Measuring verb similarity using binary coefficients with application to isiXhosa and isiZulu

Zola Mahlaza
University of Cape Town
Cape Town, South Africa
zmahlaza@cs.uct.ac.za

C. Maria Keet
University of Cape Town
Cape Town, South Africa
mkeet@cs.uct.ac.za

ABSTRACT

Natural Language Processing (NLP) for underresourced languages may benefit from a bootstrapping approach to utilise the sparse resources across closely related languages. This brings afore the question of language similarity, and therewith the question of how to measure that, so as to make informed predictions on potential success of bootstrapping. We present a method for measuring morphosyntactic similarity by developing Context Free Grammars (CFGs) for isiXhosa and isiZulu verb fragments that are relevant for the use case of weather forecast generation. We then investigate morphosyntactic similarity of the CFGs using parse tree analysis and four binary similarity measures. In particular, we selected four binary similarity measures from other domains and adapted them to our data, which are the word sets generated from the respective CFGs. The similarity measures together with the parse tree analysis are used to study the the extent to which both languages can be generated by a singular grammar fragment. The resulting 52 rules for isiXhosa and 49 rules for isiZulu overlap on 42 rules. This supports the existing intuition of similarity, as they are in the same language cluster. The morphosyntactic similarity measured with the binary coefficients reached 59.5% overall (adapted Driver-Kroeber), with 99.5% for the past tense only. This lower score cf. the structure of the CFG is attributable to the small differences in terminals in mainly the prefix of the verb. The parse tree analysis and binary similarity measures show that a modularised set of rules for the prefix, verb root, and suffix would allow the generation of the two languages with a single grammar where only the prefix requires differentiation.

ACM Reference Format:
Zola Mahlaza and C. Maria Keet. 2018. Measuring verb similarity using binary coefficients with application to isiXhosa and isiZulu. In *Proceedings of 2018 Annual Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT'18).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

South Africa is a multilingual country with 11 official languages. All the country's languages, with the exception of English, are underresourced despite being spoken by approximately over forty-five million people [31]. Investment in Information and communications technologies (ICTs) that have Southern African language interfaces is increasingly becoming important, as a number of surveys have shown that South Africans have relatively low English language proficiency and literacy skills [27, p.5], which is unlikely to change soon for a number of socio-economic and geo-political reasons specific to the country's history (the now defunct Apartheid system). Multinational companies such as Google Inc., Facebook, and Canonical Ltd. have started offering limited support for indigenous South African languages in their products. Generally, the state of Human Language Technologies (HLTs) for the indigenous languages is still poor [14, 15], however. It is known that the these languages are related and can be classified into two main groups[1]: the Nguni cluster with isiXhosa, isiZulu, siSwati, and isiNdebele and the Sotho-Tswana cluster with Sesotho, Setswana, and Sepedi [21]. This suggests that the few resources that exist for one language, such as linguistic annotation ontologies, morphological analysers, etc., may be leveraged to benefit other languages within the same group. Moreover, one may also be able to exploit their similarities to target all languages in a group when building new HLTs. There have been attempts to bootstrap a Bantu language's existing resources for another language within the same group (e.g [1] bootstrap morphological analysers for Nguni languages). However, to be able to bootstrap resources effectively for computational support, one has to have insight into the similarities of the languages in question beyond just intuition from linguists and speakers of the respective languages. This raises the question of how to measure similarity and, in fact, what aspect(s) of the languages is going to be measured on their similarity. While such questions are relevant in general for any pair of languages—e.g., in what way are, say, Spanish and Italian similar, and to what extent?—we focus here on the major language group, Nguni, which is the first/home language of over 30 million people. More specifically, for this context we seek to answer the following two questions:

(1) *How morphosyntactically similar are isiXhosa and isiZulu verbs?*
(2) *Can a single merged set of grammar rules be used to produce correct verbs for both languages?*

In order to answer them, we first develop CFGs for the isiZulu and isiXhosa verbs that are relevant for the use case of weather forecast generation. The weather domain is chosen because the resulting grammars can be used within a surface realizer when building an Natural Language Generation (NLG) system. The respective grammars are analysed using parse trees and binary similarity measures. A total of 51 distinct rules for isiXhosa and isiZulu were developed, observing some differences in the respective prefixes. Due to their

[1] We set aside Tshivenda and Xitsonga as they are the only South African languages in their respective language groups.

SAICSIT'18, September 2018, Port Elizabeth, South Africa

high rule similarity with respect to the variables, the parse tree similarity analysis does not provide much more insight into the similarity. The similarity measures over the strings generated by the respective CFGs were adapted from other application domains so as to make it usable for language comparison. This did give insight into their morphosyntactic similarities. Morphosyntactic similarity reached 59.5% (Driver-Kroeber metric) and scales as it should with the other three (Sorensen, Jaccard, Sorgenfrei). The word space similarity approach taken here may be useful for other roughly related languages.

The remainder of the paper is structured as follows. Section 2 describes preliminaries regarding the languages and the nature of binary similarity measures. Section 3.1 lays out the materials and methods to compare the two languages' verb. Section 4 and Section 5 present the results and discussion of the experiments. Section 6 concludes.

2 BACKGROUND

This section begins with a brief overview of the morphology of isiXhosa and isiZulu, and, by extension, the Nguni cluster. This is is followed by a short review of the methods that have been used to measure document similarity. Lastly, we discuss binary similarity measures and the selected measures that will be adapted in this investigation.

2.1 Languages and the Nguni cluster

Natural languages' morphology is generally classified into four types: polysynthetic, isolating, inflectional, and agglutinating [30, p.38], which may overlap. For instance, an inflectional language may have slight polysynthetic features and an agglutinating language can have slight fusional features [30]. Bantu languages—to which the Nguni cluster belongs—are generally labelled as agglutinating despite that not all are strictly so [24, p.28]. IsiXhosa and isiZulu have a complex morphology that is considered to be agglutinating. Geographically, these languages belong to Zone S of the classification of Bantu languages [21]. The two languages are "verby" [24, p.21] like other Bantu languages. This means that information that would be presented through syntactical or lexical methods in other languages is presented through complex verbal agglutination. To illustrate, consider the following two examples:

(1) *ii-bhokhwe* *zi-za-ku-hamb-a*
10.goats 10.SC-IFUT-INF-walk$_{\mathrm{VR}}$-FV
'The goats will leave'

(2) *si-ya-bon-an-a*
1pers pl-CONT-see$_{\mathrm{VR}}$-REC-FV
'we will see each other'

The '10' in (1) denotes the *noun class* of the plural noun for 'goat', which then requires the subject concord of that noun class to conjugate the verb (the '10.SC'), and similarly for 1st, 2nd and 3rd sg. and pl., like the *si-* to indicate the 1st pers. pl. 'we' in the second example (alike the *-amos* conjugation for *-ar* ending verbs in Spanish, but then as prefix). Note that in Bantu languages, each noun belongs to a noun class and each class has specific subject and object concord morphemes in the verb to ensure agreement when that noun is used as a subject or object, respectively. IsiZulu has 17 noun classes and isiXhosa 15, and each noun class has a specific subject and object concord.

Other morphemes include the immediate future tense *-za-*, the infinitive *-ku-* in example (1), and the reciprocative *-an-* in example (2). More generally, the verbs in the two languages can be inflected for aspect, mood, tense, and subject & object agreement (among other things) in the prefix to the verb root and extensions after the verb [19]. These verbs can be built from a (mostly fixed-order) slot system [20, 21], of which a simplified example is given in Figure 1.

The agglutination of the elements (clitics) in the slot system may require phonological conditioning, especially for vowel-commencing verb roots; e.g., *zi + akha = zakha*. We deem the rules for phonological conditioning orthogonal to the rules for the grammatical constituents of the verb.

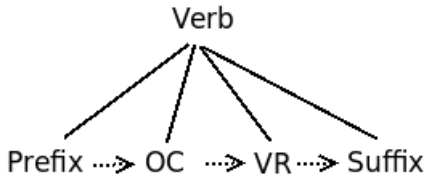


Figure 1: A (simplified) example of the verb slot system. Prefix is a slot that contains a number of elements such as the subject concord; OC is the object concord, VR is the verb root, and Suffix is slot that contains a number of elements such as the passive and final vowel.

2.2 Language similarity

Let us consider *resemblance* of two languages [2]: for any two documents X and Y, the function $0 \leq r(X, Y) \leq 1$ calculates the similarity of two documents. There are three types of similarities that can be measured for any two documents: lexical (and phrasal), syntactic, and semantic similarity. The comparison of documents need not directly make use of their text but it can use their typological features as done by Georgi et al. [12]. Lexical and semantic similarity can be measured by using a number of well established instruments such as latent semantic analysis, Spearman's correlation coefficient, Leacock-Chodorow, and Resnik [16, 34]. However, one can also make use of custom metrics [17]. Syntactic similarity has been compared at the sentence level through the use of w-shinglings [2, 25]. To the best of our knowledge, there has not been work that quantifies the verbal morphosyntax between two related agglutinating languages. The only work that is close to that goal was carried out by Chavula and Suleman [4] where the authors developed a weighted similarity measure for the purpose of Bantu language stem-based cluster induction using 3-shinglings. The next section, therefore, takes a step back in order to consider several generic similarity measures.

2.3 Similarity measures

Binary similarity measures are useful methods for measuring similarity between two sets of objects, where these measures may be part of techniques of measuring containment and resemblance [2]. Similarity coefficients are used broadly, ranging from botany [28]

to software fault localisation [6], to determine the similarity of binary feature vectors. A recent comprehensive survey collected 76 measures and classified them using hierarchical clustering [5] and another list can be found in Todeschini et al. [33].

We are interested in coefficients that we will adjust for our purposes; these are the Jaccard [18], Sorenson [7], Driver-Kroeber [11], and Sorgenfrei [29] (as cited by Todeschini et al. [33]) coefficients. These asymmetric metrics were chosen because they were well-documented and the time-consuming task of determining the nature of the other metrics given in Choi et al. [5] did not yield any results.

The Sorenson measure is also called the Sorenson-Dice or Dice measure, and the Driver-Kroeber is alternatively named the Ochiai measure [36]. The Sorenson measure was developed by Dice [7] to study the association of two species in a geographical region. Given two species Γ and Σ that exists primarily in two general regions, X and Y, one can determine the association of Σ to Γ as the ratio of the 'shared space' of Γ and Σ to size of the space in which Γ is found, which can be represented as $\Sigma \otimes \Gamma = \frac{|X \cap Y|}{|X|}$. It is complemented by the association of Γ to Σ, calculated with $\Gamma \otimes \Sigma = \frac{|X \cap Y|}{|Y|}$. This association differs for any two species based on which species is used as a base. Dice [7] devised a coincidence index in order to generalise the association index to ensure that association between two species was not variable based on which species is used as the base. The new index is the ratio of the sizes of the shared spaces[2] to the total number of species in both sets ($\Gamma \diamond \Sigma = \Sigma \diamond \Gamma = \frac{2|X \cap Y|}{|X|+|Y|}$). This is the measure that is known as the Sorenson-Dice coefficient.

The Jaccard index was first introduced as the 'coefficient of community' [18]. It measures the ratio of shared items to the total number of items that exist in two sets, and was initially applied to study the distribution of flora in the Alps. It can be calculated using the formula $\Sigma \diamond \Gamma = \frac{|X \cap Y|}{|X|+|Y|}$.

The Driver-Kroeber measure was developed in ethnology to measure the cultural traits that exist between two groups of people. Driver and Kroeber [11], unlike Dice [7], do not double the weight of the shared space. Instead, they merge the two indices by calculating the geometrical mean of the two association indices. The Driver-Kroeber metric is calculated with $\Sigma \diamond \Gamma = \frac{|X \cap Y|}{\sqrt{|X||Y|}}$. The Sorgenfrei metric consolidates the two association indices by multiplying them together to obtain $\Sigma \diamond \Gamma = \frac{|X \cap Y|^2}{|X||Y|}$.

3 SIMILARITY ASSESSMENT

The main aim of the similarity assessment is to find out how similar isiXhosa and isiZulu are with respect to their respective verbs and the secondary aim is to assess how well the selected similarity measures perform.

3.1 Materials and Methods

The assessment contains several steps, which are mainly the CFG development, adaptation of the similarity measures, and carrying out the comparisons, which are detailed in this section.

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

3.1.1 CFG development process. Because the verb structure is complex and the grammar for isiXhosa and isiZulu is neither well documented nor fully studied, we choose to formalise an interesting subset of it that also may be of practical use: those verb features used in weather forecasts. A corpus of English weather forecasts is collected from the South African Weather Service (SAWS)[3], which is then translated into isiXhosa by a member of the School of African Languages and Literature at the University of Cape Town. Verbs are manually extracted from the isiXhosa text to determine the grammatical features used. This was complemented with a literature intensive approach for designing the CFG, using [9, 10, 13, 22, 23, 26, 32] to collect detailed information about how the verb's mood, aspect, and tense function in the two languages. The quality of the rules is evaluated during development by one of the authors who speaks both languages and, for indicative purpose, by two linguists (one for each language). For the latter, a random sample of 100 strings were extracted from the total number of strings that the respective CFGs generate, and each linguist is asked to annotate them on syntactic and semantic correctness and any comments they may have.

3.1.2 Similarity measures assessment. We adapt the measures described in Section 2 to the setting of measuring similarity between natural languages. We begin by defining, for any two sets A and B of natural language's tokens, the variables $a = |A \cap B|$, $b = |B - A|$, and $c = |A - B|$. If we let B be one language (e.g., the set of isiZulu verb strings generated by its CFG) and A be another language (e.g., the set of isiXhosa verb strings form its CFG) then a is the number of verbs shared by the 'languages', b is the number of verbs that exist in B but not A (i.e., in isiZulu but not isiXhosa), and c is the number of verbs that exist in A but not B (i.e., isiXhosa but not isiZulu). The definition of these variables means that we can rewrite the binary similarity measures discussed in Section 2.3 in order to obtain Equations 1-4. Effectively, any instance of $|X \cap Y|$ in the original measures listed above is substituted with a, $|X|$ with $a + b$ and $|Y|$ with $a + c$.

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

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

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

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

Thus, the original measures' 'spaces' or 'regions'—physical in ecology and abstract in ethnology—containing elements (organisms social traits, respectively) are recast as spaces/regions that are filled with CFG-generated strings of a language.

While the formulae for the measures are different, they may behave in the same way given a certain input, and therewith be the same after applying a conversion ratio. In order to assess this, we generate 1024 cases of triples (a,b,c) using the Numpy[4] discrete uniform distribution random integer generator with the constraint

[3]http://www.weathersa.co.za/
[4]http://www.numpy.org/

$a + b + c = 1024$. We then calculate the difference between them for each of the 1024 triples and check the ratios obtained against the values we obtain with the four measures.

3.1.3 IsiXhosa and isiZulu comparison. It is common knowledge to anyone familiar with isiXhosa and isiZulu that they share some verbs. This opens up the ability to compare the verbal rules by comparing the verbs that are generated when the verb roots and concords are fixed. This comparison is based on the premise that a similarity evaluation on the shared language space can give an indication similarity between the two rule sets. We use the developed CFG rules, Python, and Natural Language Toolkit (NLTK) to generate two sets of verbs, and compare the resulting isiXhosa and isiZulu strings sets to each other. More precisely, we first select a verb root that exists in both isiXhosa and isiZulu, i.e.: *-zol-*, which means 'become calm'. We then select a subject concord of one of the noun classes, *li-* (for noun class 5), and leave the optional slot for the object concord empty. Then the production rules of the CFG are used to fill in the other slots to generate a set of strings for the two languages. That was done in a way to form four clusters of rules: using (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 respective resulting string sets are inputted to the natural language-adjusted binary similarity measures (Eqs. 1–4) to quantify similarity.

Further, 25 verb stems are extracted from an English-isiZulu dictionary [8], five a-commencing roots, 5 b-commencing roots etc. such that they also exist in isiXhosa. These verbs are used to determine whether there is a variation in the binary similarity results when different verb roots are used. Each verb root is again paired with the subject concord *li-* and an empty object concord, and the resulting strings are also provided as input to the binary similarity measures.

The subject and object concord were fixed in the aforementioned similarity calculations. It may be possible that the similarity could be different when a different pair of concords is used. We assess this by randomly picking five subject-object concord pairs existing in both isiZulu and isiXhosa that are then paired with the verb root *-zol-*; they are[5] (*a, zi*), (*i, wa*), (*i, yi*), (*lu, bu*), and (*u, yi*). The complete set of verb rules (aforementioned rule cluster 1) is used to generate strings with each of the concords and the resulting verbs are provided as input to the binary similarity measures.

4 RESULTS

This section describes the results obtained with the CFG and similarity comparisons.

4.1 CFG development and comparison

Twelve weather reports were collected from SAWS, one for each month to cover all seasons. They had an average of 488 words per document. Four sentences from each document were selected and translated into isiXhosa. The isiXhosa translations contained 53 verbs, with 27 unique verbs. The spread of the verbs based on mood is such that there were 22 indicative, 2 participial, and 3 subjunctive. The verb extensions associated with the weather are perfect, causative, neuter, and reciprocity. As a starting point for CFG development, we selected two verbal aspects (progressive and exclusive) and three tenses (past, present, and future) for inclusion in the grammar. The number of rules is shown in Table 1.

[5]In each tuple, the first element is the subject concord, and the second is the object concord.

Table 1: Total number of rules and intersection size of isiZulu and isiXhosa CFG rules. 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.

Language	Total	Terminal	Excl.	Excl. & Morphotact.	Morphotact.
isiZulu	49	13	6	8	22
isiXhosa	52	12	9	8	23
Intersection	42	11	6	8	17

IsiXhosa

Indicative & Participial

(x0.) $Verb \rightarrow SC \quad PC \quad OC \quad VR \quad S_{np}$

(x1.) $Verb \rightarrow A_{pe} \quad OC \quad VR \quad S_{np}$ a

Subjunctive

(x2.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np}$

(x3.) $Verb \rightarrow A_{pes} \quad OC \quad VR \quad S_{np}$ a

(x4.) $Prefix \rightarrow A_{es} \quad PC_1$

IsiZulu

Indicative & Participial

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

Subjunctive

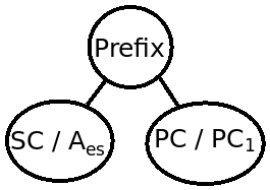
(z1.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_p$

(z2.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np}$ a

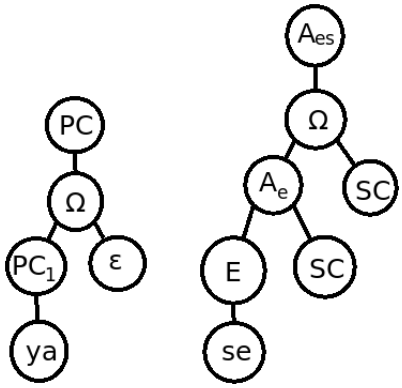
(z3.) $Prefix \rightarrow SI \quad SC \quad | \quad SC$

Figure 2: Rules that have differences between isiXhosa and isiZulu'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.

As can be seen in Table 1, there is a high overlap in rules overall. A difference is that isiXhosa has three additional rules that encode exclusivity in the use of (1) present tense and non-present continuity ($C \rightarrow PC|NPC$), (2) the progressive aspect and the remote past ($PRP \rightarrow P|PR$), and (3) the progressive aspect and the present continuity morpheme ($PCP \rightarrow P|PC$). For the terminals, the main difference lies in isiZulu's infinitive ($F \rightarrow uku$) and imperative present tense ($SI \rightarrow ma$), whereas isiXhosa has $F \rightarrow ku$ for the infinitive. Potentially interesting diverging rules are listed in Figure 2. IsiXhosa present indicative and participial rules have



(a) Tree representation of the combined isiXhosa/IsiZulu (Xh/Zu) prefix rule.



(b) Left: Subtree representation of an isiXhosa-only optional present continuity; Right: Subtree representation of an isiZulu-only presence of the exclusive aspect.

Figure 3: Tree representations of the isiXhosa and isiZulu's indicative and participial moods prefix (rule 0 in Figure 2). The Ω node represents mutual exclusiveness for its subtrees.

an additional rule (rule x1 in Figure 2) cf. isiZulu has two differences in the prefixes: 1) isiXhosa uses a fixed present continuity indicator (PC) whereas in isiZulu it can be empty (PC_1) (see rules x0 and z0), and 2) the isiZulu prefix incorporates the exclusive aspect. These differences are minor, as the underlying structure of the two rules is the same (see Figure 3). This is because the isiZulu rule is a 'super-rule' of its isiXhosa equivalent when considering the variables only.

Rules xh2 and zu1 for the subjective mood (see Figure 2) have minor differences in the suffix, where isiZulu requires the perfect suffix only (S_p) but isiXhosa also requires the neuter extension (S_{np}). This does not affect the overall structure of both rules, for there is only one mutually exclusive morpheme and the rest of the slots are equivalent. Rules x3 and z2 differ in their prefixes: isiXhosa deals with the simple, exclusive, and progressive aspects whereas the isiZulu rule deals with only the mandatory simple aspect. The rules defining the general prefix for the present subjunctive mood (x4 and z3 in Figure 2) differ in that isiXhosa incorporates continuity, whereas isiZulu does not.

Thus, the manual and, by extension parse tree, analysis of the CFGs does show a high degree of similarity at least at the variable-level. While still mostly qualitative, the precision with the rules developed here confirms more accurately the hitherto informal perceptions of similarity.

The indirect quality assessment by the two linguists yielded limited results. The isiXhosa linguist evaluated all 99 strings (with one having been removed due to an error) and the isiZulu linguist evaluated only 69 of the 99. IsiXhosa syntactic and semantic correctness were 52% and 58%, respectively, and for isiZulu they were 23% and 25%, respectively. While this may not look good, it must be noted that they are the first verb CFG to include tense other than present tense and they explicitly do not cover phonological conditioning. Further, this difference is partially due to the isiZulu linguist being more experienced in evaluating CFG outputs than the isiXhosa linguist and partially because isiZulu suffers from more phonological conditioning noise in the strings. This is especially apparent with afore-mentioned difference in immediate future tense (*uku-* vs *ku-*).

4.2 Similarity measures for isiZulu and isiXhosa

Having shown that the grammar rules are very similar, we now proceed to the quantitative results with the adapted binary similarity measures of Jaccard, Sorenson, Driver-Kroeber, and Sorgenfrei. 504 unique strings were generated for the first assessment with *-zol-* and 12600 with the 25 shared verb roots. This resulted in the similarity measures listed in Table 2.

Table 2: Calculated 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 abide by the relation $0 \leq Sorg(A,B) \leq J(A,B) \leq S(A,B) \leq DK(A,B) \leq 1$, for all binary vectors A, B [33, 35]. The results in Table 2 show that there is little difference between the measured similarities for each tense cluster (past, present, and future) within each binary coefficient. The past tense cluster, for each measure, has the highest similarity values when compared to the other two tenses regardless the measure used.

The results obtained by using different subject and object concords paired with *-zol-* as well as modifying the verb roots are the same as for *-zol-* with the fixed subject concord listed in Table 2 (results omitted). Thus, the actual vocabulary and slots selected in the similarity assessment had no effect on the similarity measure.

Recall that the Sorgenfrei index is mathematically similar to the Jaccard (Sorg ≈ J) and the Sorenson index is similar to the Driver-Kroeber (S ≈ DK) (Eqs. 1-4). This closeness is confirmed by the results in Table 2: the Jaccard and Sorgenfrei measures differ by at most 0.071, and the Driver-Kroeber and Sorenson measures differ by at most 0.007 for all the compared verb sets.

4.3 Similarity measure behaviour

To better interpret the similarity values obtained for the language spaces of the isiXhosa and isiZulu words generated, as well as for

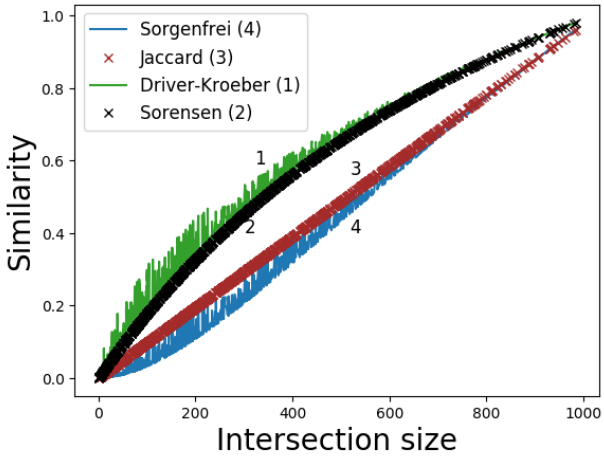


Figure 4: Difference in four binary similarity methods when the size of the intersection between two sets increases and the sets' complement decreases. Similarity is measured with value between zero (Different) and one (Equivalent).

any pair of languages one may wish to assess likewise, we now proceed to the results of the measures' behaviour.

The maximum and average differences between the Sorenson and Sorgenfrei measure are 0.250 and 0.166, respectively, which is higher than the differences between the Driver-Kroeber and Jaccard ones (0.247 and 0.142). The Jaccard and Driver-Kroeber measures are more intuitive for language spaces[6], because they are sensitive to the fact that the two 'languages' may have different sizes and uses this to generate a better measure.

To convert one into the other, we need a more detailed representation of the relationship between the four metrics, which is shown in Figure 4 that was obtained with the 1024 (a, b, c) triples. While it varies by the size of the sets, a Jaccard measure can be obtained from a Sorgenfrei one by adding about 0.04, the Driver-Kroeber from the Jaccard by adding about 0.14, and the Sorensen from the Driver-Kroeber by subtracting about 0.02. These rescalings of the metrics are dependent on the size of the intersection. When the intersection size is small and the sets' complement sizes is large then the difference between the metrics is low, as illustrated in Figure 4. The behaviour of the metrics, relative to each other, functions in the same manner irrespective of the difference in tense of the rules. Applying this to Table 2, then, 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.

Thus, while the actual similarity measure values differ, they are interchangeable modulus the conversion factor, and at least for the word space of the words generated by the respective CFGs of isiXhosa and isiZulu, they behave accordingly.

[6]Recall that the Jaccard is able to deduce the ratio of the shared items to the complete set of verbs and the Driver-Kroeber gives an average measure of the similarity of the each set to the other

5 DISCUSSION

We now return to the questions posed in Section 1. First, *How similar are isiXhosa and isiZulu verbs?*: the results show that the most intuitive similarity measures are Jaccard and Driver-Kroeber, reporting 42% (Jaccard) and 59.5% (Driver-Kroeber) similarity for the isiZulu and isiXhosa verb fragment investigated. Their difference is close to the average difference between the two measures (recall Section 4.3) and the 0.175 difference is due to the difference in the formulation of the two metrics and not the languages. Or: when scaled, these metrics are effectively interchangeable.

The similarity value may turn out higher or lower for other fragments or larger fragment, and we caution against a blank generalisability at this stage. The respective CFGs exhibit overgeneration, and the quality of the isiXhosa rules may be better than for isiZulu, as indirectly evaluated by two linguists. The difference is mostly due to the large number of 'incorrect' spaced/compound future tense verbs in isiZulu. Its impact is that the combined past+future rules cluster have a low similarity score unlike the present tense rules cluster where there are less incorrect isiZulu verbs. This also indicates that adding phonological conditioning likely will have the largest impact on attempts to reduce the overgeneration.

Regarding the second question, *Can a single merged set of grammar rules be used to produce correct verbs for both languages?*, the answer is that it may be feasible but at the cost of some accuracy and complexity of maintainability. The investigation into and formalisation of the part of the verb grammar showed that the suffix and final vowel in the two languages operate in the same manner. However, it cannot be excluded that there may be differences when other suffixal features are considered. The analysis suggests that a modularised rule set for the prefix, verb root, and suffix would enable the exploitation of the similarities between the two languages, and for these two languages only the prefix module would be differentiated. Such an approach would, in theory at least, then also be extensible to other closely related languages in the Nguni group, such as isiNdebele and siSwati, and perhaps also other Bantu languages, as bootstrapping between geographically distant Bantu languages was deemed feasible for at least one experiment in knowledge-to-text natural language generation [3].

Although the paper focused specifically on trying to find similarity measures that would work for isiZulu and isiXhosa, the same approach can easily be used for other pairs of related languages. That is, not just by comparing grammar rules, but especially the word spaces of the strings generated by the grammars to obtain a value for comparison that, in turn can inform potential for bootstrapping for a range of NLP tasks. For instance, present and past tense generation is similar in Spanish, Portuguese, and Italian where some of the terminals are the same or have a single letter permutation, such as 3rd pers. sg. *-a* among all three and five among Spanish and Portuguese (differing 3rd pers. pl. *-an* vs *-am*), and past tenses *-ato* in Italian vs *-ado* in Spanish and Portuguese, respectively, for *-are*/*-ar* ending verbs. The approach presented in this paper can assist with quantifying the intuition that Spanish and Portuguese are more similar than Spanish and Italian.

6 CONCLUSION

Language space similarity was assessed with four similarity measures, where the Driver-Kroeber metric returns the highest value, and the other three—Jaccard, Sorensen, and Sorgenfrei—can be rescaled to it. For isiXhosa and isiZulu, it showed a 59.5% (Driver-Kroeber) morphosyntactic similarity of the words generated by their respective CFGs of a fragment of the verb. The rules for the verb fragments were developed for this task, noting that the 49 isiZulu and 52 isiXhosa rules share 42 rules. The differences that are found 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.

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 isiNdebele verb grammar.

ACKNOWLEDGMENTS

This work is based on the research supported by the Hasso Plattner Institute (HPI) Research School in CS4A at UCT and the National Research Foundation (NRF) of South Africa (Grant Number 93397).

REFERENCES

[1] Sonja E. Bosch, Laurette Pretorius, Kholisa Podile, and Axel Fleisch. 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*. European Language Resources Association.

[2] Andrei Z. Broder. 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*. 21–29.

[3] Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2016. Bootstrapping a Runyankore CNL from an isiZulu CNL. In *Controlled Natural Language - 5th International Workshop, CNL 2016, Aberdeen, UK, July 25-27, 2016, Proceedings (Lecture Notes in Computer Science)*, Brian Davis, Gordon J. Pace, and Adam Z. Wyner (Eds.), Vol. 9767. Springer, 25–36.

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

[5] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. 2010. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics* 8, 1 (2010), 43–48.

[6] Valentin Dallmeier, Christian Lindig, and Andreas Zeller. 2005. Lightweight Defect Localization for Java. In *ECOOP 2005 - Object-Oriented Programming, 19th European Conference, Glasgow, UK, July 25-29, 2005, Proceedings (Lecture Notes in Computer Science)*, Andrew P. Black (Ed.), Vol. 3586. Springer, 528–550.

[7] Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26, 3 (1945), 297–302.

[8] C.M. Doke, D.M. Malcolm, J.M.A. Sikakana, and B.W. Vilakazi. 1990. *English-Zulu/Zulu-English dictionary*. Witwatersrand University Press.

[9] Clement Martyn Doke. 1931. *Textbook of Zulu grammar* (2rd ed.). Longmans Southern Africa.

[10] Clement Martyn Doke. 1992. *Textbook of Zulu grammar* (6th ed.). Maskew Miller Longman.

[11] Harold E. Driver and Alfred L. Kroeber. 1932. *Quantitative Expression of Cultural Relationships*. University of California Press.

[12] Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing Language Similarity Across Genetic and Typologically-based Groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 385–393.

[13] Lewis Grout. 1859. *The IsiZulu: A Grammar of the Zulu Language; accompanied with a historical introduction, also with an appendix*. James C. Buchanan. May & Davis. Trübner.

[14] Aditi Sharma Grover, Gerhard B Van Huyssteen, and Marthinus W Pretorius. 2010. South African human language technologies audit. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation: LREC 2010*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.). European language resources distribution agency.

[15] Aditi Sharma Grover, Gerhard B Van Huyssteen, and Marthinus W Pretorius. 2011. The South African human language technology audit. *Language resources and evaluation* 45, 3 (2011), 271–288.

[16] David Huffaker, Joseph Jorgensen, Francisco Iacobelli, Paul Tepper, and Justine Cassell. 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 (ACTS '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 15–22.

[17] Kishore V. Indukuri, Anurag A. Ambekar, and Ashish Sureka. 2007. Similarity Analysis of Patent Claims Using Natural Language Processing Techniques. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, Vol. 4. 169–175.

[18] Paul Jaccard. 1912. The Distribution of the Flora in the Alpine Zone. *The New Phytologist* 11, 2 (1912), 37–50.

[19] C. Maria Keet and Langa Khumalo. 2017. Grammar rules for the isiZulu complex verb. *Southern African Linguistics and Applied Language Studies* 35, 2 (2017), 183–200. https://doi.org/10.2989/16073614.2017.1358097

[20] Langa Khumalo. 2007. *An analysis of the Ndebele passive construction*. Ph.D. Dissertation. University of Oslo, Norway.

[21] Jouni Maho. 1999. *A comparative study of Bantu noun classes*. Acta Universitatis Gothoburgunsis.

[22] James McLaren. 1936. *A Xhosa Grammar, revised and re-written in the new orthography, edited by G. H. Welsh*. Longmans, Green and Company.

[23] James McLaren. 1955. *A Xhosa Grammar, revised and re-written in the new orthography, edited by G. H. Welsh*. Longmans, Green and Company.

[24] Derek Nurse. 2008. *Tense and aspect in Bantu*. Oxford University Press.

[25] Álvaro Pereira Jr. and Nivio Ziviani. 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)*. 194–200.

[26] Ann M Peters. 1966. *A computer oriented generative grammar of the Xhosa verb*. Ph.D. Dissertation. University of Wisconsin, USA.

[27] Dorrit Posel and Jochen Zeller. 2011. Home language and English language ability in South Africa: Insights from new data. *Southern African Linguistics and Applied Language Studies* 29, 2 (2011), 115–126.

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

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

[30] Andrew Spencer. 1991. *Morphological theory: An introduction to word structure in generative grammar*. Wiley-Blackwell.

[31] Statistics South Africa. 2012. Census 2011 : Census in brief. http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf. Accessed: 23 November 2017.

[32] PC Taljaard and SE Bosch. 1988. *Handbook of isiZulu*. JL van Schaik (Pty) Ltd.

[33] Roberto Todeschini, Viviana Consonni, Hua Xiang, John D. Holliday, Massimo Buscema, and Peter Willett. 2012. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *Journal of Chemical Information and Modeling* 52 (2012), 2884–2901.

[34] Martin Warin and Martin Volk. 2004. *Using WordNet and Semantic Similarity to Disambiguate an Ontology*. Technical Report. Institutionen för lingvistik, Stockholms Universitet.

[35] Matthijs J. Warrens. 2008. Bounds of Resemblance Measures for Binary (Presence/Absence) Variables. *Journal of Classification* 25, 2 (2008), 195–208.

[36] Kok-Seng Wong and Myung Ho Kim. 2013. Privacy-preserving similarity coefficients for binary data. *Computers & Mathematics with Applications* 65, 9 (2013), 1280 – 1290. Advanced Information Security.