$S(A,B) = \frac{2a}{2a+b+c} \tag{2}$$

$$DK(A,B) = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{3}$$

$$Sorg(A,B) = \frac{a^2}{(a+b)(a+c)} \tag{4}$$

This meant that what the four original measures considered region/culture is now language (or a subset of it) and what was considered to be species/cultural traits are now words that either present or abstract the languages being compared.

The above functions have differences in their formulation, however, one may be able to obtain the value of one function from the other by applying some conversion ratio. We investigated the relationship between the functions by generating 1024 cases of triples (a,b,c) using the Numpy[7] discrete uniform distribution random integer generator with the constraint $a+b+c=1024$ that were then fed into each function. We then calculated the difference between the functions for each of the 1024 triples and checked the ratios obtained against the values we obtain with the four measures.

3.1.4 Xhosa and Zulu comparison

The reformulation from the previous section can be used to measure more than lexical divergence by exploiting obvious similarities between the two languages. In particular, since the two languages share some verbs (e.g. *hamba*) then morphosyntax can be measured by generating sets of verb using a fixed shared verb root and concords, and then feeding the sets as input to the functions from the previous section. This was done by randomly sampling a verb root (*-zol-* meaning 'calm') from the list of extracted verbs from the Xhosa corpus. As a starting point, we paired the verb root with a subject concord (*li-*) (for noun class 5) and an empty object concord. These values were inserted into the two CFGs. In order to investigate the morphosyntactical difference between languages for each tense, we split the rules into four clusters: (1) the complete set of rules, (2) present tense rules only, (3) all verb rules excluding present tense rules, and (4) past tense rules only. The results sets of verbs for each cluster that were generated using Python and NLTK were then fed into the similarity measures (Eqs. 1–4).

The verb root and concords used in the investigation above are fixed hence it is possible that a different verb root and concords may provide different values. In order to determine whether the morphosyntactical similarity measured is not different when the verb root and concords are changed; we collected different verb roots and concords, inserted them into the CFGs, generated new sets of

[7] http://www.numpy.org/

Table 1: Four types of verb roots in Zulu and Xhosa. The type size refers to the number of verb roots that belong to each type.

Verb root type	Description	Size
A	Verb roots that begin with a vowel	α
B	Verb roots that end with a vowel	β
C	Verb roots that start and end with a vowel	γ
D	Verbs roots that do not start or end with vowel	δ

verbs, and applied the measures on the resulting sets. Specifically, regarding the selection of the verb roots, we re-used the verb roots extracted from a dictionary as detailed in Section 3.1.2. Each of these roots was paired with the *li-* subject and empty object concord when generating verb sets for comparison. Regarding the selection of the concords to investigate; we randomly selected 5 concord pairs from the set of concords that exist in both languages. The selected concords are (*a, zi*), (*i, wa*), (*i, yi*), (*lu, bu*), and (*u, yi*) where the first item in each tuple is the subject concord and the other the object concord. Each of the concord pairs was paired with the *-zol-* verb root, and each pairing was inserted only into the complete set of rules (rule cluster 1) to generate verb sets that will be compared.

3.1.5 Phonological conditioning quantification

In investigating the impact of phonological conditioning, we quantify the degree to which there are consecutive vowels that need to be conditioned in both languages. In particular, a set of semi-abstract verbs was created by substituting abstract variables for the subject concord (X), object concord (Y), and the root (Z) in the two grammar rules. For each of the three variables, we consider all the the possible types of the three variables based on whether they have a trailing or leading vowel. For the subject (X) and object concord (Y), in Xhosa and Zulu, there are there are 8 and 6 possible combinations with regards to starting or ending with a vowel as shown in Table 2. There are four types of verb roots for the two languages and these are shown in Table 1.

For each subject/object concord pair, we create a formula to count the number of verbs with consecutive vowels based on the four kinds of verb roots and these formulae are included in Table 2. This is achieved by populating the values for the concords in the semi-abstract verbs for each concord pair and counting the number of instances where vowel-targeting phonological rules will be required for each type of verb root for all the possible concord pair in each language. For instance, in Xhosa's concord pair category 3, there are 491 instances where verb roots that belong in A, B, and C may need phonological conditioning and there are 232 instances for verb roots that belong in class D hence $f_3(\alpha,\beta,\gamma,\delta) = 491(\alpha+\beta+\gamma)+232\delta$ where the function's arguments denote the number of each verb root type as defined in Table 1. We then combine all the functions of each language and created a single function for each language to produce the total ratio of verbs that would need

https://doi.org/10.18489/sacj.v31i2.698

Table 2: Information about the six Zulu and eight Xhosa concord pairs. Example pairs are provided for each type of concord pair. Functions that calculate the number verbs that have consecutive vowels for each concord pair are listed in the “Affected verb function" column. The number of semi-abstract verbs for each concord pair type are shown under the rightmost column. The ϵ symbol represents the empty string.

	Subj or Obj pair I.D.	**Subj type**	**Obj type**	**Example pair**	**Affected verb function**	**Semi-abstract verb count**
Zulu	1	Ends with vowel	Ends with vowel	ngi, ngi	$439(\alpha+\beta+\gamma)+243\delta$	439
	2	Ends with vowel	No vowel	ngi, m	$439\beta+243\delta$	439
	3	Ends with vowel	Empty	ngi, ϵ	$439(\alpha+\beta+\gamma)+243\delta$	439
	4	Vowel Only	Ends with vowel	u, ngi	$439(\alpha+\beta+\gamma)+368\delta$	439
	5	Vowel Only	No vowel	u, m	$439\beta+368\delta$	439
	6	Vowel Only	Empty	u, ϵ	$439(\alpha+\beta+\gamma)+368\delta$	439
Xhosa	1	Vowel Only	Empty	u, ϵ	$491(\alpha+\beta+\gamma)+232\delta$	491
	2	Vowel Only	Vowel only	u, e	$491(\alpha+\beta+\gamma+\delta)$	491
	3	Vowel Only	Ends with vowel	u, ndi	$491(\alpha+\beta+\gamma)+232\delta$	491
	4	Vowel Only	No vowel	u, m	$491\beta+232\delta$	491
	5	Ends with vowel	No vowel	ndi, m	$491\beta+24\delta$	491
	6	Ends with vowel	Ends with vowel	ndi, ndi	$491(\alpha+\beta+\gamma)+24\delta$	491
	7	Ends with vowel	Vowel only	ndi, e	$491(\alpha+\beta+\gamma+\delta)$	491
	8	Ends with vowel	Empty	ndi, ϵ	$491(\alpha+\beta+\gamma)+24\delta$	491

https://doi.org/10.18489/sacj.v31i2.698

phonological conditioning regarding verbs. These functions are given below:

$$RCV_{xh}(\alpha,\beta,\gamma,\delta) = \frac{\sum_{n=1}^{8} f_i(\alpha,\beta,\gamma,\delta)}{\sum_{n=1}^{8} c_i \times (\alpha+\beta+\gamma+\delta)} \times 100 \quad (5)$$

$$RCV_{zu}(\alpha,\beta,\gamma,\delta) = \frac{\sum_{n=1}^{6} f_i(\alpha,\beta,\gamma,\delta)}{\sum_{n=1}^{6} c_i \times (\alpha+\beta+\gamma+\delta)} \times 100 \quad (6)$$

We then use these two functions for each type of verb root that is listed in Table 1 to calculate the ratio of verbs that have consecutive vowels among all the strings that can be generated by the two developed CFGs. Furthermore, since Zulu and Xhosa verb roots are not uniformly distributed among the four types in any given text (i.e. the values for $\alpha,\beta,\gamma,\delta$ are unlikely to be equal), we approximate a fair distribution by analysing the Xhosa[8] and Zulu[9] verb roots distributed with the National Centre for Human Language Technology (NCHLT) lemmatisers (Eiselen & Puttkammer, 2014). The data is made up of 4354 Xhosa and 5839 Zulu verb roots extracted from a corpus that was “sourced from South African government websites and documents, with some smaller sets of news articles, scientific articles, magazine articles and prose" Eiselen and Puttkammer, 2014, p. 3699. In the data, we see that out of a total of 5839 Zulu verb roots, 5624 roots do not start or end in a vowel (Type D) and the remaining 215 verbs only start with a vowel (Type A). Furthermore, out of a total of 4354 Xhosa verb roots, 4233 roots do not start or end in a vowel (Type D) and the remaining 121 verbs only start with a vowel (Type A). This distribution serves as a basis when determining the ratio of verb root types in weather forecast documents. In particular, we assume that for each verb root of Type A there will be twenty verb roots of Type D in an Zulu weather forecast and thirty type D verb roots for each type A in Xhosa. These ratios are used to quantify the ratio of of verbs with consecutive vowels in documents that have between 0-500 verbs ($0 < a + d \leq 500$ and $b = c = 0$) where a verb root of type A is always used. This is done by calculating all the possible values for $\alpha,\beta,\gamma,\delta$ under the constraints and applying the Functions 5-6.

4 RESULTS

We present the results in the order of the methods described in the previous section.

4.1 Weather verbs

We first assessed informally the translations into Xhosa, which showed that the translator did indeed translate the meaning of the forecasts, rather than a literal translation of each sentences. For instance, the phrase “fine and warm" was translated into “Lihle kwaye litshisa": the Xhosa has a subject concord added to *tshisa* (to make *litshisa*), which indicates an implied subject. Such grammatically correct insertions enables examining the verbs in their correct forms. The Xhosa translations of the

[8]https://hdl.handle.net/20.500.12185/310
[9]https://hdl.handle.net/20.500.12185/317

12 SAWS weather reports contained 53 verbs, of which 27 were unique strings; see Table 3. Assessing their mood, 22 verbs were in the indicative, 2 in the participial, and 3 in the subjunctive. Recalling that Xhosa has 'verb extensions' (see Section 2.1), the ones used in the weather forecast translations are the perfect, causative, neuter, and reciprocity[10]. To this, the intensive form is added, because the collected corpus contains the ultraviolet (UV) index, which may also need to be communicated as intensity of UV radiation. The list of extracted verbs is given in Table 3.

4.2 CFG development and comparison

We commenced with the development of the CFG with two verbal aspects, begin the progressive and exclusive, and three verb tenses, being the past, present, and future. This resulted in 49 rules for Xhosa and 52 for Zulu, with a breakdown as listed in Table 4, where the production rules are partitioned into 1) terminal productions, 2) those that encode exclusive-morpheme-use only, 3) those that encode exclusive-morpheme-use and morphotactics, and 4) those that encode morphotactics only.

The intersection of the Xhosa and Zulu rules, being the number of rules that are the same in the two languages, is large with 42, or about 80%; the potentially interesting rules that differ are included in Figure 2. One main difference regarding the productions with non-terminals only, is that Xhosa, unlike Zulu, has three rules to encode exclusivity: one for present tense and non-present continuity (encoded as $C \rightarrow PC|NPC$), for the progressive aspect and the remote past (encoded as $PRP \rightarrow P|PR$), and for the progressive aspect and the present continuity morpheme (encoded as $PCP \rightarrow P|PC$). Also, the Xhosa present indicative and participial rules differ, besides having an additional rule (see rule x1 in Figure 2). 1) Xhosa uses a fixed present continuity (variable PC in rule x0), which may be empty in Zulu (denoted with PC_1 in rule z0), and 2) the Zulu prefix incorporates the exclusive aspect. Overall, though, they are minor differences, as the underlying structure of the two rules is the same (see also Figure 3) and, effectively, z0 is a 'super-rule' when one considered the variables only.

Obviously, there are a few differences in the terminals. The one difference that affects further processing of verbs most (discussed below), is that Zulu's infinitive is `uku` and it has a special imperative present tense morpheme `ma` and on other hand, Xhosa does not have a special imperative present tense morpheme and its infinitive is `ku`.

There are minor differences in the suffix of the subjective mood (see rules x2 and z1 in Figure 2): Zulu requires only the perfect (S_p) whereas Xhosa also requires the neuter extension (S_{np}). This difference keeps the overall structure of both rules intact, however, given that it amounts to only one (mutually exclusive) morpheme with all else being equivalent. Further, observe that rules x3 and z2 differ in their prefixes: there is the simple, exclusive, and progressive aspects in Xhosa whereas there is only the mandatory simple aspect in Zulu. Finally, the general prefix for the present subjunctive mood differs, because Xhosa includes continuity but Zulu does not (x4 cf. z3 in Figure 2).

In sum, the manual and, hence, any parse tree, analysis of the CFGs demonstrated a high degree of

[10]henceforth, we exclude reciprocity, because it is used in verbs that refer to a geospatial description that is not available.

Table 3: Full list of the verbs extracted from the Xhosa weather corpus.

Xhosa string	**Root**	**English description**
ezimelelene	*-m-*	indicative 'facing each other'
ilindelekile	*-lind-*	indicative 'expected'
kulindeleke	*-lind-*	indicative 'expected'
kuphole	*-phol-*	subjunctive 'cool/chill'
kuyakubakho	*-kh-*	indicative 'the will be'
kuyakuthi	*-th-*	indicative 'it will be/do'
libanda	*-band-*	subjunctive 'it is cold'
libeneziphango	*-b-*	participial 'there will have storms'
lipholile	*-phol-*	indicative 'it is calm/cool'
lithi	*-th-*	indicative 'it will be/do'
litshisa	*-tshis-*	indicative 'it is hot'
litshise	*-tshis-*	participial 'it was hot'
liyakuthi	*-th-*	indicative 'it will become'
liyakutshisa	*-tshis-*	indicative 'it will be hot'
lizakuthi	*-th-*	indicative 'it will become'
lizolile	*-zol-*	indicative 'it is calm'
ovela	*-vel-*	indicative 'comes'
oyakuye	*-kuy-*	indicative 'will go'
uhlaziya	*-hlaziy-*	indicative 'renews'
ukusuka	*-suk-*	indicative 'starting'
ukuya	*-y-*	indicative 'goes'
usiba	*-b-*	subjunctive 'becomes'
uyakuphola	*-phol-*	indicative 'calm/cool'
uye	*-y-*	indicative 'goes'
zilindelekile	*-lind-*	indicative 'expected'
ziyakulindeleka	*-lind-*	indicative 'will be expected'
ziyakuthi	*-th-*	indicative 'will become'

Table 4: Aggregate number of production rules of Zulu and Xhosa CFGs, respectively, and and intersection size (i.e., number of rules that are the same).

	Total	**Terminal**	**Exclusive**	**Exclusive & Morphotactics**	**Morphotactics**
Zulu	49	13	6	8	22
Xhosa	52	12	9	8	23
Intersection	42	11	6	8	17

https://doi.org/10.18489/sacj.v31i2.698

Xhosa
Indicative & Participial
(x0.) $Verb \rightarrow SC \quad PC \quad OC \quad VR \quad S_{np}$
(x1.) $Verb \rightarrow A_{pe} \quad OC \quad VR \quad S_{np} \quad \texttt{a}$
Subjunctive
(x2.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np}$
(x3.) $Verb \rightarrow A_{pes} \quad OC \quad VR \quad S_{np} \quad \texttt{a}$
(x4.) $Prefix \rightarrow A_{es} \quad PC_1$

Zulu
Indicative & Participial
(z0.) $Verb \rightarrow A_{es} \quad PC_1 \quad OC \quad VR \quad S_{np}$

Subjunctive
(z1.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_p$
(z2.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np} \quad \texttt{a}$
(z3.) $Prefix \rightarrow SI \quad SC \quad | \quad SC$

Figure 2: CFG production rules that differ between Xhosa's and Zulu's present tenses. SC: subject concord; PC/PC_1: present continuous; OC: object concord; VR: verb root; $A_{pe}/A_{es}/A_{pes}$: progressive, exclusive, and simple aspect (used exclusively); S_{np}/S_p: neuter and perfect verb extensions; SI: present imperative morpheme.

similarity of Xhosa and Zulu at the variable-level. This formally and precisely confirms the—hitherto informal—perceptions of similarity of the two languages.

4.3 Linguist grammar evaluation

The outcome of the evaluation by the linguists is shown in Table 5. Of the 99 Zulu verbs, 30 were only partially annotated with True/False for the syntactical and semantic correctness fields; therefore the correctness percentage has been calculated over 69. The Xhosa syntactic and semantic correctness were 52% and 58%, respectively, and for Zulu they were 23% and 25%, respectively. The quality difference is at least in part due to the leniency of the Xhosa linguist unlike their Zulu counterpart, which can be observed from the number of Xhosa strings that were considered as being semantically correct despite being syntactically incorrect. Xhosa has 25 strings annotated as such while Zulu has only 1 (*isazahluka* was deemed syntactically wrong due to the *za* cf. *zo*). Another reason is that the Zulu strings require more phonological conditioning, which was not resolved at this stage of the evaluation, hence there was more noise in the strings. This is the case in particular for the aforementioned difference in *uku-* vs *ku-* for immediate future tense.

Comparing the two languages using the verbs that are properly annotated by the linguists, shows that the syntactic and semantic quality between the two languages differs substantially. In particular, there is a significant statistical association between the syntactic (two-tailed p=0.0001, Fisher's exact test) and language. The same holds for semantic correctness and language (two-tailed p=0.0023, Fisher's exact test).

https://doi.org/10.18489/sacj.v31i2.698

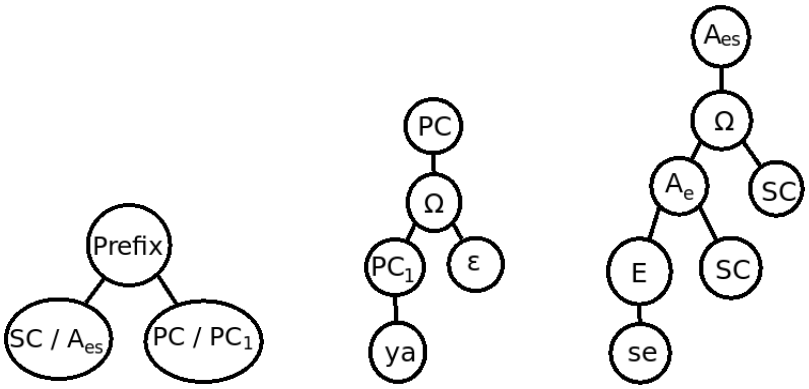

Figure 3: Tree representations of the Xhosa and Zulu's indicative and participial moods prefix (rule 0 in Figure 2). Left: the combined Xhosa/Zulu (Xh/Zu) prefix rule; middle: Xhosa-only optional present continuity; right: Zulu-only presence of the exclusive aspect. The Ω node represents mutual exclusiveness for its subtrees.

Table 5: Summary of the linguists' semantic and syntactic correctness evaluation of the Xhosa and Zulu generated strings.

		Pct. correct	Correct	Incorrect	Total
Xhosa	**Syn.**	52%	51	48	99
	Sem.	58%	57	42	
Zulu	**Syn.**	23%	16	53	69
	Sem.	25%	17	52	

4.4 Similarity measures for Zulu and Xhosa

Having demonstrated grammatical similarity by means of the CFGs comparison, we now proceed to the quantitative results of the strings generated with the grammars, using the adapted binary similarity measures (Funcs. 1-4). The first assessment with *-zol-* generated 504 unique strings the second one with the 25 shared verb roots generated and 12600 strings. The computed similarity measures are provided in Table 6.

Table 6: Binary similarity measure values (rounded) for the verb sets generated by the respective fragment of the CFG, using *-zol-*.

Rule Cluster	Sorg	J	DK	S
Complete	0.354	0.423	**0.595**	**0.595**
Present tense	0.376	0.435	**0.613**	0.606
Past and future	0.341	0.412	**0.584**	**0.584**
Past tense	0.990	0.990	**0.995**	**0.995**

The results adhere to an ordering of $0 \leq Sorg(A,B) \leq J(A,B) \leq S(A,B) \leq DK(A,B) \leq 1$, for all binary vectors A, B (Todeschini et al., 2012; Warrens, 2008). They also show that there is only a

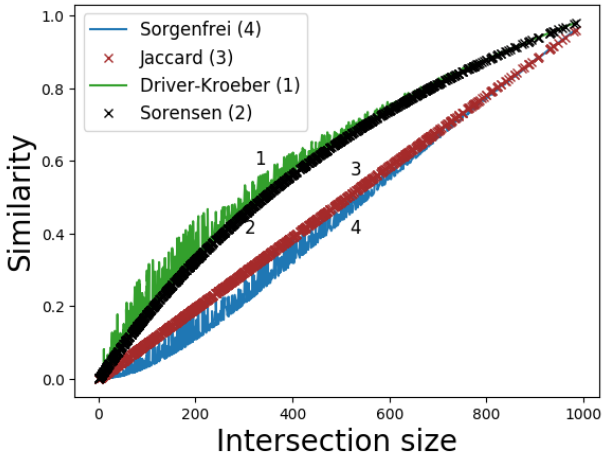

Figure 4: Difference in four binary similarity methods when the size of the intersection between two sets increases and the sets' complement decreases. Similarity is measured with value between zero (Different) and one (Equivalent).

small difference between the measured similarities for each tense cluster (past, present, and future) within each measure. The past tense cluster has the highest similarity values of the three, on all measure.

Turning to the results for the set with different subject and object concords paired with *-zol-* and modifying the verb roots, we note that they are the same as for *-zol-* with the fixed subject concord (table omitted; see Table 6). This demonstrates that the actual vocabulary and slots (as selected in the similarity assessment) thus had no effect on the similarity measure.

The Sorgenfrei and Jaccard (Sorg $\approx$ J) are mathematically similar, as are the Sorenson index and the Driver-Kroeber measure (S $\approx$ DK) (Eqs. 1-4). Their similarity also holds in our data (Table 6): the values of the former differ by at most 0.071 and the values of the latter pair differ by at most 0.007.

4.5 Similarity measure behaviour

In other to make the similarity measures and their values more meaningful for language comparisons, be this for the language spaces of the Xhosa and Zulu strings generated or any other pair of languages, we now proceed to the results of the measures' behaviour.

The differences between the Sorenson and Sorgenfrei measure are 0.250 (max.) and 0.166 (avg.), which is higher than the differences between the Driver-Kroeber and Jaccard ones (0.247 and 0.142, respectively). The latter two are more intuitive for language 'spaces' (words in the language), because they take into account that the two sets that are being compared may have different sizes and factors that in into the calculation. That said, it is possible to convert one into the other, so that some comparison will be possible anyhow. To be able to do so, we examine in detail the relationship between the measures, using the 1024 (a, b, c) triples. The results are shown in Figure 4. The

differences vary by the size of the sets, yet, a value of the Jaccard measure can be computed from a Sorgenfrei one by adding about 0.04, a representative value for Driver-Kroeber from the Jaccard by adding about 0.14, and an approximation of the Sorensen from the Driver-Kroeber by subtracting about 0.02. Note that the behaviour of the metrics, as clearly illustrated here, functions in the exact same way independent of any differences in tense of the rules. This can be cross-checked with the values we obtained (Table 6): e.g., the difference between Jaccard's and Driver-Kroeber is 0.595-0.423=0.172, which is somewhat higher than average, as the intersection size is substantial. Likewise, the difference between the Sorenson and Sorgenfrei (0.595 - 0.354 = 0.241) is slightly higher than the average for the same reason. One can thus conclude that even though the actual similarity measure values differ, they are *de facto* interchangeable when the conversion factor is applied. This also hold for the word spaces of the words generated by the Xhosa and Zulu CFGs.

4.6 Phonological conditioning

The count of verbs that have at least one consecutive vowel in each language when we use a verb root to each of the four categories listed in Table 1 is shown in Table 7. Moreover, when we assume that the ratios of the verb types in weather forecasts is similar to the ratio of verb root types in the NCHLT corpus as detailed in Section 3.1.5, then we see that 69% Zulu and between 46%-58% for Xhosa verbs generated by the CFGs have consecutive vowels.

Table 7: Ratio of verbs that may need phonological conditioning out of the total number of generated verbs for each verb root type. Root types are define in Table 1. Ratios are calculated using formulae 5 and 6.

Language	**A**	**B**	**C**	**D**
Xhosa	75	100	75	45
Zulu	67	100	67	70

5 DISCUSSION

While the quality of the developed rules is mixed, it must be noted that they are the first verb CFGs to include other tenses other than present tense and they explicitly do not cover phonological conditioning. Also, the linguists' evaluation is based on a small random sample; hence, may be biased towards the incorrect strings, whereas the evaluation conducted by the authors used the entire generated set tailored to the weather domain. Also, the linguists' evaluation is preliminary, and a deep analysis also reveals issues typical of under-researched languages. For instance, two strings generated with the Xhosa grammar (*seiyazacebeka* and *seiyazahlulisa*) were marked as not being Xhosa, and were identified as having a high likelihood of being Zulu, despite their roots existing in Xhosa (*ceba* Nabe, Dreyer, and Kakana, 1976, p 388 and *hlula* Nabe et al., 1976, p 445 can be found in an Xhosa dictionary). Likewise in Zulu, the strings *isazadibanisisa* and *beizadibanekisa* were annotated as being Xhosa and not Zulu despite their root existing in Zulu (*dibanisa* C. Doke et al.,

1990, p 144 can be found in an Zulu dictionary). They may have been classified as incorrect due to morphotactics, which is not encoded in the grammar, or dialect use and is an issue for linguists to investigate. Certain vowel combinations (*-ei-*) may also reveal an inadequacy in phonological conditioning documentation since its not found among the rules documented in C. M. Doke (1992) and Gowlett (2014).

We now go back to the research questions that were posed in Section 1. Regarding the first research question, *How grammatically similar are Xhosa and Zulu verbs?*: our results reveal that the most intuitive similarity measures are Jaccard and Driver-Kroeber. These two metrics gave 42% (Jaccard) and 59.5% (Driver-Kroeber) similarity for the verb fragment that was under investigation. We also observe that the difference between the two measures is close to the average difference between them (recall Section 4.5). The minuscule difference (0.175) is a result of difference in formulation of the two functions and not a result of differences in the two languages. In other words, these metrics are in effect interchangeable.

It should be noted, however, that the similarity value may be higher or lower for fragments that consider additional or different features, and we caution against a universalization of the results at this stage. The developed CFGs overgenerate, and the quality of the Xhosa CFG may be significantly better than the Zulu CFG, as indirectly evaluated by two linguists. The difference is mostly due to the large number of 'incorrect' spaced/compound future tense verbs in Zulu and strictness of the Zulu linguist in evaluation. The impact of this difference is that combined past+future rules cluster have a low similarity score unlike the present tense rules cluster where there are less incorrect Zulu verbs.

Regarding the second question, *Can a single merged set of grammar rules be used to produce correct verbs for both languages?*: yes, it is achievable. However, creating a single merget set of grammar rules for the two languages may result in a decrease in accuracy and make it difficult to maintain the rules when the quality is to be improved or new features added to the fragment. Our investigation has also revealed that the rules responsible for the suffix and final are similar in the two languages. The answer to this question may not hold for fragments that have additional/different features as they may exhibit differences. Current results of the investigated fragment show that a rules set that is modularised prefix, verb root, and suffix would easily allow the exploitation of similarity in Zulu and Xhosa. In that setup, the prefix module would still be different between the two languages. It is unclear to what extent this would hold if the alternation rules were not considered to be orthogonal to the CFGs. The modularisation method would, theoritically, also be extendable to other related languages in the Nguni group. Moreover, it may also extend to other Bantu langauges outside the Nguni group since Byamugisha, Keet, and DeRenzi (Byamugisha et al., 2016a) have shown that it is possible to bootstrap between geographically distant Bantu languages in knowledge-to-text NLG.

Regarding the third question, *What is the degree of improvement in verb correctness can be observed after the introduction of conditioning rules that target vowels on the CFGs?*: we see that the amount of strings, out of all the CFG generated strings, a significant amount of strings have consecutive vowels. For example, the minimum ratio of strings that will would need phonological conditioning for all the four types of verb roots that are listed in Table 1 is 45% for Xhosa and 67% for Zulu. Furthermore, if we assume that when the number of verbs that will be used in the weather forecast forecasts increases using the ratios of 20:1 in Zulu and 30:1 for Type D verb roots for each Type A verb root as influenced

by the NCHLT corpus then we see that at least 69% Zulu and at least 46% Xhosa strings generated by the CFGs will have consecutive vowels hence would need phonological conditioning. At first glance, these high values may suggest that phonological conditioning can substantially improve the quality of verbs used in a weather forecasts. However, phonological conditioning may not have great impact of the quality of verbs used in the weather forecast domains because (1) the probability of using each each verb, from all the possible verbs that can be generated from the rule sets, is not equal to all other verbs, (2) one is unlikely to generate a document with highly variable verbs even in the case of large documents in the domain, and (3) verbs with no consecutive vowels may be frequently used numerous times than verbs with consecutive vowels. Furthermore, not all the verbs generated by the CFGs exist in the two languages, and therefore these values are only accurate for the verb correctness level that has been reported. The degree of improvement brought on by phonological conditioning on consecutive vowels may be high, however, that is unlikely since weather forecasts may re-use a small set of verbs that do not have consecutive vowels.

6 CONCLUSION

We have presented an Xhosa corpus that can be used to determine the features of weather forecasts in Southern Africa and verb grammar rules that are the first to cover more than one tense. Evaluation within the weather domain showed a high syntactic and semantic correctness for the Xhosa and Zulu rules, whereas the linguist's evaluation with more verbs indicated room for improvement, partially due to strictness of the Zulu evaluator and Zulu being affected by phonological conditioning more than Xhosa. Nonetheless, this process already has uncovered an inconsistency in grammar textbooks and unspecified morphotactics. The four similarity measures (Funcs. 1-4) were applied on the two 'language' sets discussed in Section 3.1.3 and we obtained highest value with the Driver-Kroeber. The remaining three, namely Jaccard, Sorensen, and Sorgenfrei, can be rescaled to that measure. In particular, the Driver-Kroeber indicates 59.5% morphosyntactic similarity for the considered verb fragment based on the developed CFGs. There are 49 Zulu and 52 Xhosa CFG rules and they share 42 rules. The observed differences are present in the rules that encode morphotactics.

The verb rule differences of the variables are minor with respect to the structure. The three diverging terminal-generating rules have a substantial impact on the similarity measure value. The current versions of the CFGs generate a significant number of the verbs that have consecutive vowels thus indicating that if they were to be used in a large number of domains such that each CFG is used then then phonological conditioning rules may improvement the quality of the CFGs greatly. They are unlikely to offer significant improvement to verbs used in a single domain such as weather forecasts as a single domain may use the CFG generated verbs with consecutive infrequently, if at all.

Current and future work includes investigating the verb's prefix-suffix cross dependency and adding phonological conditioning to refine the grammar, and, pending linguistic advances, a comparison with Ndebele verb grammar.

https://doi.org/10.18489/sacj.v31i2.698

ACKNOWLEDGEMENTS

The first author acknowledges support from the Hasso Plattner Institute (HPI) Research School in CS4A at the University of Cape Town and the second author acknowledges support from the National Research Foundation (NRF) of South Africa (Grant Number 93397).

References

Bayram, S., Avcibas, I., Sankur, B., & Memon, N. D. (2005). Image manipulation detection with Binary Similarity Measures. In *13th European Signal Processing Conference, EUSIPCO 2005, Antalya, Turkey, September 4-8, 2005* (pp. 1–4). IEEE.

Bosch, S. E. [Sonja E.], Pretorius, L., Podile, K., & Fleisch, A. (2008). Experimental fast-tracking of morphological analysers for Nguni languages. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco* (pp. 2588–2595). European Language Resources Association.

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Proceedings of the International Conference on Compression and Complexity of Sequences, Positano, Salerno, Italy, June 11-13, 1997* (pp. 21–29). The Institute of Electrical and Electronics Engineers.

Byamugisha, J., Keet, C. M., & DeRenzi, B. (2016a). Bootstrapping a Runyankore CNL from an isiZulu CNL. In B. Davis, G. J. Pace, & A. Z. Wyner (Eds.), *Controlled Natural Language - 5th International Workshop, CNL 2016, Aberdeen, UK, July 25-27, 2016, Proceedings* (Vol. 9767, pp. 25–36). Lecture Notes in Computer Science. Springer.

Byamugisha, J., Keet, C. M., & DeRenzi, B. (2016b). Tense and aspect in Runyankore using a context-free grammar. In *Proc. of the Ninth INLG, September 5-8, 2016, Edinburgh, UK* (pp. 84–88).

Chavula, C., & Suleman, H. (2017). Morphological cluster induction of Bantu words using a weighted similarity measure. In M. Masinde (Ed.), *Proceedings of the South African Institute of Computer Scientists and Information Technologists, SAICSIT 2017, Thaba Nchu, South Africa, September 26-28, 2017* (6:1–6:9). ACM.

Cheetham, A. H., & Hazel, J. E. (1969). Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, *43*(5), 1130–1136.

Choi, S., Cha, S., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, *8*(1), 43–48.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Doke, C., Malcolm, D., Sikakana, J., & Vilakazi, B. (1990). *English-Zulu/Zulu-English dictionary*. Witwatersrand University Press.

Doke, C. M. (1931). *Textbook of Zulu grammar* (2rd). Longmans Southern Africa.

Doke, C. M. (1992). *Textbook of Zulu grammar* (6th). Maskew Miller Longman.

Driver, H. E., & Kroeber, A. L. (1932). *Quantitative expression of cultural relationships*. University of California Publications in American Archaeology and Ethnology. University of California Press.

Eiselen, R., & Puttkammer, M. J. (2014). Developing text resources for ten South African languages. In *Proc. of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, May 26-31, 2014*. (pp. 3698–3703).

Georgi, R., Xia, F., & Lewis, W. (2010). Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 385–393). COLING '10. Beijing, China: Association for Computational Linguistics.

Gilbert, N., & Keet, C. M. (2018). Automating question generation and marking of language learning exercises for isiZulu. In B. Davis, C. M. Keet, & A. Wyner (Eds.), *Controlled Natural Language - Proceedings of the Sixth International Workshop, CNL 2018, Maynooth, Co. Kildare, Ireland, August 27-28, 2018* (Vol. 304, pp. 31–40). Frontiers in Artificial Intelligence and Applications. 10.3233/978-1-61499-904-1-31

Goddard, C. (2011). *Semantic analysis: A practical introduction*. Oxford University Press.

Gowlett, D. (2014). Zone S. In D. Nurse & G. Philippson (Eds.), *The Bantu languages* (Chap. 30, pp. 609–636). Routledge.

Griesel, M., & Bosch, S. [Sonja]. (2014). Taking stock of the African Wordnet project: 5 years of development. In *Proc. of the Seventh Global Wordnet Conference, Tartu, Estonia, January 25-29, 2014* (pp. 148–153).

Grout, L. (1859). *The IsiZulu: A grammar of the Zulu language; accompanied with a historical introduction, also with an appendix*. James C. Buchanan. May & Davis. Trübner.

Grover, A. S., Van Huyssteen, G. B., & Pretorius, M. W. (2011). The South African human language technology audit. *Language resources and evaluation*, *45*(3), 271–288.

Huffaker, D., Jorgensen, J., Iacobelli, F., Tepper, P., & Cassell, J. (2006). Computational measures for language similarity across time in online communities. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech* (pp. 15–22). ACTS '09. New York City, New York: Association for Computational Linguistics.

Indukuri, K. V., Ambekar, A. A., & Sureka, A. (2007). Similarity analysis of patent claims using natural language processing techniques. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)* (Vol. 4, pp. 169–175). The Institute of Electrical and Electronics Engineers.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *The New Phytologist*, *11*(2), 37–50.

Keet, C. M. (2016). An assessment of orthographic similarity measures for several African languages. *CoRR*, *abs/1608.03065*.

Keet, C. M., & Khumalo, L. (2017a). Grammar rules for the isiZulu womplex verb. *Southern African linguistics and applied language studies*, *35*(2), 183–200. 10.2989/16073614.2017.1358097

Keet, C. M., & Khumalo, L. (2017b). Toward a knowledge-to-text controlled natural language of isiZulu. *Language resources and evaluation*, *51*(1), 131–157. 10.1007/s10579-016-9340-0

Keet, C. M., Xakaza, M., & Khumalo, L. (2017). Verbalising OWL ontologies in isiZulu with Python. In *The Semantic Web: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 - June 1, 2017* (pp. 59–64). 10.1007/978-3-319-70407-4_12

Khumalo, L. (2007). *An analysis of the Ndebele passive construction* (Doctoral dissertation, University of Oslo, Norway).

Mahlaza, Z., & Keet, C. M. (2018). Measuring verb similarity using binary coefficients with application to isiXhosa and isiZulu. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists* (pp. 65–71). SAICSIT '18. 10.1145/3278681.3278690

Maho, J. (1999). *A comparative study of Bantu noun classes*. Acta Universitatis Gothoburgunsis.

McLaren, J. (1944). *A Xhosa grammar, revised and re-written in the new orthography, edited by G. H. Welsh*. Longmans, Green and Company.

McLaren, J. (1955). *A Xhosa Grammar, revised and re-written in the new orthography, edited by G. H. Welsh*. Longmans, Green and Company.

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). SAICSIT '18. 10.1145/3278681.3278716

Msimang, C. T. (1989). *Some phonological aspects of the Tekela Nguni dialects* (Doctoral dissertation, University of South Africa).

Nabe, H. L., Dreyer, P. W., & Kakana, G. L. (1976). *Xhosa Dictionary: English, Xhosa, Afrikaans*. Educum Publishers.

Nurse, D. (2008). *Tense and aspect in Bantu*. Oxford University Press.

Pereira Jr., Á., & Ziviani, N. (2003). Syntactic similarity of Web documents. In *Proceedings of the IEEE/LEOS 3rd International Conference on Numerical Simulation of Semiconductor Optoelectronic Devices (IEEE Cat. No.03EX726)* (pp. 194–200). The Institute of Electrical and Electronics Engineers.

Peters, A. M. (1966). *A computer oriented generative grammar of the Xhosa verb* (Doctoral dissertation, University of Wisconsin, USA).

Rogers, D. J., Tanimoto, T. T. et al. (1960). A computer program for classifying plants. *Science (Washington)*, *132*, 1115–18.

Sorgenfrei, T. (1959). Molluscan assemblages from the marine middle Miocene of South Jutland and their environments. *Danmarks geologiske undersoegelse*, *2*(79), 403–408.

Spiegler, S., van der Spuy, A., & Flach, P. A. (2010). Ukwabelana - An open-source morphological Zulu corpus. In *Proc. of the 23rd International Conference on Computational Linguistics, 23-27 August 2010, Beijing, China* (pp. 1020–1028).

Taljaard, P., & Bosch, S. (1988). *Handbook of isiZulu*. JL van Schaik (Pty) Ltd.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J. D., Buscema, M., & Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of chemical information and modeling*, *52*, 2884–2901.

Warin, M., & Volk, M. (2004). *Using WordNet and semantic similarity to disambiguate an ontology*. Institutionen för lingvistik, Stockholms Universitet.

Warrens, M. J. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of classification*, *25*(2), 195–208.