Grammars for generating isiXhosa and isiZulu weather bulletin verbs

Zola Mahlaza
Supervisor : Dr. C. Maria Keet

Thesis presented for the degree of Master of Science
to
The Department of Computer Science

University of Cape Town
Cape Town, South Africa
January 2018

Grammars for generating isiXhosa and isiZulu weather bulletin verbs

Abstract

The Met Office has investigated the use of natural language generation (NLG) technologies to streamline the production of weather forecasts. Their approach would be of great benefit in South Africa because there is no fast and large scale producer, automated or otherwise, of textual weather summaries for Nguni languages. This is because of, among other things, the complexity of Nguni languages. The structure of these languages is very different from Indo-European languages, and therefore we cannot reuse existing technologies that were developed for the latter group. Traditional NLG techniques such as templates are not compatible with 'Bantu' languages, and existing works that document scaled-down 'Bantu' language grammars are also not sufficient to generate weather text. In pursuance of generating weather text in isiXhosa and isiZulu – we restricted our text to only verbs in order to ensure a manageable scope. In particular, we have developed a corpus of weather sentences in order to determine verb features. We then created context free verbal grammar rules using an incremental approach. The quality of these rules was evaluated using two linguists. We then investigated the grammatical similarity of isiZulu verbs with their isiXhosa counterparts, and the extent to which a singular merged set of grammar rules can be used to produce correct verbs for both languages. The similarity analysis of the two languages was done through the developed rules' parse trees, and by applying binary similarity measures on the sets of verbs generated by the rules. The parse trees show that the differences between the verb's components are minor, and the similarity measures indicate that the verb sets are at most 59.5% similar (Driver-Kroeber metric). We also examined the importance of the phonological conditioning process by developing functions that calculate the ratio of verbs that will require conditioning out of the total strings that can be generated. We have found that the phonological conditioning process affects at least 45% of strings for isiXhosa, and at least 67% of strings for isiZulu depending on the type of verb root that is used. Overall, this work shows that the differences between isiXhosa and isiZulu verbs are minor, however, the exploitation of these similarities for the goal of creating a unified rule set for both languages cannot be achieved without significant maintainability compromises because there are dependencies that exist in one language and not the other between the verb's 'modules'. Furthermore, the phonological conditioning process should be implemented in order to improve generated text due to the high ratio of verbs it affects.

Contents

List of Figures

for Nokulunga Mahlaza.

Acknowledgments

Ndiyabulela kuGqirha C. Maria Keet ngenxaso kunye nolwazi andiphungulele lona. Ndiyabulela kusapho lonke lwakwaMahlaza ingakumbi uSophakama ngenxaso. Enkosi kuSomila Fuma kunye noMaserame Magakoe ngokundiboleka indlebe xa bekunzima. Ndithi maz' enethole kumalungu eqela le-Digital Libraries kwiDjunivesithi yaseKapa ngeengcebiso kunye neenkqubo zabo ezindincedileyo ekucacelweni kwam yisayensi. Ndiyabulela nakuGriqha Langa Khumalo, kunye noMnumzana uZukile Jama ngokundinceda ekuvavanyeni izenzi zesiZulu kunye nezesiXhosa.

1 Introduction

Information about the state of the weather is integral to our daily routines. Weekday morning news, on television and radio, features weather reports that prepare us for the day ahead. This information is required by individuals who work in different fields and may also speak different languages. There are attempts to improve the way these reports are prepared and delivered to consumers. The Met Office[1], the United Kingdom's national weather agency, together with Arria NLG[2] have been investigating the use of natural language generation (NLG) technologies in preparing weather reports for many areas in a short period of time [118]. Earlier efforts to achieve this goal were done by, to name a few, the USA's Environmental Science Services Administration (ESSA) Weather Bureau [37] and Canada's Department of the Environment [12]. NLG is the study of techniques involved in the production of natural language texts from structured representations of data, information, or knowledge. These representations range from databases, raw numerical data to formal representations such as ontologies.

The focus of our work is natural language generation for a subset of the so-called Bantu languages. Bantu languages are spoken in approximately 27 African countries [73, p.1]. They are said to be related, however, finding a set of features that uniquely identifies these languages as belonging to one group is very difficult. This difficulty arises due to the large number of these languages and the variations that exist as a result of their speakers' geographical spread [73]. Nurse and Phillipson's [73] work reveals that the one 'feature' these languages share is being "underdescribed" [73, p.4], that is to say, it is difficult to obtain work detailing each languages' grammar.

[1] http://www.metoffice.gov.uk/
[2] https://www.arria.com/

The target languages of this work belong to the Guthrie Zone S, and these are the Bantu languages largely spoken in Southern Africa. The specific languages we consider are isiZulu and isiXhosa, languages that belong to the Zunda variety of Nguni languages (a subset of Southern Bantu languages). They are the largest South African languages by number of first language speakers (home-language speakers) [119]. They are predominantly spoken in the in-land province of Gauteng and the coastal provinces of the Western Cape, Eastern Cape, and KwaZulu-Natal.

At the time of writing, Nguni weather reports that are consumed by the South African Nguni population, are produced by the South African Broadcasting Cooperation (SABC). The company's TV channel (SABC 1) produces a daily report in isiZulu/isiXhosa at 19h00 South African Standard Time (SAST) and a daily isiNdebele/-siSwati report at 17h30 SAST. It also provides reports through its Nguni language radio stations (e.g Umhlobo Wenene[3], Ukhozi[4], etc). These reports give a very brief outline for predetermined locations. This approach is sufficient for a uniform audience that requires the forecasts for non-technical reasons. However, it falls short for a diverse audience that may not only require the current form but also detailed agricultural, or marine forecasts. Generally, investment in Nguni language generation technology is increasingly becoming important because a number of surveys have shown that South Africans have low English language proficiency and literacy skills [78, p.5]. This is unlikely to change for a number of reasons, for instance, children in rural schools have been found to never speak, read, or write English outside the formal school environment [105]. Furthermore, South Africa is still segregated (to a degree) as a result of the now defunct laws such as the Black Homeland Citizenship Act of 1970 and Group Areas Act of 1950. For instance, most of the Eastern Cape's 9 municipal regions are disproportionately made up of isiXhosa first language speakers, with the exception of Cacadu that has a 43.6% Afrikaans speaker presence [120]. A similar pattern can be observed with KwaZulu-Natal's population distribution [121, p.38]. This means that many people in these areas have no need to be fluent in English in their daily lives. It then makes sense for us to invest time developing NLG solutions that will present information in the languages with which they are comfortable. It is for this reason that we attempt to create NLG resources for isiXhosa and isiZulu.

We consider NLG an important field because its applications include the creation of computer tools that are able to explain medical data to patients, summarise statistical data, etc. [94, p.2]. The Bateman & Zock [3] list of NLG systems shows a variation of systems that exist in a number of fields ranging from systems that produce flight information [2] to systems that produce biographies [56]. The same list shows that the most popular application of NLG is health-care/medicine followed by the automation of the production of weather summary text. NLG has the potential to empower individuals who are not experts in a particular area, people who do not have the capacity to interpret the raw data, to understand certain datasets. Furthermore, it reduces the human effort required in the creation of reports. For instance, Arria NLG, one of the most successful NLG companies, boasts that their system can produce a neonatal health-care report in real-time whereas it takes a human author approximately 2 hours. This technology could have a positive impact in the way data and corresponding reports are produced and consumed in South Africa. This is especially true if there were NLG methods and technologies to handle the 9

[3] http://www.umhlobowenenefm.co.za/
[4] http://www.ukhozifm.co.za/

official languages of South Africa (excluding English and Afrikaans) that receive less research investment.

The two languages in question are under-resourced with respect to human language technologies (HLTs) despite being the most widely spoken in South Africa. South Africa's past political situation meant that prioritisation of institutional support for languages was not based on the number of language speakers. It was exclusively based on the race of the language speakers hence Afrikaans is the only South African language that is not significantly under-resourced. Moreover, Afrikaans has been able to benefit from the bootstrapping of Dutch language resources due to the similarity of the two languages [43, p.283]. HLTs, as the name suggests, are technologies that are capable and designed with the intention of working with human languages. We get them from areas such as NLG, document processing, parsing, machine translation, etc. South Africa's indigenous languages' status of being under-resourced is not something to be celebrated and as such, there have been attempts to address the problem. The government has made attempts to facilitate the advancement of indigenous languages in compliance with its national language policy framework [40]. Academic work towards HLTs has been funded by the Department of Arts and Culture (DAC), Department of Science and Technology (DST) and the National Research Foundation (NRF) [43]. However, this funding does not result in uniform investment across all areas of HLTs. An important area such as text generation is of low priority to HLT experts [43, p.277] and it receives the least activity for South African languages [42].

We cannot simply use an existing NLG system as-is because a language's features have a crucial impact on the way data is stored and processed in NLG. The GenNext [102] system, for instance, uses a corpus for determining the structure of the text. The corpus is also used to determine the various templates for the important sentences. However, we know that templates cannot be used with all natural languages [50]. Furthermore, verbalisation systems such as OntoVerbal [59] [60], for instance, assume that the labels used in ontologies are not random but are English strings that can be exploited in the verbalisation process. We also know that numerous existing NLG systems are designed to generate Indo-European languages. For instance, the Bateman & Zock [3] list has a total of 272 English NLG systems out of 388 systems. The other popular output languages for NLG systems in the list are German, French, Dutch and Spanish (in that order). Bantu languages are different from all these languages. The main two features that differentiate Bantu languages from the aforementioned languages are the systems of noun classes and concordial agreement. These two complexities are the reason why existing NLG systems are not suitable for Bantu languages, and hence, it makes sense to investigate the use of grammars for the generation of isiXhosa and isiZulu [50]. Complete natural language grammars are time-consuming and difficult to develop, especially for under-described languages such as isiZulu and isiXhosa. Therefore, in effort to build an NLG system that makes use of a grammar for surface realisation we begin by developing grammars for the verb. This is because it is a complex part of speech as Bantu languages are 'verby' [72, p.21]. Moreover, we put emphasis on understanding the grammatical similarity between the two languages with the hope of exploiting it in order to build 'bi-lingual' systems efficiently, if possible. This work will restrict the proposed grammars to ones that are able to generate weather reports for isiZulu and isiXhosa.

1.1 Problem statement

In our examination of the current state and use of Nguni languages, we have observed that there is no fast and large scale producer, automated or otherwise, of textual weather summaries in said languages. This is due to several factors, such as (1) There are multiple Nguni languages, each of which has numerous dialects and hiring human authors to interpret weather data and produce these summaries is expensive and inefficient, (2) There is no automated system to achieve the stated goal because of, among other things, the complexity of Nguni languages (that is due to their noun class systems and concordial agreement) and the small number of computer scientists working with Nguni languages. Existing tools cannot be re-used as-is due to the nature of these languages. Furthermore, to our knowledge, the grammatical similarity between isiZulu and isiXhosa has never been formally quantified.

1.2 Aim

The aim of this work is to develop grammars that will used in the surface realisation of weather report verbs in isiZulu and isiXhosa. These grammars can only be developed once we understand the set of isiZulu and isiXhosa verb grammatical features that exist in weather reports in Southern Africa for all seasons. Furthermore, we will study the feasibility of using a single high level verb grammar in an NLG system for the purpose of generating texts in two closely related languages. We do this by studying the similarity of isiZulu and isiXhosa weather forecast verb grammars. Lastly, we investigate the degree to which a set of phonological conditioning rules can improve the context free grammar (CFG) thus giving NLG developers for Nguni languages more information on how to prioritise phonological conditioning.

1.3 Research questions

In this work we develop CFG rules in order to investigate the following questions:

- How grammatically similar are isiZulu weather forecast verbs with their isiXhosa counterparts?
- Can a singular merged set of weather forecast verb grammar rules be used to produce correct verbs for both languages?
- What is the degree of improvement in grammatical correctness that can be brought on by the introduction of phonological conditioning rules on the CFGs?

1.4 Methodology

The grammatical features of the verb that will be considered are dependent on the nature of weather forecasts. We will follow the strategy of using a corpus to determine the output text requirements. This approach was made popular by Reiter and Dale [21]. The absence of an isiXhosa or isiZulu weather corpus for the South African climate necessitates the building of a corpus. The weather corpus will be collected from the South African Weather Service (SAWS) and translated into isiXhosa by members of the School of African Languages and Literature at

the University of Cape Town. Translation of the collected data is necessary since the set of grammatical features in Bantu languages and English need not be the same.

The weather corpus is used to extract grammatical features that are present in describing the weather. These features are needed when developing the grammar rules for isiZulu and isiXhosa. These rules generate the verb as it is the most complex item in Bantu languages. A literature intensive approach for designing the rules will be taken. This approach uses grammar textbooks to collect detailed information about the corpus-extracted grammatical features. An incremental development will be taken as it allows easier management of numerous rules. Furthermore, it allows greater control in verifying correctness. The evaluation of the quality of the rules will use an expertise-oriented approach. This is the dependence upon professional expertise to judge the quality of the rules [99, p.117] [98, p.483]. The expert-oriented approach is used to ensure that a 'formal' language is used thus catering for all the variable consumers of weather text. The chosen experts are linguists, and they will annotate the correctness of verbs that are provided to them.

IsiXhosa and isiZulu will be compared through verb rule parse trees and an indirect comparison approach. The indirect approach entails calculating the similarity of the languages by comparing the language spaces that are generated by the rules. In other words, the verb sets that will be generated using the verb rules will be compared. The comparisons will be done using binary similarity measures. These measures are popular for measuring co-occurrence of species in two locations. Studying the number of times phonological conditioning is required will be done through a mathematical quantification of incidents in which it could be required. This will save time as existing rules for fixing verbs will not be implemented. Furthermore, it allows us to also quantify possible incidents that may require conditioning and may have been unaccounted for in the literature.

1.5 Outline

The thesis is structured such that Chapter 2 will describe natural language generation, and will discuss how NLG systems are built, whilst putting a strong emphasis on surface realisation. We will discuss the evolution of the surface realisation tools over years and discuss the reasons for selecting a CFG for modelling the verb rules. In the chapter, we will also discuss the existing literature in which computational methods for generating and parsing Southern African languages have been documented. The chapter ends with a definition of four binary similarity measures, their formulation and behaviour. These similarity measures will be used to compare similarity of isiXhosa and isiZulu. The development of CFG rules requires knowing the grammar features of verbs that exist in weather forecasts. Chapter 3 defines the development process of weather corpus that will be used to discover these features. After the verb's features are extracted from the corpus, the process of developing verb rules and the resulting rules for isiXhosa and isiZulu are described in Chapter 4. Once the CFG rules exist, we need to evaluate their correctness and use them to address the research questions. Hence, Chapter 5 will report the methods and materials that we used to (1) evaluate the correctness of the isiXhosa and isiZulu rules, (2) compare the similarity of the isiXhosa and isiZulu rules, and (3) determine the number of verbs that require phonological conditioning.

The third (3) task is done by developing equations that can quantify the number of consecutive vowel incidents in isiZulu and isiXhosa strings. Moreover, Chapter 5 also reports the results of the aforementioned three processes, details how the research questions have been answered, and suggests improvements that can be made in the methodology and evaluation process. Lastly, Chapter 6 concludes the thesis by detailing the findings of this work and describes what could be done in future work.

2

Background and Related Work

This chapter begins by highlighting a contention surrounding the name of the language family to which the considered languages belong and a brief introduction to the two languages in Section 2.2. In Section 2.3, we then define natural language generation by looking at its role in natural language processing and focus on the process of surface realisation. Section 2.4 discusses the various weather forecast generation system that have been built over the years. Section 2.5 discusses the works that have investigated computational methods for Southern African languages, in particular, those which focus on generating and parsing natural languages. Finally, section 2.7 explains what binary similarity measures are and describes the behaviour of four binary similarity functions.

2.1 On the term 'Bantu'

IsiXhosa and isiZulu are languages that are spoken in Southern Africa, and they fall into the so-called Bantu language family. We say "so-called" because the naming of this particular language group is subject to criticism and opposition. The term Bantu and its other forms means 'people' in the languages it is used to identify. The naming of this language group in this manner started in the middle of the 19th century by Bleek, a German linguist, who has been considered by some as the 'father' of African philology [132]. Objections to the use of the identifier are not new and are probably as old as the naming itself. Modern opposition to the name is due to its derogatory connotation that was common during Apartheid in South Africa. Nonetheless, its use has persisted because, apparently, "its meaning [is] clear, and [...] it is easily pronounced" [132] (The age and the nature of the book making this argument leads one to assume that the author means that it is pronounceable by Europeans). This

justification is analogous to the situation documented by the popular South African poem *My Name* by Magoleng wa Selepe [104, p11]. This work continues the use of this term due to its ubiquity and our lack of knowledge of its alternatives. It is hoped that the use of the term will allow the reader to expeditiously identify the language group being referred to, without causing offence. We have considered the use of a tempo-invented word within this thesis, however, an undertaking of that nature would work against the goal of efficient identification of the language group in question. Furthermore, the exercise of coining an alternative term to use in this work is not of great significance here as we are not focusing on sociolinguistics.

2.2 Bantu languages: IsiZulu and isiXhosa

There are four major families of languages in the African continent, and these are Niger-Congo, Nilo-Saharan, Afro-asiatic, and Khoisan [58, p13]. IsiZulu and isiXhosa, the languages considered here, belong to the Niger-Congo family within a subfamily named Bantu. Geographically, these languages belong to Zone S of the classification of Bantu languages [61]. They are also found outside that Zone, in the South Western area of South Africa. The two languages have the highest number of first languages speakers in South Africa and they can be found as part of the top 3 home/first languages in 7 of the 9 provinces in South Africa [119]. These two languages, together with siSwati, and isiNdebele, make up the Nguni languages — a group whose verbs have complex morphology and nouns belong to specific classes. Generally, there are four categories of natural languages' morphology: polysynthetic, isolating, inflectional, and agglutinating [112, p.38]. The Bantu languages, to which the Nguni group belongs, are generally labelled as agglutinating despite that not all strictly are [72, p.28]. To illustrate the verbs complexity, consider the following two examples:

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

(2) *ba-sa-si-neth-isis-a*
3pers pl-ASP$_{P}$-OC-rain$_{VR}$-INT-FV
'they are still causing it rain intensely on us'

The '10' in example (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 *ba-* to indicate the 3rd pers. pl. 'they' in example (2). Note that in Bantu languages, each noun belongs to a noun class (classes shown in Table 2.2.1) 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.

Other morphemes, from the examples above, include the immediate future tense *-za-*, the infinitive *-ku-* from example (1) and, from example (2), the intensive verb extension *-isis-*, object concord -si-, and progressive aspect *-sa-*. 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 [52]. These verbs can be built from an (mostly fixed-order) slot system that can vary based on the verb features under consideration [61, 55].

Tense and aspect are grammatical features that apply to the verb and are used to reflect time. Blyth [10] defines

Table 2.2.1: Combined noun classes for isiXhosa and isiZulu. The prefix of nouns per class for the two languages is given. There are classes (3a, 9a, 17) that isiXhosa does not have, whereas isiZulu does. (Sources: [110, p.210] [50]).

	Class																	
	1	1a	2	2a	3	3a	4	5	6	7	8	9	9a	10	11	14	15	17
isiXhosa	um	u	aba	oo	um	-	imi	i(li)	ama	isi	izi	in	-	ii	u(lu)	ubu	uku	-
isiZulu	um(u)	u	aba	o	um(u)	u	imi	i(li)	ama	isi	izi	i(n)	i	izi	u(lu)	ubu	uku	ku

these two features by contrasting them since they are closely related. The author points out that tense is a feature that "encodes the time at which an action" [10, p. 57], that is being referred to by the verb, takes place. Aspect, on the other hand, does not pay attention to the "temporal points of reference" [10, p. 57] but is concerned with the ways in which "the internal temporal constituency of a situation" [10, p.57] can be viewed. For instance, the words *ndisahamba* ('I am still walking') and *ndihamba* (I am walking) both describe an action that is happening presently, however, the former word encodes extra information on how the action had been previously occurring. The differences between the two concepts are illustrated in Figure 2.1. The main timeline (black and bold horizontal line marked 'Time') shows the progression of time, with the center most point (labelled B) indicating the present. The elliptical shapes denote areas of a particular tense, from remote past (far left) to remote future (far right). There is an overlap between the areas because in Nguni languages tense is not discrete. The thin black horizontal lines, parallel to the main timeline, denote the various aspects. The progressive aspect is used to indicate that the action referred to by the verb was taking place and still continues to do so. Exclusive aspect indicates that an action that was not taking place before has now started taking place. Imperfect aspect indicates that the action is now is starting to be in effect. It is also used to indicate that an action is habitual. Finally, the perfect aspect is used to show that an action is finished or we have arrived at an action because of previous finished actions.

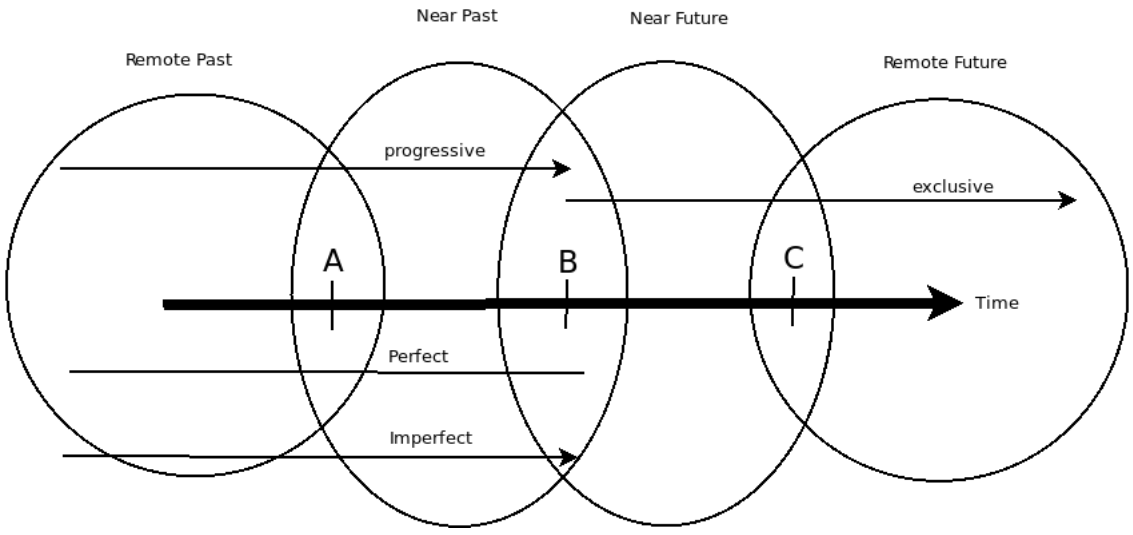


Figure 2.1: Tenses and aspects in isiZulu and isiXhosa. Thin horizontal lines denote aspect. The elliptical areas denote tense. The center-most bold horizontal line denotes the progression of time. The points labelled A, B, and C denote the perceived 'general points' of the past, present, and future (left to right).

Verbal mood, on the hand, refers to the verb's ability to signal the state of reality. The indicative mood is used to indicate a fact, opinion or statement. The subjunctive mood is used to signal the hypothetical action, and the participial mood is used to signal simultaneously occurring actions. Understanding which moods, tenses and aspects are present in real-world weather forecast verbs allows us to consider a sizeable amount of features to regard for the grammar, as full grammars for both languages are impractical (if not impossible) given the scope of this work. Moreover, this restriction is important because isiXhosa and isiZulu are "underdescribed" [73, p.4] thus an attempt to capture all features for this work would have to be preceded by writing a grammar textbook for each language.

IsiXhosa and isiZulu both require a process called phonological conditioning. This is the removal of consecutive vowels in words. Nguni languages do not allow consecutive vowels in words [107]. IsiXhosa, unlike isiZulu, has some exceptions. For instance, borrowed words such as *iorenji* (an orange) [63, p5] are allowed to have consecutive vowels. There are other special cases such as the plurals of nouns, for instance, *ootata* (fathers) is also a valid word. The difference in the number of consecutive vowels between isiXhosa and isiZulu can be seen in the Universal Declaration of Human Rights document where isiZulu has 0 and isiXhosa 30 consecutive vowels [49]. The processes used to resolve consecutive vowels are coalescence, gliding/consonantalization, glide formation and glide deletion, and vowel deletion [107]. For instance, in isiZulu one can use the following coalescence rules $a + i = e$, $a + u = o$, etc [27]. In isiXhosa, one can use the coalescence rules $a + e = e$, $a + o = o$, etc [63].

2.3 Natural language generation

NLG is a field of computer science that concerns itself with the creation of computer systems that are capable of generating text in natural languages such as isiXhosa, Setswana, Yoruba, etc. The text is generated from a predetermined non-linguistic representation [21] and this is done to satisfy some communicative goal [20]. Another way of understanding NLG is to consider the general functions of natural language processing (NLP). NLP is a field within computer science in which computers are used to analyse and process natural languages. The ultimate goal is to create software capable of 'communicating' using and 'interpreting' a natural language. It is for this reason that NLP is thought of as being made up of exactly two sub fields: natural language understanding (NLU) and NLG [21, p.3]. The former is the field where we investigate the methods and tools for mapping or converting natural language to a precise structured representation of the same information which would allow computer interpretation and other forms of computation ('*interpreting*'). NLG, on the hand, is the field in which we study methods for converting such structured computer processable representations of data into natural language text ('*communicating*'). These two fields would intertwine in the possible use case, for instance, where there exists an SQL database that keeps track of the number of purchased foods at a university's food-court/cafeteria. NLU techniques make it possible to translate/map the natural language question "What food is most popular at the university?" to an SQL query that can retrieve the answer. NLG techniques, by contrast, make it possible to generate a natural language sentence such as "The popular food is Cake and is sold at the A shop" that answers the original question when the SQL query returns the following possible fields (Shop_Name='A', Food_Item='Cake'). Individuals familiar

> The month was cooler and drier than average, with the average number of rain days. The total rain for the year so far is well below average. There was rain on every day for eight days from the 11th to the 18th, with mist and fog patches on the 16th and 17th. Rainfall amounts were mostly small, with light winds.

Figure 2.2: An example of a weather summary generated by the WeatherReporter NLG system (Source: Dale and Reiter [21, p.50])

with natural language interfaces for databases may also be familiar this application of both NLU and NLG. The conversion of abstract representations to natural language texts requires a number of intermediate subtasks, and these tasks are discussed in Section 2.3.1 and Section 2.3.2.

2.3.1 Document planning and microplanning

The first task in NLG is the determination of what should be communicated by the output text. This is sometimes influenced by the data source and context in which text is generated [21, p.51]. One needs to be able to determine which information is available as-is in the input, which data is computable, and which is unavailable [21]. Additionally, the data-to-text mapping process also needs to account for the structure of the final output. This is particularly important because text that aims to communicate a point is not arbitrarily ordered [21, p.51]. There needs to be a logical structure to the text. For instance, it is easy to see the relationship between the sentences (labelled 1 below) for a human, unlike a computer. The message communicated by these sentences can only be understood when one takes into account their structure (ordering) and its implication.

(1.) The goat is dead. The goat was sick.

Relationships of this nature in text exist not only between sentences, but they also exist between other types of document components. A popular theory used for such an endeavour is rhetorical structure theory (RST). It was developed around 1980 and it allows the analysis of language components such as sentences, clauses, etc. and the relationships between them [16]. RST takes text, segments it by rhetorical function and creates an ordered tree in which the terminal nodes represent the text parts and inner nodes represent the relationship between their children. It has been used in a number of different fields such as NLG, automatic text analysis, and interpretation [16, p.589]. Essentially, RST was developed to allow the analysis of the structure of text that communicates certain ideas. An example of a text document structure that was provided by Reiter and Dale [21] for the text given in Figure 2.2 is shown in Figure 2.3. There are other theories and techniques by which to perform document structuring [122, p.13-p.31].

Once a structure for the text has been established, further manipulation is necessary. This is to introduce linguistic information into the document structure. This process is called lexicalisation and it is part of a number of processes that are collectively known as micro-planning. Another important process is called aggregation, and it

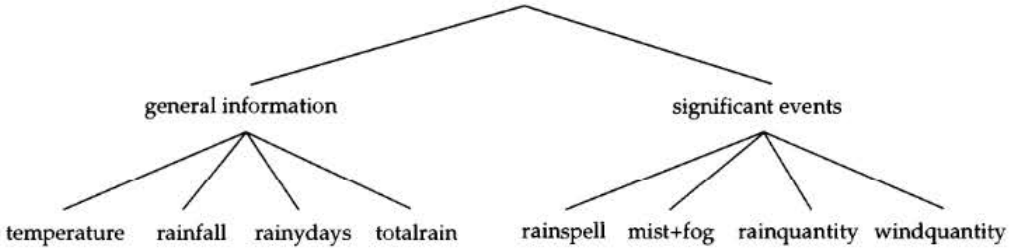


Figure 2.3: Document structure of the weather summary generated by the WeatherReporter NLG system provided in Figure 2.2 (Source: Dale and Reiter [21, p.53]).

is the joining of multiple rhetorical components that would have yielded multiple sentences. For instance, the two sentences "The cabbage is **green**. The cabbage is **fresh**." are better when presented as one unified sentence "The cabbage is **green and fresh**". This process results in concise and human-friendly text that does not unnecessarily repeat details. Furthermore, the structure may also undergo a process called referring expression generation. This is when one decides when to use a noun vs. pronoun. It is important because human authors are less likely to write sentences such as "The goat is dead. **The goat** was sick" as opposed to "The goat is dead. **It** was sick".

2.3.2 Surface realisation

The last process is essentially the transfer of the semantic representations to syntactic structures in order to allow the application of word formation rules that will result in text [21, p.71]. This process is referred to as surface realisation. There are multiple ways in which one can build a realiser. For instance, in the context of dialogue systems (but not limited to it), linguistic realisers can be achieved through templates and grammars [125] and these two approaches are considered to be mutually exclusive. Nonetheless, one can also implement surface realisation through a direct mapping [44, p.16]. Section 2.3.2.1 explains what templates are and how they function. Section 2.3.2.2 provides a short introduction to grammars and discusses the underlying linguistic theory used for modern grammars and their corresponding tools.

2.3.2.1 Templates

Templates, in their simple form, are strings with slots for the various input values. They also allow basic string manipulation. An example of such a system is the way text is manipulated by Microsoft Word's mailmerge.

> Thank you for being part of the community's conservation efforts. Because of your contribution of <insert amount here>, we are pleased offer you a <insert amount here> discount the next time you visit our online store. [129]

There are two slots in the above text and they are enclosed with chevrons. Templates can have a large impact with little effort, especially in providing natural language interfaces to databases. This is illustrated best by the toy

Bus Number	Origin	Destination	Duration
A023	Polokwane	Oudtshoorn	9 hours
B881	Upington	Tshwane	10 hours
C009	Thohoyandou	Umlazi	6 hours

(a) Example of a database table with South African domestic bus schedules.

The bus <bus number> departing from <origin> reaches <destination> in <duration>.

(b) Example of template for describing the bus schedules.

The bus A023 departing from Polokwane reaches Oudtshoorn in 9 hours.

(c) Example of generated sentence.

Figure 2.4: A database, template, and text illustrating template-based surface realisation using a toy example (Adapted from [44, p.20]).

bus schedule database example that is shown in Figure 2.4. The simplicity of templates comes at the cost of maintenance. The requirements for text that is to be generated by NLG systems change over time, albeit slowly, and templates make it difficult to update larger NLG systems [38, p.53]. Furthermore, templates are not compatible with agglutinative languages whose morphology is complex, such as Bantu languages [50] [53]. The maintenance challenge is not unique to template-based systems. The SumTime-Mousam system [115], for instance, does not depend on templates. It uses a simple ordering of lexicalised phrases to realise its output [115]. Nonetheless, it needed to be maintained on a number of fronts ranging from its database (constant updating could have been avoided by decoupling the I/O interface [116, p.4]), user interface, etc [116]. We are not aware of any comparison of the effort required to maintain the various types of NLG systems. A comparison would be very informational because it has been argued that it is not clear as to why templates should always be harder to maintain [24, p.19].

2.3.2.2 Grammars

The perceived "opposite" of template-based generation is "real NLG" [24]. This is the approach that makes use of linguistic knowledge. "Real" NLG systems are easier to maintain and generate high quality text due to their elaborate processes that cater for context planning, sentence planning, and syntactic realisation [91]. The realisation in such systems make use of a natural language's grammar. These grammars are often constrained because full language grammars are complex, and sometimes may not be properly documented. The latter point is especially important if one develops an NLG system while assuming the traditional view of studying a language. This is the view that one requires a large set of grammatical rules, and a corpus of the various patterns that exist for the language [4]. The type of grammar that is used for realisation is highly dependent on the underlying linguistic theory. For instance, one can categorise modern theories of language based on the choice between syntagmatical relations

and paradigmatical relations for linguistic units [4]. If one uses the former, then they are likely to use Markov's finite-state grammar, Chomsky's universal grammar, etc. Otherwise, if one chooses the latter, then they are likely to use Halliday's systematic functional linguistics, for instance [4]. There are multiple popular language modelling theories (and grammars, by consequence) that have been used for generation. For instance, the popular forecast generator (FoG) system made use of Meaning-Text Theory (MTT)[1]. The motivation for early "real NLG" surface realisers such as FUF/SURGE, KPML, RealPRO, NIGEL, and MUMBLE is that they were designed to be broad and domain-independent [21] [32, p20].

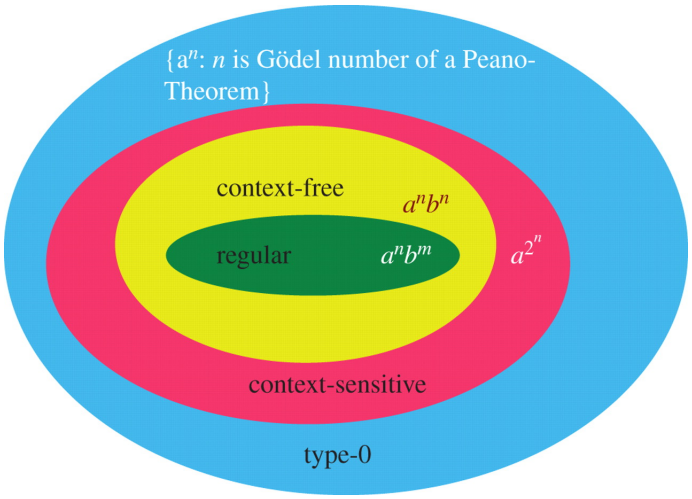


Figure 2.5: The Chomsky–Schützenberger hierarchy for formal languages [46].

The general reusability came at a cost of maintainability hence different approaches were sought. There was the use of large corpora in the creation of probabilistic grammars or reranking a small handcrafted grammar's output using corpus-based statistical information [32], however, these approaches are not suitable for underresouced languages. There was also the shift towards the use of formal language theory (FLT) as it can provide effective results with less development and maintenance effort. FLT defines language as merely a possibly infinite set of strings over some finite alphabet [134]. In simpler terms, this view of language is attractive because it ignores meaning [46]. This means that grammars are concerned with setting out the rules that form strings that constitute a subset of a natural language. This lack of focus on meaning does not mean that it is not important for the strings and their use in the real world. It simply removes the burden of encoding it in the grammar explicitly. A grammar $\mathcal{G}$, in this context, is defined as the quadruple (Σ, NT, S, R) where Σ is the set of terminals, NT is the set of non-terminals, S is the start symbol, and R is the set of rules [46]. There a number of types of languages that can be generated by these formal grammars and these are shown in Figure 2.5. Formal languages are not the same as natural languages and it has been argued, for instance, that a context free grammar cannot fully describe natural languages [106]. Fortunately, both types of languages intersect and grammars need not be able or designed to generate an entire natural language since a carefully chosen subset of a natural language is sufficient. In that spirit,

[1] One of the linguistic theory's foundations is that language is made up of meanings and texts. Meanings are shown with semantic representations and texts are shown through phonetic representations [47]

we make use of a context free grammar for modelling isiXhosa and isiZulu weather verb fragment. Moreover, we take the position that phonological conditioning should be an orthogonal process, that is, morphological alternation should not be embedded in the grammar. This is advantageous as the separation of concerns means that there will be no explicit modelling of certain complexities directly in the grammar. This separation of concerns design is also achieved, to some degree, by the Xerox finite state tools as morphotactics are modelled through *lexc* and morphological alternation through *xfst*. However, this work is interested in software independant models since the resulting morphological generator will be part of a larger realisation engine. The Xerox tools approach, unlike a CFG approach, is not software independent hence may not be fully compatible with the engine. The Xerox tools approach is best for a standalone generator, however, it would make it difficult to pursue an engine that has interactive feedback with respect to information flow, for instance. Keet and Khumalo [52] and Byamugisha et al. [17] have shown that it is feasible to make use of a CFG to model a fragment of the verb.

2.3.3 Natural language libraries

The move away from large coverage and domain-independent surface realisers has had implications on how the realisation process is viewed and should be supported by software. For instance, Gatt and Reiter [33] decomposed the realisation process and clarified that it is made up of two processes. The first process is tactical generation: the making of linguistic choices based on the provided semantic input. The second process is only concerned with "a syntactic representation, applying the right morphological operations, and linearising the sentence" [33]. It is then possible for the latter process to be the responsibility of the so-called realisation engines. These engines are the only parts that are considered to be re-usable. This is a departure from the old thinking of building packages/systems that carry out both processes. SimpleNLG[2] is a Java library that is to be used for linearisation and morphological operations of English. This means that the handling of the input to the surface realisation module is properly decoupled from its function of linearising the semantic input. SimpleNLG has also been adapted to handle other languages [62] [74] [130] [11]. SimpleNLG is not suitable for Bantu languages and it is also not possible to re-use some of its components because of its inherent tight-coupling that is a result of the encoding of English's rules in the programming language.

There has been an engine built based on SimpleNLG's framework for Telegu [30]. It is of particular interest because the Dravidian language is agglutinative and requires agreement between sentence constituents. Dokkara and Sripada [30] store the rules required by the morphology engine in external files. This feature is attractive unlike the approach taken by SimpleNLG and oxyGen of encoding the rules in the programming language because it leads to an adaptable engine. Nonetheless, the paradigm-based approach taken by Dokkara and Sripada [29] [30] is unsuitable for isiZulu and isiXhosa. IsiZulu and isiXhosa verbs cannot be formed by appending three affixes (phonetic alternation, tense mode suffix, personal suffix) as done by Dokkara et al [29] [30]. The two language's verb's morphotactics are more complex than Telugu verbs.

[2] https://github.com/simplenlg/simplenlg

The idea of creating a 'base' or 'engine' that can be reused in systems in multiple domains is not new. For instance, the LOLITA (Large-scale Object-based Linguistic Interactor Translator and Analyser) system that was developed in the early 90's was said to be a "general purpose base"[109] that had been used by a number of prototypes. If one uses so-called engines, they can have greater control of the manipulation of the language's entities. Another resource that acts as a 'library' for a natural language's grammar is the Grammatical Framework (GF) [88]. GF's resource grammar library has also been referred to as a software library for natural languages and compared to Java's API [90]. GF is a formalism that is based on type theory and is used for developing natural language grammars. The term GF is generally used to refer to the theory, programming language, and software package for processing developed grammars [89]. The two main functions of a GF grammar are linearisation and parsing [15] [89]. Ranta [88] views natural language as being made up of an abstract and concrete syntax as shown by the example in Figure 2.6. This means it can be modeled using rules defining the language's hierarchical structure and a collection of the words that exist in the 'language' together with their relationships to the abstract parts [88, p146]. In particular, the abstract syntax is defined by language categories or entities (**cats**) and functions (**funcs**) for building these entities. The concrete syntax of a specific language, on the other hand, is made up of information (**lincats**) required for the linearisation of the language's entities, and ways (**lin**) of combining the various linearisations. GF's two functions are used in a variety of applications. The web service built by Bringert et al. [15], for instance, allowed the construction of a question and answering system by relying on other reasoning technology, and a limited natural language translator that has auto-complete functionality. The framework has not been used to model most Bantu languages. The only Bantu languages, to our knowledge, whose grammars have been developed using GF are kiSwahili [71] and Setswana [83].

The management of grammar rules can be onerous hence GF is modularized. It is built such that its library is made up of two layers, the user API and a core grammar. Moreover, for any language (*Lang*), there exists three modules Syntax*Lang*, Paradigms*Lang*, and Extra*Lang* [90, p.14]. These modules separate the rules that specify the language's syntax, inflection of various entities, and extend the core grammar respectively. Ng'ang'a [71] defines a few concrete and abstract kiSwahili modules. The features that exist in kiSwahili, as identified by Ng'ang'a [71], also exist in the Nguni languages under scrutiny. IsiXhosa and isiZulu nouns have classes, and the verbs in each sentence need to agree with the sentence's object and subject through specific morphemes. Moreover, the sentences for both languages also have kiSwahili's "fairly fixed word order (SVO)" [71, p.217]. The GF grammar is designed such that the kiSwahili verb is inflected for tense, gender (noun class), animacy, and person. Moreover, the verb forms are dependent on a single noun's gender [71] [70]. This is limiting because a kiSwahili verb needs to agree with both subject and object. Moreover, the gender does not provide the concords but a class noun's prefixes [71]. This shows that the parameter system is not designed fully. Most importantly, the author's preliminary work does not present a comprehensive MorphoSwa module that is responsible for constructing verbs. The only MorphoSwa [70] module that has been presented defines functions for generating only pronouns and adjectives. The limited grammar presented by Ng'ang'a [71] [70] does not investigate the limits of GF's morphological paradigms for Bantu languages. GF's biggest limitation when modelling the grammar of 'Bantu' languages is the

```
abstract Hello = {

  flags startcat = Greeting ;

  cat Greeting ; Recipient ;

  fun
    Hello : Recipient -> Greeting ;
    World, Mum, Friends : Recipient ;
}
```

(a) Abstract 'Hello World' grammar module.

```
concrete HelloXh of Hello = {

lincat Greeting, Recipient = {s : Str} ;

lin
  Hello recip = {s = "bhota" ++ recip.s} ;
  World = {s = "lizwe"} ;
  Mum = {s = "mama"} ;
  Friends = {s = "zihlobo"} ;
}
```

(b) Concrete isiXhosa 'Hello World' grammar module.

```
concrete HelloEng of Hello = {

  lincat Greeting, Recipient = {s : Str} ;

  lin
    Hello recip = {s = "hello" ++ recip.s} ;
    World = {s = "world"} ;
    Mum = {s = "mum"} ;
    Friends = {s = "friends"} ;
}
```

(c) Concrete English 'Hello World' grammar module.

Figure 2.6: Illustration of abstract, isiXhosa, and English Grammatical Framework (GF) grammars using the traditional computer science 'Hello World' program (Adapted from Ranta [89]).

implicit assumption that the languages share a significant number of grammatical features with Indo-European languages. There is a reason to believe that we should not be limiting Bantu language modelling by using GF [83, p.163]. This limitation is very clear in the modelling of Setswana tenses using GF [83].

2.3.4 Summary

Authors have presented templates and 'real' NLG as very distinct methods where one is simplistic and the other is complex [24]. Specifically, template-based systems are considered to be application-dependent methods that have no theoretical foundations. 'Real' NLG approaches, on the other hand, are considered to be more general approaches that possess solid theoretical foundations [24]. However, the caricatured versions of both approaches used by authors who advance this argument lie at different ends of the spectrum. There exists systems that belong in the middle and can be considered hybrid [91] [24]. Nonetheless, it has been shown that basic templates are not compatible with Bantu languages [50] [53]. Therefore, a pressing challenge for Bantu languages is the development of grammars and investigating the hybrid techniques that can be used to generate these languages, if any are possible.

2.4 Weather bulletin generation

The automation of the production of weather forecasts, according to the Bateman and Zock [3] list, is the second most popular application of NLG systems. It follows behind health-care/medicine. In particular, the list shows that there are 15 systems that have been built in the weather domain. An updated list of such systems is given in Table 2.4.1. There are two classes of systems that generate weather. There are time series data summarisation systems and traditional NLG systems that do not necessarily perform extensive processing of the input. The second class of systems has access to a few data points, and these are all transferred into the text. Modern systems of the first class generally use the architecture discussed by Reiter [92]. They require significant effort in the interpretation of time-based weather predictions and, at times, signal analysis.

There are different realisation techniques used by existing systems. There are systems that make use of grammars, a few grammar-like rules, statistical methods, templates, and other simple methods. An example of a system that takes a unique approach for its realisation is the Forecast Generator (FoG). It exists within a larger system whose goal was the automation of routine aspects of weather reporting in order to allow forecasters to focus on "scientific questions"[38, p.45]. It is unique due to its use of MTT models for realisation. The SumTime [117] [115] systems, on the other hand, make of a simpler approach to realisation. These systems are built to summarise time-series weather information and they make use of rules to "[order] phrases" [115, p.5]. These rules are derived from an English sublangage called weatherese [96]. This cannot be done in Nguni languages without inflecting the verb phrase because it needs to agree with the implied sentence subject. Another approach can be observed in the work done by Winkler et al. [133] where the authors use a catalogue-based system to achieve a multilingual generation of avalanche warnings. The system generates four languages and uses a collection of sentence templates

Table 2.4.1: List of NLG systems that have been developed to produce weather forecasts

System name	Establishing literature	Realisation method	Languages	Year
WMO-based and NATURAL	[36]	SimpleNLG	English	2016
CBR-METEO	[1]	String manipulation[3]	English	2015
Winkler-Kuhn-Volk's system[4]	[133]	Catalogued phrases	German,French,Italian,English	2014
Zhang-Wu-Gao-Zhao-Lv's system	[137]	Not implemented	Chinese	2011
pCRU	[7]	Statistical methods	Possibly all	2007
SumTime-Mousam	[117]	"Grammar"	English	2003
SumTime	[117]	"Grammar"	English	2001
Mitkov's system	[67] (as cited by [108])	-	-	2001
Autotext[5]	-	-	-	2000
MLWFA	[136]	Grammar	English, German, Chinese	2000
Siren[5]	-	-	-	2000
Scribe[5]	-	-	-	1999
TREND	[14]	FUF/SURGE	English	1998
Multimeteo[5]	-	-	-	1998
ICWF	[101]	Grammar	English	1993
IGEN	[100]	Grammar	English	1992
Kerpedjiev's system	[54]	Grammar	English	1992
Weathra	[108]	Grammar	English, Swedish	1992
FoG	[12]	MTT Models	English, French	1990
MARWORDS	[39]	Grammar	English, French	1988
RAREAS	[57]	-	English, French	1986
Glahn's system	[37]	Templates	English	1970

where each sentence is split into at most 10 segments. This approach is an advanced form of templates or a flexible version of canned text. We can, therefore, deduce that it will suffer the same constraints as templates when it comes to Bantu languages.

There are two prominent challenges when building weather NLG systems : (1) deciding which professional words to use when describing weather concepts, and (2) how to generate text in two languages from the same input. Some authors have dealt with the first challenge by using words that were decided upon by the forecasters [38], and others have relied on corpus analysis [96]. The second challenge has been dealt with by introducing an abstract interlingua that will capture the syntax irrespective of language [38]. This interlingua approach is only possible, however, for languages that have the same hierarchical structure. Another approach towards dealing with the second challenge is the use of separate grammars and lexicons [136]. Lastly, a solution that requires less effort with the increase in the number of languages is the use of statistical methods [7]. However, it can only be effectively used for resourced languages.

NLG systems, at times, need to be able to account for the possible imprecision in weather forecasts, especially systems that generate text for numerous days. The work done by Sripada et al. [118], for instance, for the UK's

[3] Method depends on expert derived rules and a corpus with only one unique idiolect
[4] This is not a traditional NLG system. It is a multilingual weather warning generator.
[5] System is listed in Bateman/Zock list. Unfortunately, online searches did not yield the English paper in which it was presented.

national weather service generates weather predictions for numerous days. Their system accounts for the loss in accuracy in the text it generated to make sure that users of the system are not mislead. This is done by slightly rewording the forecast for days subsequent to the first day of the forecast. For instance, they make use of the word "expected" when discussing the temperature changes for the third day [118, p.3]. It is also possible to account for the loss of accuracy by following World Meteorological Organisation (WMO) guidelines, or studying a corpus of real world forecasts [36]. The latter approach is not easy because Sripada et al. [118] have found that collecting a corpus may be difficult, and one may need to supplement the collected text with a domain language such as weatherese[6]. Nonetheless, most weather NLG systems use a weather corpus for determining informational units communicated by weather reports. The effectiveness of this practice, especially its manual analysis, has been questioned because it can be expensive in some domains due to expert consultations, human errors in the text, and systems that use a corpus can feel restricted (non-portable) [95] [6]. In order to improve the restrictive nature of corpus-based systems, Belz [6] has argued that a corpus should be viewed as being made up of multiple idiolects that are equally valid. We can then make use of statistical methods to favour the 'overlap between idiolects' [6, p.2] without disregarding the others. The goal is not only to build portable systems but also reduce the need for expert consultations. A corpus should not be treated as a gold standard. Therefore, any problems it may have such as spelling errors must be removed. Furthermore, in the event that there is no human authored corpus, NLG system generated corpora should not be used unless the goal is reverse engineering another NLG system. This is particularly important when one uses techniques such as case-based reasoning [1] that put emphasis on 'reusing' forecasts. These system can be built in different ways. The following section will discuss their various NLG system architectures.

2.5 Southern African languages and generation

There are morphological generators that have come out of computational lingistics fields that are not natural language generation. For instance, a number of generators have been by-products of the morphological analyser creation process that makes use of finite state methods. This is because such analysers make use of transducers that "are indifferent as to the direction in which they are applied" [48] thus being able to act as analyers and generators. The creation of these analyser/generators, for the most part, is facilited by the Xerox finite-state tools; xfst, lexc, and twolc [5]. The first South African language generator developed using these tools, to the best of our knowledge, is the ZulMorph prototype [80]. Its verb coverage falls short for weather generation due to only catering for two tenses and ignoring mood and aspect. Moreover, the ZulMorph rules have not been made publicly available. Unlike, Pretorius and Bosch [80, p195] who claim that the object concord is the only affix that is dependent on the class of a sentence's noun[7] for agreement, we consider the verb's subject concord as also dependent on the sentence's subject class. Pretorius and Bosch [80] created a single transducer responsible for generating isiZulu by taking advantage of the observation that "a pair of transducers with one tape in common is equivalent to a single transducer operating on the remaining tape"[48]. The Pretorius and Bosch [80] approach is most suitable for

[6] https://sites.google.com/site/weatherese/home

[7] The object of the sentence in the case of the object concord.

analysers/generators that are standalone and not so for ones that will be part of a realisation engine that will be responsible for choosing the verb's features. It is very difficult to incorporate such generators in an engine. This is an important obversation because the architecture given by Dokkara et al [30] where the rest of the system has direct access to the morphology engine is best.

The general challenge for building computational tools for Southern African langauges is the lack of foundational resources such as corpora. The unavailability of corpora, in particular, has led researchers to investigate different ways of building these computational tools. For instance, Getao and Miriti [35] have suggested the use of machine learning techniques such as reinforcement learning. However, data is still required for machine learning and the amount is dependent on the supervision level on the learning algorithm. Spiegler et al. [113] have investigated the impact of the different levels of supervision in the learning of morphological rules for isiZulu and they have found that supervised approaches perform better than their counterparts. There is a growing number of resources (e.g. Ukwabelana corpus [114] and isiZulu National Corpus and Term Bank [128]) that will enable one to take advantage of data driven approaches. In the absence of resources, one should investigate other options. For instance, when learning a language's morphology one needs to be able to reward agents for producing correct words (e.g. Gikũyũ compounds [35]) hence human judgements can be utilized unless one has an update-to-date computational dictionary. Nevertheless, current South African langauge generation approaches are tend not to make use of machine learning. For instance, the only work whose primary focus is the generation of a Nguni language [52] [50] [53] is CFG and 'pattern'-based. It depends on language expert consultations, and stitching together language information from antiquated literature. A similar approach of language expert consultation/collaboration is seen in the development of Setswana grammar rules using the LFG formalism [8] [9]. Setswana is a language that has a distinctive disjunctive orthography and is spoken in South Africa and Botswana. This collaboration with a linguist can also be see in Pretorius et al. [79] [80] [81] [82] in their development of morphological analysers for Nguni languages. Linguists, especially grammarians, are not at the disposal of every researcher due to not being able to offer sufficient incentives for collaboration hence we aim to slightly differ by minimising the research-linguist collaboration and follow a literature intensive approach.

Keet and Khumalo [50] have focused on the verbalisation of ontologies. They have found that simple traditional NLG approaches are not feasible, and advocate the use of verbalisation patterns for logical constructs such as subsumption, conjunction, etc. In particular, they have found that patterns cannot be realized with a template-based approach [53]. Each logical construct can have more than one pattern, and certain patterns are preferred more than others [51]. The concordial nature of isiZulu has led Keet et al. [17] [52] to make use of grammars in modelling verbs. They also consider fewer tenses thus their work is not sufficient to generate weather verbs. This is because their rules were developed for a different use case and domain. Furthermore, they also consider other features that are not of interest to our work. Pretorius et al. [84] [87] have also built a morphological analyser for another South African language, Setswana. They focused only on the noun and compound nouns. Setswana belongs to a different language group (Sesotho–Setswana), and the most prominent feature that sets it apart from Nguni languages is its disjunctive orthography. They have since expanded their work to also include verbs [85]

[86]. However, the coverage of the verbs' feature includes one tense and does not consider mood. Moreover, bootstrapping language rules from a 'Bantu' language that belongs to a different group may not improve development time. It is possible that it may require more time due to figuring out the intricate differences in the languages, and how that affects bootstrapping. The largest computational grammar for isiZulu, to our knowledge, was developed by Spiegler et al. [114], and they cater for only a few verb tenses [52]. It is also not sufficient for generating weather text because it was developed for a different purpose. Generally, the Bantu language computational resource development approach differs for various contexts. However, most existing grammar development approaches rely on language experts and literature [52].

2.6 IsiXhosa and isiZulu surface realisation

The lack of a large-scale producer of textual weather summaries isiZulu and isiXhosa can be solved by first building a grammar-based surface realiser. A grammar is to be used to achieve this because basic templates are not compatible with the two languages [50] [51]. Moreover, we take the view that there should be a seperation of concerns in the surface realisation process, similar to Gatt and Reiter [34], hence there should a dedicated module for tactical generation and a realisation engine. This is deviation from the old approach to surface realisation. The verb morphological generator (i.e. grammar) should be part of the realisation engine in order have access to verb feature selection capability. A morphological generator needs to have two major functions, that is, morphotactics and alternation [80]. These functions need not be modelled at the same time in the grammar. The modelling of changes in vowels that are surrounded other morphemes that may start or end with vowel (i.e. alternation), for instance, will require a formalism capable of context-sensitivity such as tree-adjoining grammars, combinatorial categorial grammars, or lexical functional grammars. Other formalisms that are much more expressive such as the head-driven phrase structure grammar, or lexical functional grammar can be also used in such a case. However, we take the position that phonological conditioning, and morphophonological alternation in general, should be applied as an orthogonal process in order to separate concerns. This choice also avoids significantly increasing the number of grammar rules. Moreover, this separation of concern means that the grammar will be responsible for morphotactics hence a simplistic approach to encoding language rules should be sufficient. There have been variable approaches that have been used to model segments of Southern African languages over the years. They include type theory and grammatical framework [83], lexical functional grammar and XLE [8] [9], finite state methods and Xerox tools [80], definite clause grammars and Prolog [113], and context free grammars [52]. All these are capable of modelling Nguni language morphotactics, however, the first four are not chosen due to their attachment to specific tools or programming languages, and the expressive power of the lexical functional grammar is more than what is necessary hence we overlook it. We have decided to use a CFG because it is simple and independent of tool and programming language thus allows greater choice for the realisation engine for which it is designed. The realisation engine is the subject of future work and will be not built here. Treating phonological conditioning as an orthogonal process also allows us to create grammars that can be used to quantify its necessity for a limited domain such as weather forecasts.

2.7 Binary similarity measures

Similarity coefficients are developed and used in a number of fields, ranging from botany [97] to software fault localisation [22]. Their function is to determine the similarity of binary feature vectors. In other words, they are used to measure the similarity of 'documents' that have numerous attributes, where each attribute can only have a present/absent value. These 'attributes' manifest themselves in numerous forms. For instance, they can appear in the form of questions [76] whose answers are binary values. There are numerous binary similarity measures. A recent comprehensive survey collects 76 measures and classifies them using hierarchical clustering [19]. Another list of measures, albeit it has fewer measures, is done by Todeschini et al. [126].

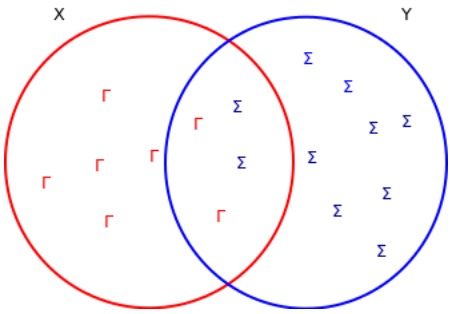


Figure 2.7: Representation of species and their habitat whose co-occurrence can be measured using binary similarly measure. The two regions (X and Y) have two distinct species (Γ and Σ). The intersection of the two circles shows the area in which these species co-exist.

2.7.1 Selected measures

Our work is interested in 4 well-documented coefficients, which are the Jaccard [45], Sorenson [25] (sometimes called Sorenson-Dice or Dice), Driver-Kroeber [31] (sometimes called Ochiai [135]), and Sorgenfrei [111] (as cited by Todeschini et al. [126]) coefficients. We have considered picking numerous different measures, in addition to these four, in order to obtain a broader picture with respect to similarity. More specifically, we attempted picking one measure from each cluster of the dendrogram developed by Choi et al. [19]. This was abandoned due to the strenuous nature of the process of determining details about each measure. These details include whether a measure is normalized, etc. All the four chosen measures are included in the work done by Todeschini et al. [126] in their analysis of binary similarity measures.

The Jaccard index was first introduced by Jaccard as the 'coefficient of community' [45]. It measures the ratio of shared items to the total number of items that exist in two sets. It was originally created to study the distribution of flora in the Alps. The Sorenson measure was developed by Dice [25] using what the author had termed the association index. This index was developed to be used by ecologists to study the association of two species in a geographical region. Specifically, given two species Γ and Σ that exists primarily in two general regions that we

will name X and Y (Illustrated in Figure 2.7), 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. This is represented by the formula $\Sigma \otimes \Gamma = \frac{|X \cap Y|}{|X|}$. 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 devised the coincidence index that would generalize the association index to ensure that association between two species was not variable based on which species is used as the base. This is ratio of the sizes of the shared spaces[8] 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 Driver-Kroeber was developed in ethnology to measure the cultural traits that exist between two groups of people. The Driver-Kroeber measure is a different approach in consolidating what Dice termed the association index of two sets. Driver and Kroeber [31], unlike Dice, 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 for any two sets X and Y is calculated with $\mathrm{DK}(X, Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}}$. The Sorgenfrei metric, on the other hand, consolidates the two association indices by multiplying them together to obtain $\mathrm{Sorg}(X, Y) = \frac{|X \cap Y|^2}{|X||Y|}$.

$$J(A, B) = \frac{a}{a + b + c} \qquad \text{(Jaccard index)}$$

$$S(A, B) = \frac{2a}{2a + b + c} \qquad \text{(Sorenson index)}$$

$$DK(A, B) = \frac{a}{\sqrt{(a + b)(a + c)}} \qquad \text{(Driver-Kroeber)}$$

$$Sorg(A, B) = \frac{a^2}{(a + b)(a + c)} \qquad \text{(Sorgenfrei)}$$

Figure 2.8: Four binary similarity measures used to measure the 'amount' shared items of two sets A and B. The variables (a,b,c,d) are detailed in Section 2.7.2.

2.7.2 Defining measures for language

In order to be able to use these measure for natural languages, we begin by defining, for any two sets A and B, the variables $a = |A \cap B|$, $b = |B - A|$, $c = |A - B|$, and 'd' as the size of the complement of the union of both sets. If we let B be the set of isiZulu generated strings, and A be the set of isiXhosa strings then a is the number of verbs shared by the 'languages', b is the number of verbs that exist in isiZulu but not isiXhosa, c is the number of verbs that exist in isiXhosa but not isiZulu. Lastly, d is the number of strings that do not exist in the two generated sets but exist in isiXhosa and isiZulu. The definition of these variables means that we can rewrite the binary similarity measures discussed in Section 2.7.1 in order to obtain the functions listed in Figure 2.8.

[8] There is one shared space, it is only counted twice.

2.7.3 Measure behaviour

Todeschini et al. [126] investigated the mathematical properties of the similarity measures and any functional inter-relationships they may have. This was done by first generating 100 000 (a,b,c,d) quadruples randomly, with the constraint a + b + c + d = 1024. The authors compared the relationship between the values generated by the measures by creating a Hasse diagram. A Hasse diagram is a tool for illustrating the partial order of a set under a given operation. In particular, the operation that was chosen for the comparison is "coefficient superiority".

Definition (Coefficient superiority.) Let X be a set of (a,b,c,d) quadruples. Let $\mathcal{A}$ and $\mathcal{B}$ be binary similarity measures. We then say $\mathcal{A} \succ \mathcal{B} \iff \mathcal{A}(x,y) \geq \mathcal{B}(x,y) \quad \forall x,y \in X$.

We say that A is superior to B if the binary measure value produced by A for all pairs of sets are larger or equal to the binary measure values generated by B for every such pair. In the Hasse diagram that is generated, if a measure Y is superior to measure Z, it is placed at a higher level in the graph and an edge is drawn between the two sets. This means there is an increase in the measure values for each similarity measure that is applied to the same pair as we move upwards on the Hasse diagram. This relationship is only true for measures that have edges between each other. The levels that exist in the generated Hasse graph range between 1-7 (inclusive). We observe that the Driver-Kroeber and Sorgenfrei measures are at level 5, the Sorenson measure is at level 4, and the Jaccard measure is at level 2. However, there are no direct links between the chosen measures despite them being at different levels (hence we cannot directly compare the measures). We can determine, however, a path that shows that $0 \leq J(X,Y) \leq S(X,Y) \leq DK(X,Y) \leq 1$, for all X, Y. Furthermore, Warrens [131] provides a mathematical proof showing that $0 \leq Sorg(X,Y) \leq J(X,Y) \leq 1$ for any pair of sets, X & Y.

2.7.4 Summary

There are numerous binary similarity measures that exist. These measures have been developed in diverse fields of study. Nonetheless, they can re-purposed to be able to measure similarity between two natural languages. The relation of the four chosen measures to each other is such that they can be equal, however, that is not always the case. They have specific order relations to each other as was revealed in Section 2.7.3. For instance, the Jaccard measure is never greater than the Driver-Kroeber measure.

3

Weather forecast corpus development

This chapter discusses the process of weather text collection, its processing, translation and feature extraction because this work follows a corpus-driven approach to building NLG systems.

3.1 Data collection

The obvious source for the weather text is the South African Weather Service (SAWS)[1]. Unfortunately, initial requests made to the SAWS failed to yield data. A number of alternative avenues were tried in an attempt to obtain the forecasts. These sources were the eNews Channel Africa's weather desk (eNCA), the Met office (United Kingdom's national weather service), the South African broadcasting corporation (SABC) and People's Weather (24/7 climate, environment and weather channel on DSTV[2] channel 180). It was impossible to obtain data from all the aforementioned sources as well. ENCA, at the time when the request was made (August 2016), did not save any written reports of forecasts. Furthermore, the tweets obtained from the twitter account of their weather desk were not suitable for analysis. The text was informal, used twitter conventions such as hashtags, and relied on pictures to convey certain details. Here is a sample of the tweets that were posted between 02 August 2016 and 18 August 2016;

- Decent rains reported this morning in parts of the Karoo. #thunderstorms @eNCA https://t.co/Dc8EhhkMro

[1] http://www.weathersa.co.za/
[2] http://www.dstv.com/

- Not even spring yet and we're already forecasting t-storms! @eNCA https://t.co/zuhJXNs5wh
- What a hot day it's been along the W. Cape south coast! https://t.co/x77zZYr3ey
- Indulge us ladies. Our dashing weathermen dressed alike. Who killed it the most? #WomensDay https://t.co/9Ytg7knxoM
- City forecast for the #LocalElections tomorrow: Cold & windy in Bloem & PE. @eNCA https://t.co/TGM96pImsv

The sampled tweets above show that the account is not used to report daily information formally about the state of the weather, but it is used as a supplementary communication measure between ENCA's weather desk and its audience. The Met office provides textual summaries of historical weather data, alongside other forms of data such as graphs, maps and other data sets through its website[3]. These summaries are of past data and the tense that is used is also reflective of this. Furthermore, the areas that are covered by the Met office's reporting have a different climate from South Africa. The updated Köppen-Geiger climate type maps, as provided by Peel, Finlayson, and Mcmahon [75], show that the locations described by the Met office have a group C climate (Mesothermal climate). This is in contrast with South Africa's variable climate [75, p.467] which can be classified as group B (Arid and semi-arid climate) and group C. This makes the summaries provided by the Met office unsuitable. The SABC, at the time of writing, does not store any historical written reports. They only keep historical televised reports. Efforts to obtain these reports and transcribe them were not fruitful as the SABC charges 320 South African Rands (ZAR) per televised report. This route was not taken due to the excessive costs. The other contacted sources, that are part of the SABC, did not keep such records and others did not reply to our requests.

Further inquiries to the SAWS, directly to the head office, were fruitful. They led to a number of weather forecasts, an example of which is provided as Figure 3.2, that were sampled from the weather reports written in 2015. The complete set of data cannot be provided because of a non-disclosure agreement. The reports were short, and sometimes contained misspellings as they were human authored. The provided sampled set of reports was made up of reports that described the weather for the first day of each month in 2015 (Jan 2015 - Dec 2015). These reports are mainly used by clients of the SAWS in interpreting raw data. They are not made publicly available, as far as we know. The forecasts is a mixture of regional and term-based point forecasts. One is able to find descriptions of the type 1 and 2 as given in Figure 3.1. Regional forecasts are generally hard to wholy generate using an NLG system due to their focus on spatiotemporal changes [93]. Nevertheless, that complexity is not of concern to this work at the moment as we only focus on the verb only. Moreover, the verb we focus on is suitable for both types of forecasts as it was extracted from texts that are a combination of both types. The reports also succinctly reported weather for Namibia, Botswana, Lesotho and Swaziland. The text for the aforementioned regions was not removed as it was short and because the regions do not differ significantly in terms of climate with South Africa hence their presence does not introduce a significant change in the way the weather is described.

[3] http://www.metoffice.gov.uk/climate/uk/summaries

1 Extremely uncomfortable conditions are expected over the eastern interior of the Western Cape, southern parts of the Eastern Cape, eastern parts of KwaZulu-Natal as well as the Lowveld and escarpment of the Limpopo and Mpumalanga Provinces.
2 Partly cloudy and warm with isolated showers and thundershowers

Figure 3.1: Types of sentences in the forecasts produced by the South African Weather Service (SAWS).

3.2 Data cleaning

The sentences in all reports were processed and strings such as "The expected UVB sunburn index", and its variants, were removed should they be present in each report. A sample of the resulting processed sentences is as follows

- Morning fog patches, otherwise partly cloudy and warm with isolated showers and thundershowers, but scattered in the south.
- Cloudy at first with morning fog on the high ground and drizzle patches on the escarpment, otherwise partly cloudy and cool to warm with isolated showers and thundershowers, but scattered over the southern Highveld.
- Cloudy at first with morning fog and drizzle on the escarpment, otherwise partly cloudy and warm with isolated showers and thundershowers in the southwest.
- Partly cloudy and warm with isolated showers and thundershowers but scattered in the east. Morning fog patches are expected in the east and north.
- Partly cloudy and warm with isolated showers and thundershowers but scattered in the west.
- Partly cloudy and warm.

No other forms of processing were necessary. This cleaning came after extracting the text from the provided files.

3.3 Corpus translation

The extraction of grammatical characteristics from an English corpus as done by Byamugisha et al. [17] where they analyse 45 English medication prescriptions makes the assumption that the grammatical features that exist in the English corpus exist, as is, in its Runyankore equivalent. This assumption need not hold as these languages are different and belong to different language groups. The translation of the English weather corpus to isiXhosa is therefore deemed necessary to obtain the grammatical characteristics of the gathered corpus. A subset of the English weather corpus is selected for translation. The selection was done by randomly shuffling the set of processed sentences from each English report using Python's "shuffle" method from the "random" package. Four sentences were selected from each randomised collection (that is, 4 sentences from each month). This resulted in a total of 48 sentences that were translated into isiXhosa by members of the School of African Languages and Literature at the University of Cape Town. The sentences are not translated into isiZulu due to time and financial constraints. We operate under the assumption that isiZulu and isiXhosa, as they belong to th same Nguni language group, will

THE REGIONAL WEATHER FORECAST FOR TODAY: 2015-05-01 ISSUED AT 05:00 SAST BY THE SOUTH AFRICAN WEATHER SERVICE. THIS FORECAST WILL BE UPDATED AT 16:00 SAST.
SEVERE WEATHER ALERTS:

WARNINGS: Nil

WATCHES: Nil.

SPECIAL WEATHER ADVISORIES: Nil.

GAUTENG: Fine and warm but cool in the south.

The expected UVB sunburn Index: Moderate

MPUMALANGA: Partly cloudy in the east with morning drizzle and fog patches on the escarpment, otherwise fine and cool but warm in the Lowveld.

LIMPOPO: Partly cloudy in the east with morning fog patches and drizzle on the escarpment, otherwise fine and warm but cool on the high grounds.

NORTH-WEST PROVINCE: Fine and warm.

FREE STATE: Fine and cool.

NORTHERN CAPE: Fine and cool, but warm in the west and in the north becoming partly cloudy in the afternoon. The wind along the coast will be fresh to strong southerly to south-easterly.

WESTERN CAPE: Cloudy along the south coast at first, otherwise partly cloudy and cool to warm. Morning fog patches is expected along the south-west coast. The wind along the coast will be moderate southerly to south-easterly along the west coast, otherwise moderate westerly to south-westerly along the south coast in the afternoon. The expected UVB sunburn Index: Moderate

WESTERN HALF OF THE EASTERN CAPE: Partly cloudy and cool. The wind along the coast will be light to moderate westerly to south-westerly.

EASTERN HALF OF THE EASTERN CAPE: Fine and cool. The wind along the coast will be moderate westerly to south-westerly.

KWAZULU-NATAL: Partly cloudy with morning fog patches in the north, otherwise fine and cool to warm. The wind along the coast will be moderate southwesterly, becoming southeasterly.

The expected UVB sunburn Index: Extreme

NAMIBIA: Partly cloudy in the north-east, otherwise fine and warm. It will be cool along the coast with morning fog patches. The wind along the coast will be light to moderate south-westerly but moderate to fresh southerly to south-easterly in the south.

BOTSWANA: Fine and warm, but hot in the extreme east.

LESOTHO: Fine and cool.

SWAZILAND: Partly cloudy and warm.

Visit our website at http://www.weathersa.co.za

Figure 3.2: South African weather report produced by the South African Weather Service (SAWS). Forecast was valid for 01 May 2015.

share the verb features in such a small domain. A sample of the translated sentences is given below (A complete set of translated sentences is provided in Appendix A);

- *Liyakuthi gqabagqaba ngamafu kwaye lipholile* (Partly cloudy and cool)
- *Lipholile kumkhwezo wonxweme apho kulindeleke izibhaxu zenkungu yakusasa ngaphaya koko liyakuthi gqabagqaba ngamafu kwaye libeshushu okanye litshise kwaye libeneziphango ezithe saa emantla.* (Cool along the coast where morning fog patches are expected, otherwise partly cloudy and warm to hot with isolated thundershowers in the north)
- *Inkungu yakusasa embindini, ngaphaya koko liyakuthi fakafaka ngamafu kwaye liphole lide libande, ze lizole kwaye libande.* (Morning fog over the interior, otherwise partly cloudy and cool to cold, becoming fine and cold.)
- *Umoya kumkhwezo wonxweme uyakubaphakathi ukuya kumoya ohlaziyayo ovela ngakumantla empuma.* (The wind along the coast will be moderate to fresh northeasterly.)

3.4 IsiXhosa verb analysis

Manual analysis of the translations shows that there are 53 verbs, with only 27 unique verbs. We define a unique verb as being a unique string, without considering its semantic meaning. It is, therefore, possible to have two verbs with the same meaning, for instance, *litshise* and *litshisa* are considered different in the aforementioned set, but they both mean 'it is hot'. These unique verbs are listed in Table 3.4.1. The spread of the verbs based on mood is such that we have 22 indicative, 2 participial, and 3 subjunctive. The verb forms that are associated with the weather are perfect, causative, neuter, and reciprocical. However, we disregard reciprocity in favour of the intensive form because the reciprocity is used on verbs that refer to information we do not have. This is not unusual as a collected corpus may contain unavailable data [21]. As a starting point, we choose two verbal aspects to consider for the grammar, the progressive and exclusive aspect. These allow the ability to build a rule-set capable of considering, to some degree, the past and future state of the weather. In order to be able to generate both traditional weather forecasts and time series summaries, we consider three tenses: past, present, and future. All these verb features are used in the creation of context free grammar rules in Chapter 4.

Table 3.4.1: IsiXhosa verbs found in the translated weather forecast text. The mood, stem, and English translation of the verb is provided.

IsiXhosa string	Root	Mood	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

4

Development of verb context free rules

This section begins by outlining the incremental approach taken in the development of verbal rules. Section 4.2 discusses the suffixal components of verbs for isiXhosa and isiZulu. They operate in the same manner for both languages. Lastly, we discuss the specific process of development of the isiXhosa and isiZulu verbal rules in Section 4.3 and Section 4.4. The context free grammar rules for isiXhosa rules will be marked with (X*i*.), isiZulu rules with (Z*i*.), and shared rules with (XZ*i*.) where *i* is an integer greater than zero.

4.1 Verb rule development process

In this work, the verb is broken up into 5 building blocks (or is considered as being made up of 5 slots) for both isiZulu and isiXhosa. It is possible to break down verbs for Bantu languages further, for instance, Nurse [72, p.32-p39] considers the verb as being made up of 11 building blocks. Other detailed slot systems have been shown by Maho (as referenced by [55, p.78]) and Khumalo [55, p.79] in their work for isiNdebele. The verb components we consider are the prefix, object concord (OC), verb root (VR), pre-final vowel suffix (S_{suffix}) and final vowel (FV). The rules, for both languages, are developed incrementally, with each increment aligning to the addition of the following building block: (1) prefix, (2) prefix + OC + VR + S_{suffix}, and (3) all five slots. The verb root and pre-final vowel suffix are not added to the rules in a separate incremental step as they present no hardships due to restrictions placed upon the S_{suffix}. These restrictions are described in Section 4.2.1. The steps per incremental stage are as follows:

- Increment 0: Prefix

1. Gathering preliminary rules
2. Verb generation, correctness classification, and elimination of incorrect verbs.

- Increment 1: Prefix + OC + VR + S_{suffix}
  1. Suffix addition, verb generation and correctness classification
  2. Elimination of incorrect verbs, verb generation and correctness classification
- Increment 2: Complete verbs
  1. Investigate missing features, add missing features (where necessary), add final vowel, correctness classification
  2. Elimination of incorrect verbs, verb generation and correctness classification

The verb information for each language was taken from multiple sources and this necessitated a preliminary step of gathering this information into one informal rule set. In isiXhosa, for instance, McLaren [65, p.82] does include the object concord unlike Davey [23, p.33] when presenting how present indicative verbs are formed. Furthermore, Davey [23, p.39] includes the tokens generated by Rule (XZ11) when presenting present participial verb formation unlike McLaren [65, p.89]. The merging processes is discussed in '*Increment 0: Prefix*' for each of the two languages. In each of the listed increment, generated words are classified by syntactic and semantic correctness. All the undesirable words among the generated words are removed. This classification of words is done by a single individual (the researcher) who is an isiXhosa first language speaker.

4.2 IsiXhosa and isiZulu shared verb component rules

This section discusses the components of the verb that behave in the same manner for isiZulu and isiXhosa. These components are the verb extensions and the final vowel.

4.2.1 Verb extensions

The rules for the pre-final vowel suffix are easily handled because we will focus on the generation and not consider any special cases, as a starting point. Additionally, we shall not be considering all the possible forms of verb extensions. An example of an ignored special case is illustrated by Doke [27] concerning causative extension (rule XZ3), in the event that a stem ends with a 'k' or 'l' then there are modifications that are expected. 'k' should be replaced with 's' and 'l' should be replaced with 'z'. There are exceptions with words that end with an 'l'. The last letter in the following stems should not change. (fudumal-, khukhumal-, thwal-) [27, p.145-p.147]. The rules regarding the generation of the verb extensions are as follows:

(XZ1.) $S_N \rightarrow ek$ (Neuter)
(XZ2.) $S_I \rightarrow isis$ (Intensive)
(XZ3.) $S_C \rightarrow is$ (Causative)
(XZ4.) $S_p \rightarrow il$ (Perfect)

(XZ5.) $S_o \rightarrow S_C|S_I$
(XZ6.) $S_{nic} \rightarrow S_N|S_o|S_N \quad S_o$
(XZ7.) $S_{np} \rightarrow S_{nic}|\varepsilon$
(XZ8.) $S_{suffix} \rightarrow S_{nic}|S_p|\varepsilon$

4.2.2 Final Vowel

The final vowel is implemented through a series of 'if' statements, as shown in Algorithm 1. This approach ensures brevity of the context free grammar rules. Perfectness is shown through the use of an optional verb extension and the final vowel 'e'. The pseudo-code that will follow makes use of the following abbreviations; (fv, Final vowel), (S_p, Pre-final vowel perfect suffix), (PC/PC_1, Present continuous formatives), (A_{sa}, Spaced aspect), (A_e, Compound exclusive aspect) and (NPC_2, non-present continuity). The values of these abbreviations can be found in Figure 4.4.

Algorithm 1 If statements that generate the final vowel of a verb for isiXhosa and isiZulu. The function *is*() is used to check whether the verb is perfect or is its tense is the future tense.

if has(fv) **then return** ''
else
  if *is*(*perfect*) or *has*(S_p) and ($PC|PC_1$)! = *ya* and *isNotEmpty*(A_{sa}) **then return** 'e'
  else if *has*(A_e) **then return** 'a'
  else if *is*(*future*) and *has*(NPC_2) **then return** 'e'
  elsereturn 'a'
  end if
end if

4.3 IsiXhosa verb rules

This section begins with a discussion about a number of isiXhosa features and this is followed by a description of the isiXhosa context free rule development using the approach described in Section 4.1.

4.3.1 IsiXhosa features

The verb in isiXhosa can be inflected with respect to tense, voice, mood and form according to McLaren [64]. Form refers to the various verbal derivatives such as reciprocity, causativity, etc. These are often referred to as verb extensions. In this work, we shall not consider the inflection of the verb with respect to the voice. There are two special concords to consider when dealing with verbs, the subject (SC) and object concord (OC). The subject concord form part of what this work has named the prefix.

(X1.) $OC \rightarrow m|ba|wu|yi|li|wa|si|zi|lu|bu|ku|\varepsilon$

(X2.) $SC \rightarrow ndi|si|u|ba|i|li|a|si|zi|lu|bu|ku|\varepsilon$

These concords are dependent on the verb's noun class, table 4.4.1 lists all the object and subject concords with their respective noun class. The various isiXhosa noun classes are provided in Table 2.2.1. There is a special modification that happens to the subject concord when it is prefixed by the negative prefix ($NP_{pref} \rightarrow a$). However,

Primary

- Present
- Perfect/Near-past
- Remote-past
- Future

Secondary/Compound

- Near-past-progressive
- Near-past-perfect
- Future-in-the-near-past
- Remote-past-progressive
- Remote-past-perfect
- Future-in-the-remote-past

Figure 4.1: List of 10 isiXhosa tenses as provided by McLaren [64, p.81]. They are separated into primary and secondary categories.

we shall not be considering that as we are not considering negated verbs. This has the implication that we shall not be considering the effect of negation on the final vowel.

There has not been a consensus about the number of tenses in isiXhosa. Recent work maintains that there are only 5 tenses, however, some older work [64] [65] has suggested that there as many as 10 tenses (or 16 tenses [77]) in isiXhosa as listed in Figure 4.1. It should be noted that these works make a distinction between 'primary' and 'compound/secondary' tenses. There are only four primary tenses [64, p.81], and they are shown in Figure 4.1. Peters [77] concedes that some of the so-called primary tenses the author uses could be considered moods. Davey [23], on the other hand, splits the future tense into two separate tenses thus obtaining five primary tenses. These five are the ones shown in Figure 2.1. The five primary tenses are also subject to certain restrictions based on the mood. For instance, in their works (Davey [23] & McLaren [64] [65]) we see that the indicative and participial moods do not have distinct past tenses. That is, the remote and near past are equivalent. Furthermore, they use perfect verbs to indicate the past. There is some work that has never made a distinction between perfect and near-past tense to begin with [63]. The subjunctive mood, similarly, has only one past tense - it does not draw distinctions between the remote and immediate past [64]. Mncube [68] also reveals that the present tense in the subjunctive mood is special. It is frequently called the present tense, however, its formal name is the present-future tense. This is because the tense is compound-like, as it is the present but "usually has a future intent" [68, p.84].

Verb continuity in isiXhosa past tenses is achieved by prefixing the verb with 'be' or 'b' [103, p.193]. This is not consistent, however, with what is presented by Mncube [68] as they show that the continuity for the past tense in the subjunctive mood is achieved through 'ye'. Davey [23] loosely agrees with Schonstein [103] with respect with the future tenses, as they use "ba/be" to signal continuity. Further analysis of Davey's [23] discussion on compound tenses reveals that "ye" and "be" are interchangeable for the past tenses.

4.3.2 Increment 0 : Verbal prefix composition

There is no fixed location for all types of features in the various verb components. For instance, aspect tokens are not all found in the same position. They can precede or succeed the subject concord within the verb prefix. Nonetheless, the general constituents of a verb in Bantu languages are known [66, p.108-p.111] [72, p.32-p39].

The positions of the various constituents, as we have found them in the literature for isiXhosa, are listed in the preliminary rules below. These are preliminary prefixes that are built based on our study of the presentation made by McLaren [64] [65] and Davey [23] about the nature of the verb in isiXhosa. The base isiXhosa rules required for constructing the prefix are listed in Figures 4.2 and 4.3. In the preliminary rules, the superscript ($[0..1]$) is used to denote the presence or absence of a particular item.

INDICATIVE MOOD

- $P_{prefix} \rightarrow NPC^{[0..1]} \quad E^{[0..1]} \quad SC \quad P^{[0..1]}$ (Immediate past)
- $P_{prefix} \rightarrow E^{[0..1]} \quad SC \quad PCP_1^{[0..1]}$ (Present)
- $P_{prefix} \rightarrow SC \quad IM \quad F \quad C \quad S_{space} \quad E^{[0..1]} \quad SC \quad P^{[0..1]}$ (Immediate future)

PARTICIPIAL MOOD

- $P_{prefix} \rightarrow NPC^{[0..1]} \quad E^{[0..1]} \quad SC \quad P^{[0..1]}$ (Immediate past)
- $P_{prefix} \rightarrow E^{[0..1]} \quad SC \quad PCP_1^{[0..1]} \quad PMF^{[0..1]}$ (Present)
- $P_{prefix} \rightarrow SC \quad IM \quad F \quad C \quad S_{space} \quad E^{[0..1]} \quad SC \quad P^{[0..1]}$ (Immediate future)

SUBJUNCTIVE MOOD

- $P_{prefix} \rightarrow NPC^{[0..1]} \quad E^{[0..1]} \quad SC \quad PRP^{[0..1]}$ (Immediate past)
- $P_{prefix} \rightarrow E^{[0..1]} \quad SC \quad PCP_1^{[0..1]}$ (Present)
- $P_{prefix} \rightarrow SC \quad IM \quad F \quad C \quad S_{space} \quad E^{[0..1]} \quad SC \quad P^{[0..1]}$ (Immediate future)

The preliminary rules were re-written to eliminate the use of the superscript. The goal is the formalisation of the pseudo-rules. This allows us to list the rules in their numbers, with the various restrictions among the constituents being explicit. After this formalisation process, the preliminary prefix rules are tested, in their entirety, and a number of inconsistencies were found and corrected. This was done by generating the prefixes that are defined by the rules. Each generated string was marked with *True* (if the prefix is syntactically correct) and with *False* (if the prefix does not exist in isiXhosa). The generation was first done manually because it was assumed that it would be excessive to use an automated generator for the relatively few rules. However, this assumption was found to be false due to the tedious and laborious nature of the manual generation process. The rules were updated to prevent them from producing the prefixes marked with false. The resulting rules are listed below (X6 to X14).

(X3.)	$PCP_1 \rightarrow C \mid P$	(Continuity/Progressive aspect)
(X4.)	$PRP \rightarrow P \mid PR$	(Past remote/progressive aspect)
(X5.)	$F \rightarrow ku$	(morpheme used to indicate the future)

Figure 4.2: Context free grammar rules for generating morphemes used to indicate aspect and tense in the isiXhosa verb prefix.

Indicative mood

(X6.) $P_{prefix} \rightarrow NPC \quad A_{pes}$ (Immediate past)

(X7.) $P_{prefix} \rightarrow A_{pes} \quad PC$ (Present)

(X8.) $P_{prefix} \rightarrow SC \quad IM \quad F \quad A_{sa}$ (Immediate future)

Participial mood

(X9.) $P_{prefix} \rightarrow NPC \quad A_{pes}$ (Immediate past)

(X10.) $P_{prefix} \rightarrow A_{pes} \quad PC \quad PMF$ (Present)

(X11.) $P_{prefix} \rightarrow SC \quad IM \quad F \quad A_{sa}$ (Immediate future)

Subjunctive mood

(X12.) $P_{prefix} \rightarrow NPC \quad A_p \quad | \quad NPC \quad A_{es} \quad PR$ (Past)

(X13.) $P_{prefix} \rightarrow A_{es} \quad PC \quad | \quad A_{ps}$ (Present)

(X14.) $P_{prefix} \rightarrow SC \quad IM \quad F \quad A_{sa}$ (Immediate future)

The above rules (X6 to X14) make use of the rules for the 'compounded' exclusive and progressive aspects signaled by A_e, A_p and their combined rule as A_{pes} where 'PES' stands for progressive, exclusive and simple. These 'compounded' rules are defined in Figure 4.3.

4.3.3 Increment 1 : Verbal prefix and pre-final vowel suffix

The prefix rules (X6 to X14) were extended through the addition of the object concords (OC), verb root (VR), and the pre-final vowel suffix (S_{suffix}). The method by which the S_{suffix} is generated is detailed in Section 4.2. A set of strings was generated with those rules using the CFG module in the natural language toolkit (NLTK)[1]. A verb stem was chosen randomly (through the Python 'random' package) from the pool of strings listed in Table 3.4.1. The chosen verb root is *-zol-* and can be loosely translated to 'quiet/calm down'. The noun *izulu* was picked as a subject, as a starting point, and for simplicity, we used an empty object. The subject is in noun class 5 and therefore its subject concord is '*li*' (see Table 4.4.1). A hypothetical generated sentence that would require the chosen pair of concords is the following: *izulu liyazola* (The sky is quieting down). There are other sentences that could make use of this verb and require these concords. The final vowel was added manually when validating the correctness of the generated verbs as the rules did not include the final vowel. This means that the final set of validated verbs did contain duplicates. The resulting strings were classified into three variants, correct words (syntax and semantics), syntactically correct strings that are not used in the language, and non-existent/invalid words. The second class of words are not eliminated when evaluating the grammar. The grammar need not be prevented from generating such words because NLG system modules that use the grammar engine can easily prohibit the use of verb features that produce such words. The elimination of that class of words represents a semantic restriction [81, p.211] and can be implemented through a technique capable of discovering a verb's meaning and use it to determine whether the verb is compatible with certain verb extensions. We take the position that such restrictions should be the

[1] http://www.nltk.org/

responsibility of the realisation engine and should not be placed in the grammar. The following sections detail the outcomes for each mood-specific rule set.

Table 4.3.1: Number of correct and incorrect words generated using the first increment isiXhosa grammar (participial mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	55.6%	10	8	18
	Semantics	55.6%	10	8	18
Present	Syntax	28.7%	31	77	108
	Semantics	14.8%	16	92	108
Future	Syntax	53.1%	69	61	130
	Semantics	21.5%	28	102	130

Table 4.3.2: Number of correct and incorrect words generated using the first increment isiXhosa grammar (indicative mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	55.6%	10	8	18
	Semantics	55.6%	10	8	18
Present	Syntax	49.1%	53	55	108
	Semantics	15.7%	17	91	108
Future	Syntax	53.1%	69	61	130
	Semantics	21.5%	28	102	130

4.3.3.1 Indicative and participial

The results in correctness of the generated verbs after the addition of the OC, VR, and S_{suffix} to the existing rule set yielded the results listed in Table 4.3.1 (Participial) and Table 4.3.2 (Indicative). Examples of incorrect words (with the manually added final vowel) that were generated include *lizakuzolile*, *balizolile* and examples of correct words include *belisazolile*, *lisazolise*. Analysis of the participial mood's results showed that the low quality is due to the presence of the *PMF* morpheme within the rule for generating present tense verbs. Its presence made all the generated strings incorrect (syntactically and semantically). Our conclusion is that it should not be despite grammar books claiming it is needed. Its removal means that the rules for generating the indicative and participial moods become equivalent. The verbs were analysed further and we observed that the presence of the non-present continuous morpheme value '*ba*' resulted in incorrect words, the pre-final vowel perfect suffix cannot be used in the past tense jointly with the simple aspect. Furthermore, the use of the present continuity morpheme in the prefix when the progressive aspect is also present in the present tense results in incorrect words. There were a few other evident inconsistencies that were observed. All these inconsistencies were eliminated and the rules were updated accordingly. The updated rules for both the participial and indicative moods are as follows.

Past

(X15.) $Verb \rightarrow NPC_o \quad A_{pe} \quad OC \quad VR \quad S_p$

(X16.) $Verb \rightarrow NPC_o \quad SC \quad OC \quad VR$

Present

(X17.) $Prefix \rightarrow SC \quad PC$

(X18.) $Prefix \rightarrow A_{pe}$

(X19.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{nic}$

(X20.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_p$

Future

(X21.) $Prefix \rightarrow SC \quad IM \quad F \quad A_{sa}$

(X22.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{nic}$

The correctness results obtained from evaluating the updated rules are provided in Table 4.3.3. The results for the various tenses saw increases in syntactical correctness. The average increases for the two moods were 24.4% for the past tense, 51.7% for the present tense, and 46% for the future tense. This correctness improvement was accompanied by a decrease in the number of generated words. The past tense decreased by 8, present tense by 76, and the future tense by 88 for both moods.

Table 4.3.3: Number of correct and incorrect words generated using the second increment isiXhosa grammar (indicative and participial mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	80.0%	8	2	10
	Semantics	80.0%	8	2	10
Present	Syntax	90.6%	29	3	32
	Semantics	56.2%	18	14	32
Future	Syntax	100.0%	42	0	42
	Semantics	50.0%	21	21	42

4.3.3.2 Subjunctive

The results in the correctness after the addition of the OC, VR, and S_{suffix} to the existing rule for the subjunctive mood are listed in Table 4.3.4. Examples of incorrect words (with the manually added final vowel) that were generated include *beliazolile*, *baseliazola* and examples of correct words include *lizakube lizolile*, *seliyazola*. We observed that the presence of the of the simple aspect together with the PR morpheme and pre-final vowel perfect suffix all at the same time results in incorrect words for the past tense. Also, the value of '*ba*' for the non-present continuous morpheme also results in incorrect words for the past tense. This is despite being given by a grammar textbook as a valid value for the non-present continuous morpheme. The progressive aspect cannot be used with the perfect pre-final vowel suffix in the past tense. Finally, the perfect pre-final vowel suffix should not be used in the event that the spaced aspect is an empty string for the future tense as they result in incorrect words. These restrictions pertaining to how certain constituents should not be used at the same time is due to the formalisation process. All these inconsistencies were addressed to obtain the following updated rules;

Past

(X23.) $Verb \rightarrow NPC_o \quad A_p \quad OC \quad VR \quad S_p$

(X24.) $Verb \rightarrow NPC_o \quad A_{es} \quad PR \quad OC \quad VR$

Present

(X25.) $Prefix \rightarrow A_{es} \quad PC$

(X26.) $Prefix \rightarrow A_{ps}$

(X27.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{nic}$

Future

(X28.) $Prefix \rightarrow SC \quad IM \quad F \quad A_{sa}$

(X29.) $Verb \rightarrow Prefix \quad OC \quad VR \quad Suffix$

The updated rules yield the results listed in Table 4.3.5. The updates lead to an increase in syntactical correctness of the past tense verbs. The correctness rose by 48.6%, however, the number of generated verbs dropped by 10.

Table 4.3.4: Number of correct and incorrect words generated using the first increment isiXhosa grammar (subjunctive mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	38.9%	7	11	18
	Semantics	38.9%	7	11	18
Present	Syntax	100.0%	30	0	30
	Semantics	50.0%	15	15	30
Future	Syntax	98.0%	48	1	49
	Semantics	53.1%	26	23	49

Table 4.3.5: Number of correct and incorrect words generated using the second increment isiXhosa grammar (subjunctive mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	87.5%	7	1	8
	Semantics	87.5%	7	1	8
Present	Syntax	100.0%	30	0	30
	Semantics	50.0%	15	15	30
Future	Syntax	98.0%	48	1	49
	Semantics	53.1%	26	23	49

4.3.4 Increment 2 : Complete verbal rules

The various inconsistencies that were found for the three moods on the previous two increments necessitated an update of the rules that were listed in Figure 4.3 and Figure 4.2. The rules in Figure 4.2 were reduced to the singular rule $F \rightarrow ku$ as the other rules were no longer needed. The rules listed in Figure 4.3 were updated to obtain the rules listed in Figure 4.4. Furthermore, the presence of the various verbal features denoting continuity, aspect and other notions in each rule set was interrogated, and the status of each feature in the various rule sets can be found

(XZ9.) $IM \rightarrow za$	Immediate future (XZ9.)
(XZ10.) $S_{space} \rightarrow$ ‘ ’	Single space (XZ10.)
(XZ11.) $PMF \rightarrow si\|s\|\varepsilon$	Participial mood morpheme (XZ11.)
(XZ12.) $FV \rightarrow a\|e$	Final Vowel (XZ12.)
(XZ13.) $A_e \rightarrow E \quad SC$	Compound exclusive aspect (XZ13.)
(XZ14.) $A_p \rightarrow SC \quad P$	Compound progressive aspect (XZ14.)
(XZ15.) $A_{pes} \rightarrow A_p\|A_e\|SC$	Compound exclusive, progressive or simple aspect (XZ15.)
(XZ16.) $A_{es} \rightarrow A_e\|SC$	Compound exclusive or simple aspect (XZ16.)
(XZ17.) $S_p \rightarrow il$	Perfect suffix (XZ17.)
(XZ18.) $P \rightarrow sa$	Progressive aspect (XZ18.)
(XZ19.) $E \rightarrow se$	Exclusive aspect (XZ19.)
(XZ20.) $PR \rightarrow a$	Past remote (XZ20.)
(XZ21.) $PC \rightarrow ya\|\varepsilon$	Present continuous (XZ21.)
(XZ22.) $NPC \rightarrow be\|ba\|\varepsilon$	Non-present continuous (XZ22.)
(XZ23.) $A_{sa} \rightarrow C \quad S_{space} \quad A_{pes} \quad \| \quad \varepsilon$	Spaced aspect morpheme (XZ23.)
(XZ24.) $C \rightarrow PC\|NPC\|\varepsilon$	Continuity (XZ24.)

Figure 4.3: Figure with common morpheme rules for isiXhosa and isiZulu. Adapted from list provided by Doke [26, p.271]

(XZ25.) $E \rightarrow se$	Exclusive aspect (XZ25.)
(XZ26.) $P \rightarrow sa$	Progressive aspect (XZ26.)
(XZ27.) $PC_1 \rightarrow ya$	Prelim. Present Continuous (XZ27.)
(XZ28.) $PC \rightarrow PC_1 \quad \| \quad \varepsilon$	Present Continuous (XZ28.)
(XZ29.) $NPC_0 \rightarrow NPC_2 \quad \| \quad \varepsilon$	Prelim. Non-present Continuous 0 (XZ29.)
(XZ30.) $NPC_1 \rightarrow ba \quad \| \quad NPC_2$	Prelim. Non-present Continuous 1 (XZ30.)
(XZ31.) $NPC_2 \rightarrow be$	Prelim. Non-present Continuous 2 (XZ31.)
(XZ32.) $NPC \rightarrow NPC_1 \quad \| \quad \varepsilon$	Non-present continuous (XZ32.)
(XZ33.) $PR \rightarrow a$	Past Remote (XZ33.)
(XZ34.) $IM \rightarrow za$	Immediate future (XZ34.)
(XZ35.) $S_{space} \rightarrow$ ‘ ’	Single space (XZ35.)
(XZ36.) $A_p \rightarrow SC \quad P$	Compound progressive aspect (XZ36.)
(XZ37.) $A_e \rightarrow E \quad SC$	Compound exclusive aspect (XZ37.)
(XZ38.) $A_{es} \rightarrow A_e \quad \| \quad SC$	Exclusive or simple aspect (XZ38.)
(XZ39.) $A_{ps} \rightarrow A_p \quad \| \quad SC$	Progressive or simple aspect (XZ39.)
(XZ40.) $A_{pe} \rightarrow A_p \quad \| \quad A_e$	Exclusive or progressive aspect (XZ40.)
(XZ41.) $A_{pes} \rightarrow A_p \quad \| \quad A_e\| \quad SC$	Exclusive, progressive or simple aspect (XZ41.)
(XZ42.) $A_{sa} \rightarrow NPC_1 \quad S_{space} \quad A_{pes} \quad \| \quad \varepsilon$	Spaced aspect formative (XZ42.)
(XZ43.) $C \rightarrow PC\|NPC\|\varepsilon$	Continuity (XZ43.)

Figure 4.4: Updated context free grammar rules for isiXhosa and isiZulu common morphemes.

in Table 4.3.8. The goal of this exercise is to determine whether the rules are capable of integrating/generating that particular feature. The verb rules that had missing features were updated to include the missing features. Moreover, the final vowel was incorporated into the rules using the technique defined in Section 4.2.2. The technique of verb generation with the CFG module of NLTK and syntactic and semantic classification was carried out. The results obtained are given in Table 4.3.6 (Indicative and participial) and Table 4.3.7 (Subjunctive). The two tables show that the generated verbs have high syntactical correctness values ranging between 80% and 100% (Inclusive) for the various tenses.

Table 4.3.6: Number of correct and incorrect words generated using the third increment isiXhosa grammar (indicative and participial mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	97.4%	38	1	39
	Semantics	51.3%	20	19	39
Present	Syntax	80.0%	28	7	35
	Semantics	45.7%	16	19	35
Future	Syntax	98.6%	72	1	73
	Semantics	53.4%	39	34	73

Table 4.3.7: Number of correct and incorrect words generated using the third increment isiXhosa grammar (subjunctive mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	100.0%	26	0	26
	Semantics	53.8%	14	12	26
Present	Syntax	100.0%	30	0	30
	Semantics	50.0%	15	15	30
Future	Syntax	98.6%	72	1	73
	Semantics	53.4%	39	34	73

Table 4.3.8: Table showing whether various isiXhosa rules contain various features. True means that the element can be found in the rules.

Indicative & Participial	Simple aspect	Progressive aspect	Exclusive aspect	Causative	Neuter	Intensive	Perfect	Continuity
Past	True	True	True	False	False	False	True	True
Present	True	True	True	True	True	True	True	True
Future	True	True	True	True	True	True	False	True
Subjunctive	Simple aspect	Progressive aspect	Exclusive aspect	Causative	Neuter	Intensive	Perfect	Continuity
Past	True	True	True	False	False	False	True	True
Present	True	True	True	True	True	True	False	True
Future	True	True	True	True	True	True	True	True

Indicative and Participial

PAST
(X30.) $Verb \rightarrow NPC_2 \quad A_{pes} \quad OC \quad VR \quad S_p$
(X31.) $Verb \rightarrow NPC_o \quad A_{pes} \quad OC \quad VR \quad S_{np}$

PRESENT
(X32.) $Verb \rightarrow A_{pes} \quad OC \quad VR \quad S_p$
(X33.) $Verb \rightarrow SC \quad PC \quad OC \quad VR \quad S_{np}$
(X34.) $Verb \rightarrow A_{pe} \quad OC \quad VR \quad S_{np} \quad a$

FUTURE
(X35.) $Prefix \rightarrow SC \quad IM \quad F \quad A_{sa}$
(X36.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{suffix}$

Subjunctive

PAST
(X37.) $Verb \rightarrow NPC_2 \quad A_{ps} \quad OC \quad VR \quad S_p$
(X38.) $Verb \rightarrow NPC_o \quad A_{es} \quad PR \quad OC \quad VR \quad S_{np} \quad a$

PRESENT
(X39.) $Prefix \rightarrow A_{es} \quad PC_1$
(X40.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np}$
(X41.) $Verb \rightarrow A_{pes} \quad OC \quad VR \quad S_{np} \quad a$

FUTURE
(X42.) $Prefix \rightarrow SC \quad IM \quad F \quad A_{sa}$
(X43.) $Verb \rightarrow Prefix \quad OC \quad VR \quad Suffix$

Figure 4.5: Context free grammar rules that generate isiXhosa verbs for various moods (Inductive, Participial, and Subjunctive) and tenses (Past, Present and Future).

4.4 IsiZulu verb rules

This section begins with discussing some isiXhosa verb features and this is followed by a description of the isiZulu context free rule development using the approach described in Section 4.1.

4.4.1 IsiZulu features

IsiZulu, like isiXhosa, has two main verbal concords, the subject concord and the object concord.
(Z1.) $SC \rightarrow ngi|si|u|ni|ba|i|li|a|zi|lu|bu|ku|\varepsilon$
(Z2.) $OC \rightarrow ngi|si|ku|ni|m|ba|wu|yi|li|wa|zi|lu|bu|\varepsilon$
These concords are also noun class dependent and the pairing between noun class and concord is given in Table 4.4.2. The various isiZulu noun classes are provided in Table 2.2.1. The subject concord also undergoes modification when it is prefixed by the negative prefix ($NP_{pref} \rightarrow a$) hence resulting in the additional rule ($SC_n \rightarrow angi|asi|awu|ani|aba|ayi|ali|a|azi|alu|abu|aku$). This rule is unnecessary in our case, however, as we are not dealing with negated verbs. The isiZulu verb, similar to isiXhosa, is also considered as being made up of five elements (or slots) in our work (as was discussed in Section 4.1). Likewise, what is referred to as the prefix by this work in the verb is made up of morphemes that reflect the aspect, tense and subject concord.

Canonici [18], when defining aspect for isiZulu, mentions that it is a grammatical category that applies to the verb and it is marked by a morpheme that can be a suffix or prefix. More specifically, the author points out that the aspect is not used to show specifically when an action takes place but its duration and the type of action [18,

Table 4.4.1: IsiXhosa subject and object concords (Adapted from [13, p.17] [124]).

Type	Subject	Object
1st person singular	ndi	
2nd person singular	u	ku
3rd person singular	u	m
1st person plural	si	
2nd person plural	ni	
3rd person plural	ba	

Class		
1	u,a	e,m
1a	u,a	e,m
2	ba	
2a	ba	
3	u	wu
4	i	yi
5	li	
6	a	wa
7	si	
8	zi	
9	i	yi
10	zi	
11	lu	
14	bu	
15	ku	

Table 4.4.2: IsiZulu subject and object concords (Adapted from [13, p.17] [52]).

Type	Subject	Object
1st person singular	ngi	
2nd person singular	u	ku
3rd person singular	u	m
1st person plural	si	
2nd person plural	ni	
3rd person plural	ba	

Class		
1	u	m
1a	u	m
2	ba	
2a	ba	
3	u	wu
3a	u	wu
4	i	yi
5	li	
6	a	wa
7	si	
8	zi	
9	i	yi
9a	i	yi
10	zi	
11	lu	
14	bu	
15	ku	
17	ku	

p. 69]. This observation is consistent with Blyth's [10] view on aspect. This work's definition of aspect is slightly different to the work done by Doke [26] [27] in which the author draws distinctions into what he terms implication and manner/aspect. There are 6 types of aspect in isiZulu, according to Canonici [18]. We are not interested in all the aspects but in the simple, progressive and exclusive aspects as focused on in isiXhosa because only those are most relevant for the weather. Section 3.4 also mentioned that the verbal moods we are interested in are the subjunctive, indicative, and participial moods. Furthermore, we will also examine the present, past, and future tenses with the exclusion of remoteness for both the past and the future. This exclusion of remoteness applies to two moods but not the subjunctive since its tenses are slightly different.

A few of the auxiliaries provided by Doke [26, p. 271] are needed when discussing tense and continuity, and they can be found in Figure 4.3. Continuity in the tenses can be shown using two morphemes. The first morpheme is applicable only for the present tense (Rule XZ27). The second morpheme is introduced by Doke [26, p. 271] as being applicable for the immediate past. In this work, however, we have referred to it as the non-present tense continuity morpheme (Rule XZ32) because in our analysis of the work done by Khumalo [55, p.91-93] we see that the morpheme is also used with the future tenses. Specifically, Khumalo says continuity in the future tenses can be identified by the morpheme *-zabe-* [55, p. 91]. However, closer scrutiny reveals that this morpheme is a combination of future tense verbal auxiliaries together with the morpheme. Its long form is given by $RF/IM + F + NPC = za/ya + uku + be$. It is worth noting that the future auxiliary (F) is dropped since the morpheme is presented in the contracted form by Khumalo [55, p.91]. Furthermore, Khumalo [55, p. 91] refers to the *be* morpheme simply as the continuous morpheme, with no reference to its specificity to the immediate past, unlike Doke [26] [27].

(Z3.) $PCP_{\circ} \rightarrow PC\|P$	(Present tense continuity/Progressive aspect)
(Z4.) $F \rightarrow uku$	(Morpheme used to indicate the future)

Figure 4.6: Context free grammar rules for generating morphemes used to indicate aspect and tense in the isiZulu verb prefix.

4.4.2 Increment 0 : Verbal prefix composition

The preliminary set of rules is formed with information gathered from Doke [27] [26] and Grout [41]. The isiZulu base rules that are needed to construct the prefix are listed in Figures 4.6 and 4.3. Similar to isiXhosa, prefixes are divided based on mood and tense. The superscript ($[0..1]$) shows that a particular item can either be present or absent.

Indicative mood

- $P_{prefix} \rightarrow NPC^{[0..1]} \quad E^{[0..1]} \quad SC \quad P^{[0..1]}$ (Immediate past)

- $P_{prefix} \rightarrow E^{[0..1]} \quad SC \quad PCP_{o}^{[0..1]}$ (Present)
- $P_{prefix} \rightarrow SC \quad IM \quad F \quad NPC \quad E^{[0..1]} \quad S_{space} \quad SC \quad P^{[0..1]}$ (Immediate future)

Participial mood

- $P_{prefix} \rightarrow NPC^{[0..1]} \quad E^{[0..1]} \quad SC \quad P^{[0..1]}$ (Immediate past)
- $P_{prefix} \rightarrow E^{[0..1]} \quad SC \quad PCP_{o}^{[0..1]}$ (Present)
- $P_{prefix} \rightarrow SC \quad IM \quad F \quad NPC \quad E^{[0..1]} \quad S_{space} \quad SC \quad PCP_{o}^{[0..1]}$ (Immediate future)

The method of formalisation of the rules that was used with isiXhosa was also followed, that is, the preliminary rules were used to manually generate strings. Each of the generated strings was marked with *True*, if syntactically correct, and with *False* if the prefix does not exist in isiZulu. The decision to generate the strings manually, despite the preliminary rules being made up of rule sets for two moods only, was also laborious and tedious. The rules were updated to not produce the syntactically incorrect prefixes. The updated rules for generating the prefix are listed below (Z5 to Z10). Here, we make use of the rules for the 'compounded' exclusive and progressive aspects that are identified by A_e, A_p and their combined rule as A_{pes} where 'PES' stands for progressive, exclusive and simple. All these rules are defined in Figure 4.3.

Indicative mood

(Z5.) $P_{prefix} \rightarrow NPC \quad A_{pes} \quad | \quad A_{pes}$ (Immediate past)

(Z6.) $P_{prefix} \rightarrow A_{pes} \quad | \quad A_{es} \quad PC \quad | \quad SC \quad C$ (Present)

(Z7.) $P_{prefix} \rightarrow SC \quad IM \quad F \quad | \quad SC \quad IM \quad F \quad NPC \quad A_{pes}$ (Immediate future)

Participial mood

(Z8.) $P_{prefix} \rightarrow NPC \quad A_{pes} \quad | \quad A_{pes}$ (Immediate past)

(Z9.) $P_{prefix} \rightarrow A_{pes} \quad | \quad A_{es} \quad PC \quad | \quad SC \quad C$ (Present)

(Z10.) $P_{prefix} \rightarrow SC \quad IM \quad F \quad | \quad SC \quad IM \quad F \quad NPC \quad A_{e} \quad PC$
$\quad | \quad SC \quad IM \quad F \quad NPC \quad A_{pes}$ (Immediate future)

4.4.2.1 Subjunctive mood

There is no need for the preliminary rules that make use of the superscript ($[0..1]$) on the subjunctive mood because all the information was taken from Doke [27] [26]. This mood is used to describe unreal and hypothetical events [55, p. 102]. Here we see that there is a tense called present-future tense that is referred to as simply the present tense. This is because it "usually [has] a future intent" [27, p. 186]

(Z11.) $P_{prefix} \rightarrow SC \quad PR$ (Past)

(Z12.) $P_{prefix} \rightarrow SC \quad PR \quad SC$ (Past, continuous)

(Z13.) $P_{prefix} \rightarrow SC$ (Present)

(Z14.) $P_{prefix} \rightarrow SI \quad SC$ (Present, imperatively)

(Z15.) $P_{prefix} \rightarrow SC \quad PR \quad IM \quad F$ (Emphatic future tense)

Table 4.4.3: Number of correct and incorrect words generated using the first iteration isiZulu grammar (indicative mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	50.0%	9	9	18
	Semantics	50.0%	9	9	18
Present	Syntax	100.0%	30	0	30
	Semantics	50.0%	15	15	30
Future	Syntax	0%	0	70	70
	Semantics	0%	0	70	70

Table 4.4.4: Number of correct and incorrect strings generated using the first iteration isiZulu grammar (participial mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	50.0%	9	9	18
	Semantics	50.0%	9	9	18
Present	Syntax	100.0%	30	0	30
	Semantics	50.0%	15	15	30
Future	Syntax	0%	0	91	91
	Semantics	0%	0	91	91

4.4.3 Increment 1 : Verbal prefix, verb root, and pre-final vowel suffix

The rules (Z5 to Z15) were updated by adding the OC, VR, and S_{suffix}. A set of strings was generated with the rules using the CFG module in the NLTK. The same values for the verb root, subject concord, and object concord that were used in isiXhosa were also used for isiZulu. This is possible because the verb stem *-zol-* and noun *izulu* also exists in isiZulu. The resulting words were classified into three categories: correct strings (syntax and semantics), syntactically correct strings that are not used in the language, and non-existent/invalid words. The rules that are for generating terminals in isiZulu but also exist in isiXhosa are listed in Figure 4.4, and the terminals that exist only in isiZulu are listed as rule Z16 and Z17. The following sections (4.4.3.1 and 4.4.3.2) describe the results of the correctness evaluations and the overall issues facing each mood-specific rule set.

(Z16.) $F \rightarrow uku$

(Z17.) $SI \rightarrow ma$

4.4.3.1 Indicative and Participial

The correctness results obtained after the inclusion of the OC, VR, and S_{suffix} are listed in Table 4.4.3 (Indicative) and Table 4.4.4 (Participial). The participial mood and indicative have very similar rule sets (as can be seen in Section 4.4.2), with the exception of an additional rule that exists in the participial mood (Z10). This rule was found to generate incorrect words as it makes use of the mutually exclusive *NPC* and *PC* morphemes within the same rule. Furthermore, there was a set of words that were incorrect due to the juxtaposition of the F and A_e

morphemes when the morpheme *NPC* generated an empty string (in Rule Z7 and Z10). Eliminating the errors resulted in equivalent rule sets for the indicative and participial moods. The updated rules are follows:

Past

(Z18.) $P_{prefix} \rightarrow NPC_0 \quad A_{pe}$

(Z19.) $Verb \rightarrow P_{prefix} \quad OC \quad VR \quad S_p$

(Z20.) $Verb \rightarrow NPC_0 \quad SC \quad OC \quad VR$

Present

(Z21.) $P_{prefix} \rightarrow A_{pes} \mid A_{es} \quad PC \mid SC \quad C$

(Z22.) $Verb \rightarrow P_{prefix} \quad OC \quad VR \quad S_{nic}$

Future

(Z23.) $P_{prefix} \rightarrow SC \quad IM \quad F \quad A_{sa}$

(Z24.) $Verb \rightarrow P_{prefix} \quad OC \quad VR \quad S_{suffix}$

The updated rules (Z18 to Z24) yielded the correctness results listed in table 4.4.5 for both moods. There was some improvement for the syntactical correctness of the rules. The past tense rules saw a 30% correctness increase, albeit the verb count decreased by 8. The indicative mood's present tense rules saw an increase of 50% in correctness and a growth of 12 in verb count. The future tense saw a significant increase in correctness by 98%.

Table 4.4.5: Number of correct and incorrect strings generated using the second iteration isiZulu grammar (indicative & participial mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	80.0%	8	2	10
	Semantics	80.0%	8	2	10
Present	Syntax	100.0%	30	0	30
	Semantics	50.0%	15	15	30
Future	Syntax	98.0%	48	1	49
	Semantics	51.0%	25	24	49

Table 4.4.6: Number of correct and incorrect strings generated using the first iteration isiZulu grammar (subjunctive mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	0.0%	0	4	4
	Semantics	0.0%	0	4	4
Present	Syntax	100.0%	6	0	6
	Semantics	50.0%	3	3	6
Future	Syntax	0%	0	7	7
	Semantics	0%	0	7	7

4.4.3.2 Subjunctive

The correctness results obtained after the inclusion of the OC, VR, and S_{suffix} for the subjunctive mood's can be seen from table 4.4.6. The use of the past remote (PR) morpheme resulted in significant errors. Moreover, the rules that include aspect were missing. These issues were addressed to obtain the following rules (Z25 to Z31)

Past

(Z25.) $Verb \rightarrow NPC_o \quad A_p \quad OC \quad VR \quad S_p$

(Z26.) $Verb \rightarrow NPC_o \quad A_{es} \quad PR \quad OC \quad VR$

Present

(Z27.) $P_{prefix} \rightarrow SC$

(Z28.) $P_{prefix} \rightarrow SI \quad SC$

(Z29.) $Verb \rightarrow P_{prefix} \quad OC \quad VR \quad S_{nic}$

Future

(Z30.) $P_{prefix} \rightarrow SC \quad IM \quad F \quad A_{sa}$

(Z31.) $Verb \rightarrow P_{prefix} \quad OC \quad VR \quad Suffix$

These updated rules yielded the results listed in table 4.4.7. The syntactical correctness of the past tense rules rose by 87.5% and the verb count doubled. The correctness of the future tense rules rose by 98%.

Table 4.4.7: Number of correct and incorrect strings generated using the second iteration isiZulu grammar (subjunctive mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	87.5%	7	1	8
	Semantics	87.5%	7	1	8
Present	Syntax	100.0%	6	0	6
	Semantics	50.0%	3	3	6
Future	Syntax	98.0%	48	1	49
	Semantics	53.1%	26	23	49

4.4.4 Increment 2 : Complete verbal rules

In a similar fashion to isiXhosa, the isiZulu rules obtained for the various tenses and moods were checked to verify the presence of all necessary notions such continuity, aspects, etc. Table 4.4.8 shows the missing and present elements. The rules were updated to include the missing features and automatically include the final vowel using the method detailed in Section 4.2.2. The complete rules are listed in Figure 4.7. These complete rules were used to generate verbs, and then those verbs were then classified. The classification results are given in Table 4.4.9 (Indicative and participial) and Table 4.4.10 (Subjunctive). The two tables show that the generated verbs have high syntactical correctness values ranging between 88% and 100% (inclusive) for the various tenses.

[2] Continuity here is not shown using the various continuity morphemes used in other rules.

Table 4.4.8: Table showing whether various isiZulu rules contain various features. True means that the element is can be found in the rules.

Indicative & Participial	Simple aspect	Progressive aspect	Exclusive aspect	Causative	Neuter	Intensive	Perfect	Continuity
Past	True	True	True	False	False	False	True	True
Present	True	True	True	True	True	True	False	True
Future	True	True	True	True	True	True	True	True
Subjunctive	Simple aspect	Progressive aspect	Exclusive aspect	Causative	Neuter	Intensive	Perfect	Continuity
Past	True	True	True	False	False	False	True	True
Present	True	False	False	True	True	True	False	True[2]
Future	True	True	True	True	True	True	True	True

Table 4.4.9: Number of correct and incorrect words generated using the third increment isiZulu grammar (indicative and participial mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	97.2%	35	1	36
	Semantics	47.2%	17	19	36
Present	Syntax	88.9%	16	2	18
	Semantics	55.6%	10	8	18
Future	Syntax	98.6%	72	1	73
	Semantics	53.4%	39	34	73

Table 4.4.10: Number of correct and incorrect words generated using the third increment isiZulu grammar (subjunctive mood). Correctness is divided into semantic and syntactic categories.

		Percentage correct	Correct	Incorrect	Total
Past	Syntax	100.0%	28	0	28
	Semantics	64.3%	18	10	28
Present	Syntax	100.0%	14	0	14
	Semantics	57.1%	8	6	14
Future	Syntax	98.6%	72	1	73
	Semantics	53.4%	39	34	73

Indicative and Participial

PAST

(Z32.) $Verb \rightarrow NPC_0 \quad A_{pes} \quad OC \quad VR \quad S_{np}$

(Z33.) $Verb \rightarrow NPC_2 \quad A_{pes} \quad OC \quad VR \quad S_p$

PRESENT

(Z34.) $Verb \rightarrow A_{pes} \quad OC \quad VR \quad S_p$

(Z35.) $Verb \rightarrow A_{es} \quad PC_1 \quad OC \quad VR \quad S_{np}$

FUTURE

(Z36.) $Prefix \rightarrow SC \quad IM \quad F \quad A_{sa}$

(Z37.) $Verb \rightarrow Prefix \quad OC \quad VR \quad Suffix$

Subjunctive

PAST

(Z38.) $Verb \rightarrow NPC_2 \quad A_{ps} \quad OC \quad VR \quad S_p$

(Z39.) $Verb \rightarrow NPC_0 \quad A_{es} \quad PR \quad OC \quad VR \quad S_{np} \quad a$

PRESENT

(Z40.) $Prefix \rightarrow SI \quad SC \quad | \quad SC$

(Z41.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_p$

(Z42.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np} \quad a$

FUTURE

(Z43.) $Prefix \rightarrow SC \quad IM \quad F \quad A_{sa}$

(Z44.) $Verb \rightarrow Prefix \quad OC \quad VR \quad Suffix$

Figure 4.7: Context free grammar rules that generate isiZulu verbs for various moods (Inductive, Participial,and Subjunctive) and tenses (Past, Present and Future).

5

Expert evaluation of grammars

This chapter will describe the expert evaluation of the created context free grammar rules, the calculation of verb similarity between the two languages, and quantification of consecutive vowel incidents in strings from the two languages in order to understand the degree of improvement phonological conditioning could introduce. Section 5.1 will discuss what the expert evaluation entails, verb similarity is calculated by using binary similarity measures as is discussed in Section 5.2, and Section 5.3 details the development of equations and their use in calculating the number of consecutive vowels in the two verb fragments.

5.1 Linguist evaluation

The linguist's evaluation differs from the quality evaluation of the rules in each increment during the development as it was done by the researcher. In Section 5.1.1 we describe how the expert evaluation was performed, and report its results in Section 5.1.2.

5.1.1 Materials and methods

A number of rules that generate verbs were developed through an incremental process and evaluated by the researcher in Section 4.3 and Section 4.4. This section discusses the evaluation of the resulting verbs generated by these rules using language experts. The expert evaluation of rules is done by generating strings using Python and NLTK. In particular, twenty-five verbs that exist in isiZulu and isiXhosa are extracted from an English-isiZulu dic-

Table 5.1.1: Twenty five verbs that exists in isiXhosa and isiZulu extracted from Doke et al. [28]. The verbs are used in the measurement of the similarity of the developed rules.

Verb	Root	Meaning
aba	ab	apportion, distribute, divide out
abuza	abuz	cast off skin
ahluka	hluk	deviate from, part from
ahlula	hlul	overcome, conquer, overpower
akha	akh	build (a house), erect, construct, establish
baca	bac	hide oneself, take shelter, lie low
bacalala	bacalal	lie flat on the stomach
bala	bal	(1) scratch marks, cut incisions (2) write (letter, etc), draw (3) enter, write down (a person's name, etc)
baqa	baq	(1) light up, illuminate (2) kindle a light, cause to shine (3) come upon unexpectedly, light upon (Example *Niyokubaqwa ngubani?*)
beja	bej	(1) make a vow, take an oath, swear (2) bet
cangcatha	cangcath	(1) rap, tap quickly on top (2) knock about (as in a stick fight) (3) type, strike typewriter (4) tramp down, beat down (as hail does vegetation); wear a foot-path
cacisa	cacis	(1) cause to smash (3) Explain; make clear
cazulula	cazulul	(1) split right upon; untie, disentangle (2) unravel an affair
ceba	ceb	(1) inform against, report about, tell on, make known secret concerning another (2) devise, contrive, invent, conspire
cela	cel	(1) ask for, beg, request (2) settle for a wife
dala	dal	(1) create, bring into being, form (2) conceive (3) cause
dangala	dangal	get depressed, dejected, strength-less, languid, weary
deda	ded	get out of the way
delela	delel	(1) be satisfied for; abandon for (2) have contempt for; despise, disregard contemptuously; act without concern or constraint
dibana	diban	get mixed together
efuza	fuz	(1) strip off thatch (2) resemble
ehlela	hlel	(1) come down upon;happen to, befall; alight on (2) slope down
ekhama	kham	squeeze, throttle
embesa	mbes	clothe or cover another with
enaba	nab	(1) spread out (2) live at ease, be comfortable, be at home

tionary [28] in alphabetical order from the a-commenncing to e-commenncing words sections (five verbs from each of the five sections). All the extracted word, alongside their determined roots, are listed in Table 5.1.1. Furthermore, five pairs of subject and object concords are randomly selected. The selected pairs are (*u,yi*), (*i,wa*), (*a,zi*), (*lu,bu*), and (*i,yi*) where the first tuple item is the subject concord and the second is the object concord. The 25 roots together with the *-zol-* root are paired with the five concords pairs. The last root (*-zol-*) is re-used from previous exercises. The pairings of concords and root are inserted into the rules to generate strings in isiZulu and isiXhosa. This resulted in 49400 strings for both languages. Ninety nine[1] strings are randomly[2] taken from each language set, packaged into a spreadsheet, and sent to two linguists for evaluation. The linguists are required to annotate each word with True/False for syntactic correctness, True/False for semantic correctness, and add a comment should they have one. The strings are not subjected to phonological conditioning, and therefore there are words that have consecutive vowels. The linguists were asked to ignore this imperfection. The resulting annotated spreadsheets are converted into comma separated files, and the ratio of correct and incorrect strings is analysed using a Python script. Furthermore, Fisher's exact test is used to evaluate whether there is a statistically significant difference between syntactic and semantic correctness between the evaluations of the two languages.

5.1.2 Results

The semantic and syntactic correctness of the isiXhosa strings is shown in Table 5.1.2. Two strings were marked as being not isiXhosa, and were identified as having a high likelihood of being isiZulu, these were *seiyazacebeka*

[1] 100 verbs were sampled and a single verb was mistakenly discarded from the isiXhosa and isiZulu lists.

[2] python random module

and *seiyazahlulisa*. The roots used in the words exist in isiXhosa (*ceba* [69, p.388] and *hlula* [69, p.445] can be found in an isiXhosa dictionary). Also, there were 2 words that were marked as not existing in isiXhosa but not marked as being isiZulu : *izabejekisisa* and *seiyazahlukeka*. The problem with the former is the combination of the verb extensions and with the latter, when we ignore the consecutive vowels *ei*, is the missing *-ku-* after the *-za-*. In other words, the two morphemes used to mark the future tense in verbs have to be used together for this particular word. Closer scrutiny of the evaluated verbs shows that there are only a few cases where *-za-* can be used without the morpheme *-ku-*. There was an inconsistency with the isiXhosa linguist's evaluation. The word *luyabubacisisa* was included twice in the isiXhosa verb set, but was annotated once as being semantically true, and once as semantically false.

Table 5.1.2: Summary of the linguists' semantic and syntactic correctness evaluation of the isiXhosa and isiZulu generated strings.

		Percentage correct	Correct	Incorrect	Total
IsiXhosa	Syntax	52%	51	48	99
	Semantics	58%	57	42	
isiZulu	Syntax	23%	16	53	69
	Semantics	25%	17	52	

There were 99 isiZulu verbs, and 30 of them were not fully annotated with True/False for the syntatical and semantic correctness fields. The 30 isiZulu strings that were not annotated based on the provided instructions are listed in Table 5.1.3. Out of the 69 verbs, the semantic and syntactic correctness of the isiZulu strings is shown in Table 5.1.2. The words *isazadibanisisa* and *beizadibanekisa* were annotated as being isiXhosa and not isiZulu. This is despite the fact that the root exists in isiZulu, for instance, *dibanisa* [28, p.144] can be found in an isiZulu dictionary. Furthermore, the verb phrase *luzaukube lubudede* was determined by the linguist to not be isiZulu. The root of the verb, however, exists in the language in the form *deda* [28, p.141]. The perfect form of the verb (*dede*), combined with the supplementary part (*luzaukube lubu-*) results in a word that does not exist. The word *beseiyimbesa* was marked as being syntactically incorrect, it had no semantic correctness annotation, and the comment 'besiyimbesa would be correct' was added. This means that the linguist did not find a problem with the verb aside from the consecutive vowels. The string *lubudangalise* is syntactically correct according to the context free grammar, however, the combination of concords and verb extensions results in a non-existent isiZulu verb. It is for this reason that the linguist marked this verb as being syntactically and semantically incorrect. The linguist's incorrect annotation is consistent with the isiXhosa linguist's observation that it is sometimes confusing to separate semantic and syntactic correctness in natural languages.

The total number of verbs whose semantic correctness was properly annotated in isiXhosa and isiZulu are listed in Table 5.1.4. Likewise, the number of verbs whose syntactic correctness was properly annotated are listed

Table 5.1.3: Thirty isiZulu verbs that were not fully annotated during evaluation. They are split into 5 groups. Group 1 only have syntactical correctness annotated, and a comment was provided. Group 2 only have a comment about the future tense formative. Group 3 are semantic and syntactical annotated, however, the syntactical correctness of the verb phrase is per word. Group 4 only annotated the syntactical correctness of each word in a verb phrase, and there is no annotation of the semantic correctness. Group 5 only the syntactical correctness is annotated.

	Word	**Syntax**	**Semantic**
Group 1	seiazacangcatheka	TRUE BUT FUTURE TENSE IS -ZO-	-
	seiayicaciseka	SEYIYACACISEKA IS TRUE	-
Group 2	seiazahlela	SEYIZOHLELA (-ZO-) AS FUTURE TENSE	-
	isazaabuzisa	AA IS A . THE FUTURE TENSE IS -ZO-	-
Group 3	izaukuba isazabacalalisa	TRUE FIRST WORD/FALSE SECOND	FALSE
Group 4	uzaukuba seuyibacalalisisa	TRUE/FALSE	-
	luzaukube selubukhamisile	TRUE/FALSE	-
	luzaukuba lubudangale	TRUE/FALSE	-
	izaukuba isazadangalekisa	TRUE/FALSE	-
	luzaukuba selubuakhisis	TRUE/FALSE	-
	uzaukuba seuyiabisise	TRUE FIRST WORD/FALSE SECOND	-
	izaukuba seiyicelekisise	TRUE FIRST WORD/FALSE SECOND	-
	luzaukuba selubucazululile	TRUE FIRST WORD/FALSE SECOND	-
	izaukube isazakhameke	TRUE FIRST WORD/FALSE SECOND	-
	luzaukube selubuhlukise	TRUE FIRST WORD/FALSE SECOND	-
	izaukuba iyifuzekisise	TRUE FIRST WORD/FALSE SECOND	-
	uzaukuba seuyibacalalisa	TRUE FIRST WORD/FALSE SECOND	-
	izaukube isayicele	TRUE FIRST WORD/FALSE SECOND	-
	izaukuba izadibanisa	TRUE FIRST WORD/FALSE SECOND	-
	uzaukuba usayidangaleka	TRUE FIRST WORD/FALSE SECOND	-
	izaukuba isawaabuze	TRUE FIRST WORD/FALSE SECOND	-
	izaukuba iyidibanise	TRUE FIRST WORD/FALSE SECOND	-
	uzaukuba uyicangcatheka	TRUE FIRST WORD/FALSE SECOND	-
	izaukube seiyidangalisisa	TRUE FIRST WORD/FALSE SECOND	-
	izaukube isayicangcathekisise	TRUE FIRST WORD/FALSE SECOND	-
	izaukuba iwakhamiseke	TRUE FIRST SECOND FALSE	-
	izaukube seiwaakhekise	FIRST ONE TRUE. SECOND FALSE	-
	luzaukuba selubufuzeke	LUZAUKUBA IS TRUE. THE OTHER FALSE	-
Group 5	iawahlela	TRUE	-
	beseiyimbesa	FALSE	-

Table 5.1.4: Number of semantically correct and incorrect verbs from the total isiXhosa and isiZulu strings that were correctly annotated as per evaluation instructions.

	True	False	Total
isiXhosa	47	52	99
isiZulu	17	53	70
Total	64	105	169

Table 5.1.5: Number of syntactically correct and incorrect verbs from the total isiXhosa and isiZulu strings that were correctly annotated as per evaluation instructions.

	True	False	Total
isiXhosa	53	46	99
isiZulu	17	54	71
Total	70	100	170

Table 5.1.6: Number of syntactically correct and incorrect verbs from the isiXhosa and isiZulu strings that can be interpreted from the annotations provided by the linguists.

	True	False	Total
isiXhosa	53	46	99
isiZulu	17	78	95
Total	64	105	169

in Table 5.1.5. These values show that there is a significant statistical association between the syntactic (two-tailed p=0.0001, Fisher's exact test) and language. The same is true for semantic correctness and language (two-tailed p=0.0023, Fisher's exact test). In Table 5.1.3, the verb phrases in group 4 and 3 can be considered to be syntactically incorrect, and group 1 can be considered to be syntactically correct. If one is of that view then the number of isiZulu verbs that are syntactically correct (and incorrect) can be updated to obtain the values listed in Table 5.1.6. This updated table also shows that there is a strong statistically significant association between the syntactic correctness (two-tailed p<0.0001, Fisher's exact test) and language for the developed rules. This means that there is a significant difference in the number of correct verbs between the two languages and this should be taken into account when interpreting the results.

5.2 Verb similarity

This section begins with a description of the creation of verb sets that will be used to compare isiXhosa and isiZulu. The similarity of the verbs is analysed through (1) verb parse trees and (2) comparing the similarity of verb sets generated by the developed rules. The section will end with the obtained results.

5.2.1 Materials and methods

The number of the developed isiXhosa and isiZulu rules are shown in Table 5.2.1, and the language specific rules are given in Figure 4.5 & Figure 4.7. We will compare the open class rules to each other, and single out the rules that are different between the two languages. This comparison will be done manually by analysing the differences between the rules for each tense and mood. For rules that are not identical between the two languages for each tense and mood, we will make use of parse trees to examine the differences. The parse trees will be created by hand

on paper, and through the Dia Diagram Editor[3].

Table 5.2.1: 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.	Morpho- tact.
isiZulu	49	13	6	8	22
isiXhosa	52	12	9	8	23
Intersection	42	11	6	8	17

We will also compare the similarity of the rules by generating strings using Python and NLTK, and compare the resulting isiXhosa and isiZulu strings sets to each other. It is common knowledge, to anyone familiar with the two languages, that the two languages share some verbs. This opens up the ability to compare the verbal rules by comparing their generated verbs. This comparison is based on the premise that a similarity evaluation on the shared language space can give an indication of similarity between the two rule sets. The process in detail, we fix the root, subject and object concord for isiXhosa and isiZulu verb rules to verb root *-zol-*, subject concord *li-*, and an empty object concord. The verb rules were used to form four clusters of rules. These are (1) complete set of rules, (2) present tense rules only, (3) all verb rules, excluding present tense rules, and (4) past tense rules. We generate a set of strings for the two languages for each of the four rule clusters. The resulting string sets are served as input to the binary similarity measures that were discussed in Section 2.7 to quantify similarity. Moreover, the set of 25 isiZulu and isiXhosa shared verbs, listed in Table 5.1.1, is used to determine whether there is a variation in the binary similarity results when different verb roots are used. Each verb root from the 25 verbs is paired with the subject concord '*li-*' and an empty object concord, and these strings are also provided as input to the binary similarity measures. Lastly, the impact of the change of concords is also interrogated. We randomly picked five concords (*a,zi*), (*i,wa*), (*i,yi*), (*lu,bu*), and (*u,yi*). The verb root *-zol-* is then paired with each concord pair. The complete set of verb rules (rule cluster 1) is used to generate strings and the resulting verbs are provided as input to the binary similarity measures.

In order to understand the difference between the Driver-Kroeber and Jaccard measures when the number of verbs increases, we generated 1024 cases of triples (a,b,c) using the Numpy[4] discrete uniform distribution random integer generator with the constraint $a + b + c = 1024$. The three variables are defined in Section 2.7.2, a is the number of verbs shared by the generated siXhosa and isiZulu, b is the number of verbs that exist in isiZulu but not isiXhosa, c is the number of verbs that exist in isiXhosa but not isiZulu. We calculated the difference between the Driver-Kroeber and Jaccard metrics for each of the 1024 triples. We have determined that the maximum difference between the measures is 0.247 and the average difference is 0.142.

[3] http://dia-installer.de/

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

5.2.2 Results

In this section we present the results for the two types of similarities; parse tree similarity and languages-space similarity.

5.2.2.1 Parse tree similarity

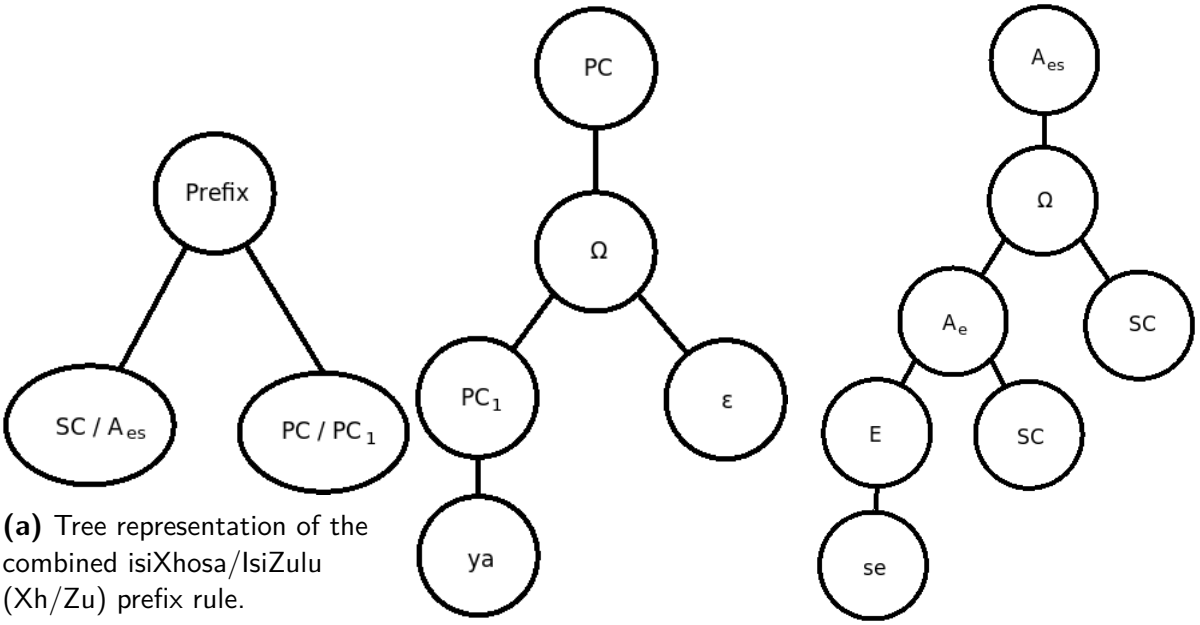
The rules that generate verbs for the past and future tenses for all moods are equivalent between the two languages. Differences exist in the present tense rules only. The rules that have differences are listed side-by-side in Figure 5.1. There are two special cases that introduce the differences: (1) The existence of an additional rule and (2) Slight differences in rules where most of the slots are the same. Specifically, we have observed that for the indicative and participial moods, isiXhosa has an additional rule ($Verb \rightarrow A_{pe} \quad OC \quad VR \quad S_{np}$ a).

IsiXhosa	**IsiZulu**
Indicative & Participial	Indicative & Participial
(xho.) $Verb \rightarrow SC \quad PC \quad OC \quad VR \quad S_{np}$	(zuo.) $Verb \rightarrow A_{es} \quad PC_1 \quad OC \quad VR \quad S_{np}$
Subjunctive	Subjunctive
(xh1.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np}$	(zu1.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_p$
(xh2.) $Verb \rightarrow A_{pes} \quad OC \quad VR \quad S_{np}$	(zu2.) $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np}$
(xh3.) $Prefix \rightarrow A_{es} \quad PC_1$	(zu3.) $Prefix \rightarrow SI \quad SC \mid SC$

Figure 5.1: Rules that have differences between isiXhosa and isiZulu's present tenses.

The rules that differ for isiXhosa and isiZulu's indicative and participial mood are labelled xho and zuo in Figure 5.1. The difference in the rules affects the prefix only. IsiXhosa uses a fixed present continuity indicator (PC), whereas in isiZulu it can be empty (PC_1). Furthermore, the isiZulu prefix, unlike isiXhosa, incorporates the exclusive aspect. These differences are not significant as the underlying structure of the two rules is the same, as can be seen in Figure 5.2. This is because the isiZulu rule is a 'super-rule' of its isiXhosa equivalent, that is, the set of words generated by the isiXhosa rule is entirely contained in the set generated by the isiZulu rule.

The subjunctive mood has 3 rules per language that have differences. Rules xh1 and zu1 in Figure 5.1 have minor differences in the suffix. IsiZulu requires the perfect suffix whereas isiXhosa deals with the neuter, intensive, and causative suffixes. This difference does not affect the overall structure of both rules as only one of their constituents are mutually exclusive and the rest being equivalent. The second rules (xh2 and zu2 in Figure 5.1) have differences in their prefixes. The prefix in isiXhosa deals with all three verbal aspects (simple, exclusive and pro-

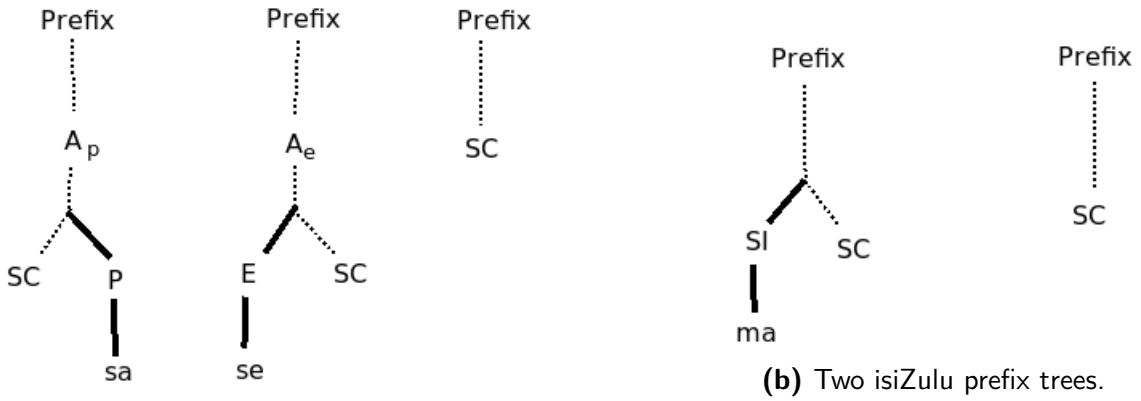


(a) Tree representation of the combined isiXhosa/IsiZulu (Xh/Zu) prefix rule.

(b) Subtree representation of an isiXhosa-only optional present continuity feature.

(c) Subtree representation of an isiZulu-only presence of the exclusive aspect.

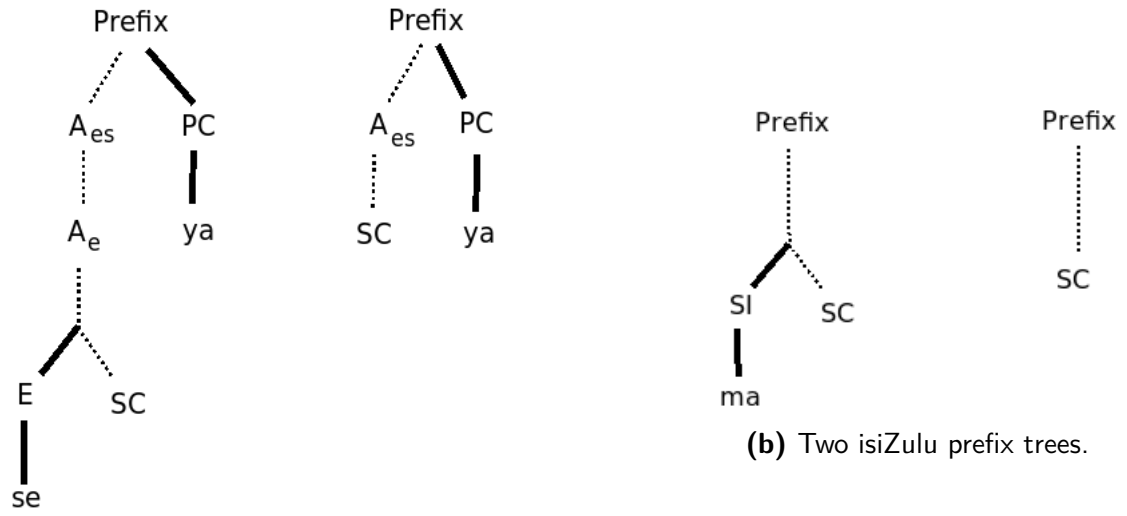
Figure 5.2: Tree representations of the isiXhosa and isiZulu's indicative and participial moods prefix (rule 0 in Figure 5.1). The Ω node represents mutual exclusiveness for its subtrees.



(a) Three isiXhosa prefix trees.

(b) Two isiZulu prefix trees.

Figure 5.3: Differences between isiXhosa (rule 2) and isiZulu's (rule 3) subjunctive moods prefix as listed in figure 5.1. The thin dotted lines show that only the subject concord is the only similar thing between the two languages.



(a) Two isiXhosa prefix trees.

(b) Two isiZulu prefix trees.

Figure 5.4: Differences between isiXhosa and isiZulu's subjunctive mood's prefix within rule 3 in figure 5.1. The thin dotted lines show that only the subject concord is the only similar thing between the two languages.

gressive) whereas isiZulu deals with only the mandatory verbal aspect (simple aspect). IsiXhosa can have three possible abstract values for its prefix, isiZulu can have two. They share a single value, and this is shown diagrammatically in Figure 5.3. The rules defining the general prefix (labelled xh3 and zu3 in Figure 5.1) for the present subjunctive also have differences. IsiXhosa incorporates continuity in the present tense, and the exclusive aspect. IsiZulu, on the other hand, does not include both features and this is shown in Figure 5.4.

5.2.2.2 Language space similarity

The binary measures that were discussed in Section 2.7 being the Jaccard, Sorenson, Driver-Kroeber, and Sorgenfrei, were used on the four sets of grammar rules (detailed in Section 5.2.1) in order to obtain the results listed in Table 5.2.2. All four measures are normalized and their range is between zero and one. A change of concords and root has no impact on the similarity values. There are two metrics that are the most intuitive with respect to similarity of languages, and these are the Jaccard and Driver-Kroeber metrics. The Jaccard is able to deduce the ratio of the shared items to the complete set of verbs. The Driver-Kroeber, on the other hand, gives us an average measure of the similarity of the each set to the other. In particular, it is sensitive to the fact that the two 'languages' have different sizes and uses this to generate a better measure. We also know from discussions in Section 2.7 that the Sorgenfrei index is close to the Jaccard (Sorg $\approx$ J) and the Sorenson index is close to the Driver-Kroeber (S $\approx$ DK). 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. The difference between the Jaccard and Driver-Kroeber for the complete set of rules is close to the average difference between the two metrics.

Table 5.2.2: Calculated 4 binary measure values for each verb pair set generated using the four rule sets (Complete set of rules, Present tense rules only, Past only, and Past and future tense rules only). The values are rounded off at 3 decimal place.

	Sorgenfrei (Sorg)	Jaccard (J)	Driver-Kroeber (DK)	Sorensen (S)
Complete set of rules	0.354	0.423	0.595	0.595
Present tense rules only	0.376	0.435	0.613	0.606
Past and future tense rules only	0.341	0.412	0.584	0.584
Past only	0.990	0.990	0.995	0.995

5.3 Phonological conditioning

In order to understand how to prioritise phonological conditioning in weather verbs, in this section we investigate the ratio of verbs that require it out of all the string generated by the CFGs. Section 5.3.1 details the development and intuition behind the equations for quantifying phonological conditioning. Section 5.3.2 will present the obtained results.

5.3.1 Materials and methods

To recall what was discussed in Section 2.2, phonological conditioning is the process by which consecutive vowels are eliminated from words in Bantu languages. This section discusses the creation of equations that will quantify the phenomenon for weather forecast strings generated by the developed CFG rules. These equations calculate the ratio of the number of consecutive vowel incidents to the total number of strings in a given CFG. These functions are deduced by first creating an abstract language using the developed CFG for both languages. This function deduction is done by learning the number of consecutive vowel incidents that are independent of actual concords and verb roots from the abstract language. This abstract language is necessary in order to be able to capture all types of words in both languages. That is, we cater for words that have all four type of roots and also have all possible types of object and subject concords.

In order to obtain the equations, we begin by defining four variables α,β,γ, and δ in Table 5.3.1 that are instrumental in the creation of equations to quantify the occurrence of consecutive vowels in strings. The variables quantify the number of verb roots, in a generated weather forecast verb 'language', that begin or end with a vowel.

IsiXhosa and isiZulu have at most 18 noun classes, and the language specific classes are listed in Table 2.2.1. Each noun class has its own subject and object concords. These concords can either be a singular vowel, or end in a vowel, etc. In particular, there are finite combinations of the two concords for each language. The concords can be paired together in at most 8 different ways for isiXhosa (Table 5.3.2), and at most 6 ways for isiZulu (Table 5.3.3). We have chosen a few concords that satisfy the conditions of being in each of the concord pair classes, and they are shown in Table 5.3.4 and Table 5.3.5 under the column labelled 'Example pair'.

Table 5.3.1: Four types of verb roots in isiZulu and isiXhosa. The type size refers to the cardinality of each verb root type, that is, the number of verb roots that are (or not) surrounded by a vowel on each size.

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

Table 5.3.2: The eight types of possible pairs of subject and object concords in isiXhosa.

Concord pair classification number	Subject concord	Object concord
1	Vowel only	Empty
2	Vowel only	Vowel only
3	Vowel only	Ends with vowel
4	Vowel only	No vowel
5	Ends with vowel	No vowel
6	Ends with vowel	Ends with vowel
7	Ends with vowel	Vowel only
8	Ends with vowel	Empty

The verb rules were used to generate semi-abstract verbs : the subject concord, object concord, and verb root were set to be the following variables $SC = X$, $OC = Y$, and $VR = Z$. This resulted in semi-abstract verbs of the form : *seXyaYZekisisa*, *seXyaYZisisa*, *beseXYZise*, etc. The other variables in the grammar are not restricted and they take on all their values. The X and Y variables were then replaced by each of the sampled concord pairs ('Example pair') to obtain semi-abstract verbs with only Z as the variable. This resulted in a total of 3928 semi-abstract isiXhosa verbs of the form : *sendiyamZisisa*, *seuaZisisa*, *sendimZisise*, and 491 semi-abstract strings for each of the 8 concord pair types. There were a total of 2634 semi-abstract isiZulu verbs of the form : *ngimZa*, *seuangiZisisa*, *beuangiZisa*, and 439 semi-abstract strings for each of the 6 concord pair types. The count of abstract verbs for each concord pair classification number are shown in Table 5.3.4 for isiXhosa and Table 5.3.5 for isiZulu. We then used Python to calculate the number of verbs that would require at least one application of phonological conditioning for the four possible classes of the variable Z (The variable can belong to the four types listed in Table 5.3.1). For instance, in isiXhosa concord pair classification number 5, there are 24 semi-abstract that will have consecutive vowels when the verb root belongs to Type D, and there are 491 semi-abstract verbs that will have consecutive vowels when the verb root belongs to Type B. Furthermore, there will be no semi-abstract strings that will have consecutive vowels when the verb root belongs to Type A and C. A similar counting process was conducted for each concord pair classification number in the two languages and this resulted in the functions for counting the number of affected verbs is shown in Table 5.3.4 for isiXhosa, and Table 5.3.5 for isiZulu in the column labelled "Count affected verbs". We can represent each of these functions with f_i where i is the concord

Table 5.3.3: The six types of possible pairs of subject and object concords in isiZulu.

Concord pair classification number	Subject concord	Object concord
1	Ends with vowel	Ends with vowel
2	Ends with vowel	No vowel
3	Ends with vowel	Empty
4	Vowel only	Ends with vowel
5	Vowel only	No vowel
6	Vowel only	Empty

Table 5.3.4: Information about the eight isiXhosa concord pairs. Example pairs are examples of concords that belong to the classification. Equations that calculate the number of verbs that have consecutive vowels for each concord pair class are listed in "Count affected verbs". The number of semi-abstract verbs (the verb root is variable Z) for each concord pair type are shown under "Number of semi-abstract verbs".

Concord pair classification number	Example pair	Count affected verbs	Number of semi-abstract verbs
1	u, ε	$491(\alpha+\beta+\gamma)+232\delta$	491
2	u, e	$491(\alpha+\beta+\gamma+\delta)$	491
3	u, ndi	$491(\alpha+\beta+\gamma)+232\delta$	491
4	u, m	$491\beta+232\delta$	491
5	ndi, m	$491\beta+24\delta$	491
6	ndi, ndi	$491(\alpha+\beta+\gamma)+24\delta$	491
7	ndi, e	$491(\alpha+\beta+\gamma+\delta)$	491
8	ndi, ε	$491(\alpha+\beta+\gamma)+24\delta$	491

pair classification number. Moreover, we can represent the count of semi-abstract verbs that belong to class i as c_i. Finally, we can calculate the percentage of weather forecast verbs that require conditioning out of the total number of verbs that can be generated by the CFG rules using the ratio of consecutive vowels (RCV) equations 5.1 and 5.2.

$$RCV_{xh}(\alpha,\beta,\gamma,\delta)=\frac{\sum_{n=1}^{8}f_i(\alpha,\beta,\gamma,\delta)}{\sum_{n=1}^{8}c_i\times(\alpha+\beta+\gamma+\delta)}\times 100 \tag{5.1}$$

$$RCV_{zu}(\alpha,\beta,\gamma,\delta)=\frac{\sum_{n=1}^{6}f_i(\alpha,\beta,\gamma,\delta)}{\sum_{n=1}^{6}c_i\times(\alpha+\beta+\gamma+\delta)}\times 100 \tag{5.2}$$

These two equations will be used to quantify the ratio of strings that have consecutive vowels for each of the various types of verb roots. A collection of verb roots sourced from the Resource Management Agency[5] will be used to determine the distribution of the various types of verb roots in the two languages that are listed in Table 5.3.1. We will then use that distribution as guide to approximate the ratio of weather forecast verbs that have consecutive vowels when the number of verb roots that are input to the CFG rules increases.

[5] https://rma.nwu.ac.za/

Table 5.3.5: Information about the eight isiZulu concord pairs. Example pairs are examples of concords that belong to the classification. Equations that calculate the number verbs that have consecutive vowels for each concord pair class are listed in "Count affected verbs". The number of semi-abstract verbs (the verb root is variable Z) for each concord pair type are shown under "Number of semi-abstract verbs".

Concord pair classification number	Example pair	Count affected verbs	Number of semi-abstract verbs
1	ngi, ngi	$439(\alpha+\beta+\gamma)+243\delta$	439
2	ngi, m	$439\beta+243\delta$	439
3	ngi, ε	$439(\alpha+\beta+\gamma)+243\delta$	439
4	u, ngi	$439(\alpha+\beta+\gamma)+368\delta$	439
5	u, m	$439\beta+368\delta$	439
6	u, ε	$439(\alpha+\beta+\gamma)+368\delta$	439

Table 5.3.6: Ratio of verbs that will require phonological conditioning out of the total number of generated verbs for each verb root type. Root types are define in Table 5.3.1. Ratios are calculated using formulae 5.1 and 5.2.

Language	A	B	C	D
IsiXhosa	75	100	75	45
IsiZulu	67	100	67	70

5.3.2 Results

We calculated the ratio of verbs that will require phonological conditioning for each type of verb root type, and we see that when one uses an isiXhosa verb root that belongs to Type A (it begins but does not end with a vowel) then 75% of the total number of generated verbs will require phonological conditioning. The ratios for each of the verb root types are listed in Table 5.3.6. IsiZulu and isiXhosa verb roots are not uniformly distributed among the four types. For instance, when we examine the isiXhosa and isiZulu verb roots distributed by the South African Department of Arts and Culture & Centre for Text Technology[6] through the Resource Management Agency we see that out of a total of 5839 isiZulu 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 isiXhosa 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 change in the ratio of generated weather forecast verbs that require phonological conditioning when the number of verbs that will appear in the forecast text increases. In particular, if we assume that for each verb root of Type A there will be two verb roots of Type D that will be used to generate verbs that will appear in a weather forecast that is made up of at most 500 verbs ($a+d \leq 500$ and $b=c=0$), then we see that 55% isiXhosa and 69% isiZulu generated weather forecast verbs will require phonological conditioning.

[6] http://www.nwu.ac.za/ctext

5.4 Discussion

This section discusses the results and shows how the research questions that were posed in Chapter 1 have been addressed. It then discusses the various ways to improve the design of the rules and grammar evaluation technique.

5.4.1 Weather forecast corpus

The limited nature of available corpora led us to develop a weather forecast corpus for the Southern African region. A number of sources were contacted in an attempt to source the corpus. We managed to obtain English forecasts from the South African Weather Services. This text was then translated into isiXhosa from English since grammatical features that exist in English are not guaranteed to exist as-is in isiXhosa. However, this approach has limitations. The manner in which weather is presented in a language might be different in another. It may also be different when presented to different audiences despite being written in the same language. This is particularly true for regional forecasts unlike point forecast that have report weather for a fixed location. For instance, some regional forecasts may prefer the use of compass directions whereas others may use landmarks to give an indication of direction. More generally, one can use altitude, absolute direction, coastal proximity, and population for providing spatial direction/position [127]. Our work extracts verbs from a corpus that has both point and regional forecasts. We do not analyse the distinction between the verbs from the two types of forecasts due lack of availability of resources. The impact of the differences in spatial references on the verbs is unclear. Other researchers are encouraged to investigate this by attempting to source naturally occurring Nguni weather forecasts. This can be done by manually recording and transcribing forecasts provided by the SABC over an extended time frame.

The English corpus that was collected in this work is made up of 12 human authored weather forecasts. These forecasts have a total of about 4334 words that collectively describe the weather. A small sample of that text was translated into isiXhosa. The complete translated sample is given Appendix A. It is a small corpus compared to the SumTime-Meteo corpus[7], for instance. The latter is made up of 1045 human authored texts. The difference in corpus size is not alarming because we have used the corpus to determine the nature of verb features. We have not used it in the traditional process of determining the exact lexical items that will be used in the generated text. Furthermore, our verbal features were representative of the various weather seasons because the forecasts were sampled from each month in a year.

5.4.2 CFG evaluation differences

In the incremental CFG rule development process in Chapter 4 there are increases in syntactic and semantic correctness between increments because the discovered inconsistencies for each increment are removed prior to proceeding to the following increment thus leading to better quality. Moreover, semantic correctness is generally lower syntactic correctness due overgeneration that is a result of the verbal extensions. The improvement of the syntactical correctness is a priority than semantic correctness because since the latter is attributable to overgeneration it can be improved outside by the grammar through the realisation engine's feature selection mechanism.

[7] https://github.com/AdeebNqo/Sumtime-Meteo-Cached

The rule development approach taken does not start with a complete and imprecise set of rules that require iterative improvements. We built the rules using an incremental approach where each increment adds a verb component and then improves the rules. In the incremental approach, we saw the highest level of correctness increase in the isiZulu future tense cluster rules, for all moods, after correcting the prefix in the first increment and including the root and verb extensions. A significant source of the errors in the first isiZulu increment was the remote past morpheme that was extracted from Doke [27] [26] that's said to also be capable of denoting the emphatic future tense (together with other morphemes). The future tense cluster is also the only tense that has the two features that have the potential to confuse one when judging the correctness of the strings: it is affected by phonological conditioning and has long form strings (i.e. verbs with a space). The isiXhosa future tense cluster rules, like their isiZulu counterparts, have long form verbs and require conditioning. However, unlike isiZulu, they they do not suffer from very low correctness values in the first increment because the isiZulu prefix was susceptible to more phonological conditioning than its isiXhosa counterpart (The presence of *uku* vs. *ku*). Moreover, we see that isiZulu future tense cluster has lower semantic correctness values in the second and last increment because the incorrectness due to the prefix is further increased by the overgeneration caused by verbal extensions.

The isiXhosa indicative and participial present tense rules are the only rules in the language that do not have 100% syntactic correctness at the first increment and the errors are due to the presence of the participial mood morpheme. Its removal, together with the encoding of other rules, reduced the number of strings generated by the rule cluster by approximately 30% but increased the syntactic quality to 90.6%. Also, the two mood's (i.e. indicative and participial) present and future rules are the only ones that do not have syntactic increases between all increments. There is a decrease for the two rule sets of at most 10.6% and it is attributable to addition of the final vowel and the missing perfect future that has since been addressed. In all the above cases, the reason why a morpheme results in incorrect strings should be the subject of future work in which it is possible to have potentially long and detailed consultations with linguists. Lastly, the increment with the largest impact for both languages is increment 2. It has large coverage because it adds the object concords, verb root, and verbal extensions to the prefix and then eliminates the errors that exist.

The syntactic and semantic evaluations in the incremental development process in Chapter 4 are significantly higher than the expert evaluations as detailed in Section 5.1 and this can be attributed to linguists being more knowledgeable and strict in evaluating string correctness. Moreover, the evaluations from Section 5.1 were done on a limited sample that had a variation of verb roots unlike the isiXhosa and isiZulu evaluations in Chapter 4 that used only one verb root and a larger set of strings. The semantic correctness values are lower in the Chapter 4 evaluations due to the difference in the number of evaluated strings.

The language-space quality assessment by one isiXhosa and one isiZulu linguist yielded limited results. The isiXhosa linguist evaluated all 99 strings 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 appear to be good, 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. When we devise a confusion

metric among the strings that were annotated as being incorrect, using phonological conditioning and the long form strings that have spaces, as follows $CONFUSION\_COUNT = number\ of\ consecutive\ vowels + number\ of\ spaces$, then we see that in isiXhosa $CONFUSION\_COUNT = 19$ and $CONFUSION\_COUNT = 59$ in isiZulu. Hence, it is not surprising that isiZulu are likely to be annotated as being incorrect as they are more 'confusing'. The difference in impact of phonological conditioning noise between isiZulu and isiXhosa is especially apparent in the immediate future tense (*uku-* vs *ku-*). Further, this difference between the quality of the two languages is partially due to the isiZulu linguist being more experienced in evaluating CFG outputs than the isiXhosa linguist. The leniency of the isiXhosa linguist, unlike their isiZulu counterpart, can be observed in the number of isiXhosa strings that were considered as being semantically correct despite being syntactically incorrect. IsiXhosa has 25 strings annotated as such while isiZulu has only 1. The lone isiZulu string is *isazahluka* and it is syntactically wrong due to the *-za-* that should have been, according to the linguist, *-zo-*. This shows that the difference between the isiXhosa and isiZulu quality differences can be attributed to phonological noise and the experience of the isiZulu linguist. The results are able to illuminate structural similarities between the two languages albeit in a limited fashion due to the causes of grammar quality differences.

5.4.3 Binary similarity measures

The language-space comparison of the isiZulu and isiXhosa CFG rules was done using four binary similarity measures. These measures have certain relationships, for instance, the maximum and average difference between the Sorenson and Sorgenfrei measure are 0.250 and 0.166, respectively, which is higher than the difference between the Driver-Kroeber and Jaccard ones (0.247 and 0.142). The Jaccard and Driver-Kroeber measures are more intuitive for language spaces[8], 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 5.5 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 5.5. 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 5.2.2, 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.

[8] 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

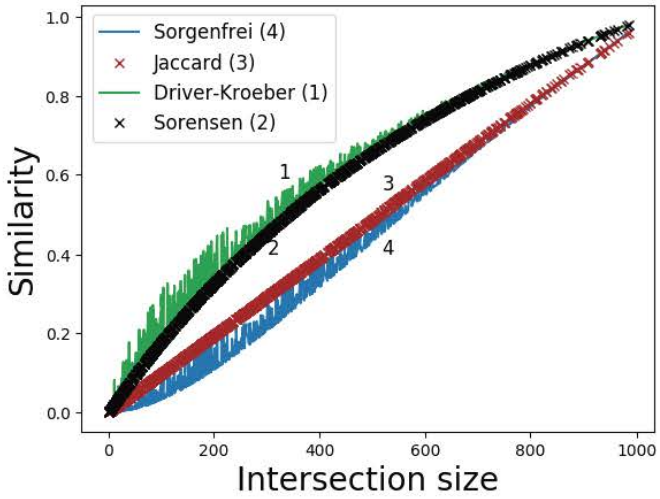


Figure 5.5: Difference in four binary similarity methods when the size of the intersection between two sets increases and the sets' complement decreases. Similarity is measured with value between zero (Different) and one (Equivalent).

5.4.4 Answering research questions

Section 1.3 listed three research questions. To answer the first question (*How grammatically similar are isiZulu verbs with their isiXhosa counterparts?*), we developed two rule sets for both languages following an incremental approach. This yielded 24 rules per language for generating the open classes. Out of these rules, we have discovered that each language differs from the other by only 5 rules. These differences are minor, and affected the prefix mostly, as discussed in Section 5.2.2.1. The quality of the developed isiZulu rules are not as good as their isiXhosa counterparts due to the isiZulu linguist not fully complying with the annotation instructions, as shown in Section 5.1.2. Nonetheless, we have found that the most intuitive measures show that isiZulu and isiXhosa are 42% (Jaccard) and 59.5% (Driver-Kroeber) similar.

With regards to the second question (*Can a singular merged set of grammar rules be used to produce correct verbs for both languages?*), we have observed from the literature that the suffix and final vowel in the two languages operate in the same manner. It is possible that there may be differences when more suffixal features are considered. At first glance, it may seem that a modularized rule set in which the prefix, verb root, and suffix would enable the exploitation of the similarities between the two languages. In particular, this would mean that only the prefix module would be differentiated between the two languages. There is a dependence between the prefix and suffix that renders such an endeavour difficult. This dependence can be observed, for instance, in the two rules for generating present subjunctive verbs for isiXhosa and isiZulu that are listed in Figure 5.6. The isiXhosa rule can only have that specific prefix (A_{pes}), and not the other present subjunctive prefix (rule X39), for that specific suffix (S_{np} a). This is different from the isiZulu rules (Z40 to Z42) where the all the prefixes (rule Z40) can be used with any suffix. The creation of a modularized system in which some modules are shared would require specific rules dictating that certain versions of a module can only appear with certain versions of the other modules. This

- $Verb \rightarrow A_{pes} \quad OC \quad VR \quad S_{np} \quad a$ (isiXhosa)
- $Verb \rightarrow Prefix \quad OC \quad VR \quad S_{np} \quad a$ (isiZulu)

Figure 5.6: Two rules for generating isiXhosa and isiZulu present subjunctive verbs. These two rules have differences.

can be achieved with additional rules, however, it may lead to bulky grammars.

With regards to the third question (*What is the degree of improvement in grammatical correctness that can be brought on by the introduction of phonological conditioning rules?*), we have conducted a mathematical quantification of the ratio of the weather forecast verbs that require grammatical improvement. Weather forecast verbs are the ones that are generated by the rules using a specific verb root, for instance, for the verb root *-band-* taken from Table 3.4.1, the isiXhosa and isiZulu context free grammar rules generate 187 isiXhosa verbs and 193 isiZulu verbs that can be used in a forecast. Our quantification determines the ratio of verbs that require phonological conditioning out of the total number of verbs that can be used in a forecast. We have discovered that the amount of strings, out of all the generated strings, that require phonological condition is high. For example, the minimum ratio of strings that will require phonological conditioning for the four types of verb roots that are listed in Table 5.3.1 is 45% for isiXhosa and 67% for isiZulu. Furthermore, when the number of verbs that will be used in the forecast increases, using a conservative ratio of two Type A verb roots for each Type D verb root, then at least 55% isiXhosa and 69% isiZulu rule generated strings will require phonological conditioning. These high values suggest that phonological conditioning can substantially improve the quality of verbs used in a weather forecasts. It should be noted that not all the generated verbs exist in the two languages, and therefore these values are only accurate for the verb correctness level that has been reported.

5.4.5 CFG development and evaluation considerations

An incremental rule development approach coupled with NLTK was followed in order to expedite rule development. This approach is effective for languages whose grammar information is scattered because it allows one to begin by piecing this information together and slowly get rid of conflicts that may exist in the various literature sources. This approach does not involve linguists in the early stages. The correctness judgements of the rules in each incremental step were done by the researcher. It uses a Nguni language speaker under the assumption that determining which words exist in the languages in the early stages is simple and should be fast for an individual who is dedicated to that particular task, unlike linguists who have other responsibilities. However, the large difference in the correctness evaluation levels from each language's third increment (detailed in Chapter 4) with each linguist's evaluation shows that non-experts should not be used for evaluating strings. Moreover, phonological conditioning noise and one's experience in evaluating CFG output significantly affects the evaluation process as discussed in Section 5.4.2. Hence at least two linguists should be used to evaluate the same strings and their agree-

ment should be measured using statistics such as Bangdiwala's B or Cohen's Kappa as linguists are also susceptible to human judgement errors. This was not done in this work as we were only able to recruit two linguists. Individuals who do not have access to a linguist should consider a computational approach that makes use of a dictionary if it available in computational form. We did not make use of this approach because there is no isiZulu/isiXhosa dictionary available in computational form and a crowd-sourced evaluation brings afore an whole different set of issues.

The grammar development process discussed in Chapter 4 revealed discrepancies between the documented rules assumed to be correct. The observed differences were expected as the literature [41, 26, 63, 65, 77, 123, 27] used to collect the verb's morpheme rules is dated. Furthermore, the theoretical linguistics approach used to establish the rules in the mentioned literature depends on the confirmation of rules by a few linguists and this is usually done through a manual process where a few examples are provided. Our work shows that a computational linguistics is able to inform us of differences that exist between traditional linguistics literature. Moreover, our approach allows us to be able state the precise morphotactics and not only provide a few examples of where a morpheme fits in the verb's slots. The reasons for the discovered discrepancies are altogether unknown and this is a challenge that needs to be addressed by linguists. Our work does not provide explanation for the various discrepancies as we were not able to recruit linguists for long feedback sessions.

Another responsive technique that ensures high-quality rules for each of the development increments is a linguist participatory verb rule design. A participatory design approach to the rules was not taken because it would require the frequent participation of the linguists thus the design and evaluation of the rules would need more time due to the communication overhead. Researchers would then conduct in-depth interviews with linguists in an iterative manner when evaluating the quality of the developed rules. Moreover, an evaluation that forces the linguists to follow instructions with respect to responses is advised. One could also show the linguist the progress they have made in the evaluation process. A completion bar in a survey as done by Keet and Khumalo [53] is sufficient. Failure to provide such indicators may result in incomplete evaluations. The linguists in this study were required to annotate two binary values (syntactic and semantic correctness) hence there were four classes to which each verb could belong. Our labelling system was chosen because it would enable us to distinguish between overgenerated strings (syntactically true and semantically false) words and the two other forms of 'wrong' words. However, we have learnt that the use of such technical labels is not advised because it is not easy to separate syntactic and semantic correctness in agglutinating natural languages, unlike programming languages. Uncomplicated labels/classes such as (grammatical + acceptable, grammatical + ambiguous, ungrammatical + understandable, and ungrammatical + unacceptable) that were used by Keet and Khumalo[53, p.150] are advised.

6
Conclusion

This chapter concludes the findings of this work and provides suggestions for future work.

6.1 Findings

The investigation into the similarity of isiXhosa and isiZulu context free grammar rules for a fragment of the verb relevant for weather forecast generation has revealed that IsiXhosa and isiZulu weather verb rules are very similar. In particular, isiXhosa and isiZulu have 42% shared strings out of all the total number of verbs. A number of differences exist in the prefix component/slot. There are also dependencies between the various verb slots and these dependencies are not shared in the two languages. For instance, the isiXhosa (X39-X41 in Figure 4.5) and isiZulu (Z40-Z42 in Figure 4.7) rules that generate present subjunctive strings have differences in the prefix component whereas the other components are the same. However, we also see that in isiXhosa there is a suffix that can only be used with one form of prefix. Moreover, if these rules were to be joined with the isiZulu rules, one would need to take into account that the isiZulu rules have an additional distinct prefix. When the number of sentence constituents or verbal features under consideration increases in the grammars, or the quality of the grammars is improved, then relationships of this nature create a significant maintainability challenge. A merged grammar is possible but may require considerable effort in maintaining.

We have devised an easy method and equations to calculate the impact of the phonological conditioning process on weather forecast verbs. This is done by quantifying the ratio of strings that have consecutive vowels from

the list of strings that can be used in weather forecasts as discussed in Section 5.4. We see that verb roots that do not start or end with a vowel (Type D in Table 5.3.6) result in weather forecast verbs that have lower rates of consecutive vowels. For isiXhosa, for instance, 45% of the total number of strings that are generated by the CFG rules using a Type D verb root require phonological conditioning. When the number of verbs that will appear in the weather forecast increases, using a conservative 2:1 distribution of the Type A and D verb roots, then the number of generated weather forecast strings that require phonological conditioning also increases to 55% for isiXhosa and 69% for isiZulu strings. This shows that a significant number of strings that can be used in weather forecasts have consecutive vowels and need phonological conditioning to improve their grammatical correctness. However, the readability of forecasts may not be significantly affected because they need not use a large number of verbs per forecast.

6.2 Implications

This is the first work, to the best of our knowledge, that presents verb CFGs that include more than one tense. We also present the first computational approach, and results, for verb similarity comparison between isiZulu and isiXhosa. The presented language-space similarity method could be applied to quantify similarity between any two languages that share at least one base-form string.

The evaluation of the developed weather verb grammars turned out limited results. In it, we see a difference in isiXhosa and isiZulu grammar quality that is attributable to the experience of the different linguists with evaluations of this nature. Nonetheless, we see significant structural similarities between the two languages, especially slot-wise. This suggests that an existing isiZulu grammar, after replacing the closed class morphemes, can generate a significant number of valid isiXhosa strings. Of course it would still require modifications in order to be fully compatible with isiXhosa. This also holds when the roles of the mentioned languages are reversed.

The similarity in corresponding slots between the two languages means that one should be able to create an abstract component, e.g. suffix, that can model both languages' component morphotactics, with the concrete values being inserted when a language is chosen. Thus a "singular" abstract grammar could be used by an engine to generate the two languages. The extent to which such a solution in tenable is mostly dependent on the prefix-suffix relationship that is yet to be investigated.

Lastly, any grammar that is built to generate isiXhosa and isiZulu weather verbs should not only model morphotactics but phonological conditioning as well. This is because the latter affects a large number of the generated string and morphological alternation affects the understandability of strings.

6.3 Future work

More work is required to improve the quality of the rules and this could be done through early stage linguist involvement when refining the rules. An insightful and labour-intensive approach of this nature would be such that linguists are responsible for examining the rules and harmonising them where possible. The rules would then be

tested computationally and then re-evaluated by linguists. This is an approach that is common in computational linguistics but has only been used sparingly so far for isiZulu and isiXhosa. This may explain, at least partially, the errors observed in the published literature that follows the traditional grammar approach of linguistics. Future work should recruit lingusists for potentially lengthy collaboration sessions in order to extensively discuss possible disagreements in Nguni grammar literature. Alternatively, the incremental approach used in this work should be followed and it should have at least two iterations and at the end of each iteration there should be in-depth linguist feedback sessions. It would also be insightful to compile a list of expected strings for each sampled set of grammar generated strings for comparison. This ensures that the grammar generates an exact string when given specific features and does not include unexpected morphemes. The list of expected strings can compiled by linguists during these feedback sessions as non-experts may not be knowledgeable enough to compile such lists. Generally, more research is required to establish an effective methodology that can result in high-quality grammar rules with less linguist involvement and this might be achieved through the crowd-sourcing of large amounts of Nguni weather forecast text in order to create corpus-induced grammars. Additional work is also necessary to accurately quantify the phonological conditioning process. For instance, the use of a computational form of a dictionary to reduce the current estimate by only considering words that exist in the language can be investigated. On a broader level, more focus should be put on building a tool with a similar architecture to Dokkara et al. [30] that can load CFG rules from external files.

A

Supplementary data

Collection of isiXhosa translated weather sentences:

1. Lipholile kumkhwezo wonxweme apho kulindeleke izibhaxu zenkungu yakusasa ngaphaya koko liyakuthi gqabagqaba ngamafu kwaye libeshushu okanye litshise kwaye libeneziphango ezithe saa emantla.
2. Inkungu yakusasa embindini, ngaphaya koko liyakuthi fakafaka ngamafu kwaye liphole lide libande, ze lizole kwaye libande.
3. Liyakuthi gqabagqaba ngamafu kwaye libeshushu
4. Umoya kumkhwezo wonxweme uyakubaphakathi ukuya kumoya ohlaziyayo ovela ngakumantla empuma
5. Umoya kumkhwezo wonxweme uyakubaphakathi ukuya kohlaziyayo womzantsi-ntshona nowomzantsi
6. Liyakuthi gqabagqaba ngamafu linezibhaxu zenkungu kumantla, ngaphaya koko liyakubalihle kwaye liphole okanye libeshushu
7. Lizakuthi gqabagqaba ngamafu kwaye lipholile lishushu likwanamafu athe gqabagqaba phezu wehehehu nakumzantsi.
8. Izibhaxu zenkungu zilindelekile kumbindi, ngapha koko amafu athe qabagqaba kwaye nobushushu obunemvula yethutyana ethe saa kwaneziphango ezithe saa kodwa ezizodwa ezikwiqondo eliphezulu emntla-mpuma.
9. Izibhaxu zenkungu, ngapha koko ezithe gqabagqaba ngamafu kwaye zipholile.
10. Umoya kunxweme uyakubaphakathi ukuya kumoya ohlaziyayo womzantsi ukuya kowentshona, kodwa iyakubangumoya okhaphukhaphu ukuya kophakathi womzantsi mpuma entshonalanga kuqala.
11. Izibhaxu zenkungu yakusasa emzantsi, ngaphaya koko liyakubalihle kwaye litshisekakhulu, liye lisithi gqabagqaba ngamafu ngemvakwemini kodwa libenamafu emva kwemini kumkhwezo wonxweme.

12. Izibhaxu zenkungu kusasa kumkhwezo wonxweme, ngapha koko liyakuthi gqabagqaba ngamafu kwaye libeshushu, kodwa liyakuthi gqabagqa ngamafu lishushu kwaye lshuhsu emzantsi.
13. Umoya kumkhwezo wonxweme kumzantsi-ntshona ukuya kumzantsi uyakubangophakathi ukuya kopho-lileyo.
14. Liyakubanamafu kwaye libeshushu lineziphango kodwa zithi saa empuma.
15. Liyakubanenkungu kumkhwezo wonxweme kuqala, ngapha koko liyakubalihle kwaye lishushu kodwa lit-shisa ngamandla kwiindawo eziphakathi embindini.
16. Liyakuthi gqabagqaba ngamafu empuma kuqala, lineziphango ezithe saa ngaphaya koko liyakubalihle kwaye lipholile, kodwa libanda emzantsi.
17. Liyakuthi gqabagqaba ngamafu kwaye litshisa ngamandla likwaneziphango ezithe saa kwimpuma esekudeni
18. Liyakuthi qabagqaba ngamafu kwaye litshisa ngamandla, linezineziphango ezithe saa kwanemvula yethutyana kwimpuma
19. Liyakubanamafu kwaye litshisa lineziphango ezithe saa neziqeleleneyo, kodwa zigqagqene empuma
20. Liyakubaneqabaka kusasa kumbindi, ngaphaya koko liyakubalihle kwaye liphole kodwa litshisa kwiindawo ezikumkhwezo wonxweme lwasentshona—liyakutshisa ngamandla kwiindawo ezikumntla-ntshona
21. Umoya kumkhwezo wonxweme uyakubakhaphukhaphu ukuya kumoya wempuma ophakathi oyakuye ubangu-moya ophakathi wentshona ukuya kumzantsi-ntshona ususela kwintshona.
22. Liyakubanezibhaxu zenkungu ekuseni ekudeni komntla-mpuma, ngaphaya koko liyakubalihle kwaye lip-hole.
23. Lihle kwaye litshisa.
24. Lihle kwaye lipholile, kodwa libanda kumzantsi
25. Iziphango ezithe gqagqa kwanemvula zilindelekile kumkhwezo womzantsi-ntshona kananjalo nakunxweme lomzantsi.
26. Kwindawo emelene nombindi apho kulindeleke imvula emandla nezikhukhula eziwingingqi zinokuhla.
27. Umoya kumkhwezo wonxweme uyakubangophakathi ukuya kumoya ohlaziyayo womzantsi.
28. Lihle kwaye litshisa , kodwa lithe gqabagqaba ngamafu ekudeni komzantsi
29. Umoya kumkhwezo wonxweme uyakuphola ukusuka emzantsi ukuya emzantsi-mpuma
30. Linamafu ukuya kumafu athe gqabagqaba kwaye lipholile kwaye likwaneziphango ezithe saa
31. Kumntla ngaphaya koko ziyakuthi saa kunxweme lomzantsi mpuma nakumbindi omelene nonxweme pho imvulala ezimandla ziyakulindeleka.
32. Liyakuthi gqagqaba ngamafu entshona nase mzantsi-ntshona kodwa liyakubanda emzantsi-ntshona lik-waneziphango ezizodwa ezithe saa kodwa zithe gqagqa phaya kwingingqi ye-Edeni, apho izikhukhula ezi-mandla ziyakulindeleka, zibuye zithe ngemvakwemini.
33. Liyakuthi gqabagqaba ngamafu kwaye lipholile
34. Iimvula ezina ngamandla ezikhokhelela kwizikhukhula ilinendelekile kwiindawo ezikumkhwezo wonxweme, kwaye nakwiindawo ezimelelene nombindi apho liyaakubanamafu kumkhwezo wonweme nalapho kuyakuthi kuphole kunemkhumezelo ethe saa kwanemvula, ngaphaya koko liyakuthi gqabagqaba ngamafu likwashushu.
35. Umoya kumkhwezo wonxweme uyakubakhaphukhaphu ukuya kophakathi kumzantsi-ntshona uye usiba

kwimpuma ngemvakwemini, liyakubanamafu kumkhwezo wonxweme lasemzantsi kuqala, ngaphaya koko liyakuthi gqabagqaba ngamafu kwaye liphole likwashushu.

36. Liyakubalihle kwaye litshisa kodwa lithi gqabagqaba ngamafu kwintshona esekudeni ngemvakwemini
37. Umoya kumkhwezo wonxweme uyakubakhaphukhaphu ukuya kopholileyo kumzantsi-ntshona
38. Kuyakubakho ubushushu obukhulu emzantsi-ntshona, ngaphaya koko liyakuthi gqabagqaba ngamafu kwaye litshise linemikhumezelo ethe saa kwaye likwaneziphango ezithe saa kodwa hayi kumntla-mpuma.
39. Liyakubanamafu libanda likwanemikhumezelo kwaneziphango ezithe saa kodwa umoya kumkhwezo wonxweme uyakubakhaphukhu ukuya kophakathi kumzantsi-ntshona kodwa uhlaziya emzantsi, liyakubalihle kwaye lipholile, kodwa libanda kumkhwezo wonxweme linezibhaxu zenkungu.

Bibliography

[1] I. Adeyanju. "Generating Weather Forecast Texts with Case Based Reasoning". In: *ArXiv e-prints* (Sept. 2015).

[2] S. Axelrod. "Natural language generation in the IBM flight information system". In: *Proceedings of the ANLP-NAACL 2000 Workshop on Conversational Systems*. Association for Computational Linguistics. Seattle, Washington, U.S.A, 2000, pp. 21–26.

[3] J. Bateman and M. Zock. *Bateman/Zock list of NLG systems*. http://www.nlg-wiki.org/systems/. Accessed: 14 April 2016. 2012.

[4] M. Bavali and F. Sadighi. "Chomsky's Universal Grammar and Halliday's Systemic Functional Linguistics: An Appraisal and a Compromise." In: *Journal of Pan-Pacific Association of Applied Linguistics* 12.1 (2008), pp. 11–28.

[5] K. R. Beesley and L. Karttunen. *Finite State Morphology*. Vol. 3. CSLI Studies in Computational Linguistics. CSLI Publications, 2003.

[6] A. Belz. "Corpus-driven generation of weather forecasts". In: *Proceedings from the Corpus Linguistics Conference Series*. Vol. 1. University of Birmingham. Birmingham, United Kingdom, July 2005.

[7] A. Belz. "Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models". In: *Natural Language Engineering* 14.04 (2008), pp. 431–455.

[8] A. Berg, R. P. L. Pretorius, M. Butt, and T. H. King. "The representation of Setswana double objects in LFG". In: *LFG 2013 - Proceedings of the 18th International Lexical Functional Grammar Conference, July 18-20, 2013, Debrecen, Hungary*. Ed. by M. Butt and T. H. King. CSLI Publications, 2013, pp. 111–130.

[9] A. Berg, R. Pretorius, and L. Pretorius. "Exploring the treatment of selected typological characteristics of Tswana in LFG". In: *LFG 2012 - Proceedings of the 17th International Lexical Functional Grammar Conference, July 28- July 01, 2012, Denpasar, Bali*. Ed. by M. Butt and T. H. King. CSLI Publications, 2012, pp. 85–98.

[10] C. Blyth. "A Constructivist Approach to Grammar: Teaching Teachers to Teach Aspect". In: *The Modern Language Journal* 81.1 (1997), pp. 50–66.

[11] M. Bollmann. "Adapting SimpleNLG to German". In: *ENLG 2011 - Proceedings of the 13th European Workshop on Natural Language Generation, 28-30 September 2011, Nancy, France*. Ed. by C. Gardent and K. Striegnitz. The Association for Computer Linguistics, 2011, pp. 133–138.

[12] L. Bourbeau, D. Carcagno, E. Goldberg, R. Kittredge, and A. Polguère. "Bilingual Generation of Weather Forecasts in an Operations Environment". In: *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*. COLING '90. Helsinki, Finland: Association for Computational Linguistics, 1990, pp. 318–320.

[13] W. Bourquin. "Notes on the concords in Xhosa and Zulu, their differences and general aspect". In: *African Studies* 11.1 (1952), pp. 16–28.

[14] S. Boyd. "TREND: a system for generating intelligent descriptions of time series data". In: *IEEE International Conference on Intelligent Processing Systems*. ICIPS1998. Gold Coast, Australia: Griffith University, 1998.

[15] B. Bringert, K. Angelov, and A. Ranta. "Grammatical Framework Web Service". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. EACL '09. Athens, Greece: Association for Computational Linguistics, 2009, pp. 9–12.

[16] K. Brown, J. Bateman, and J. Delin. "Rhetorical Structure Theory". In: *Encyclopedia of Language & Linguistics* (2006), pp. 589–597.

[17] J. Byamugisha, C. M. Keet, and B. DeRenzi. "Tense and Aspect in Runyankore Using a Context-Free Grammar". In: *INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK*. Ed. by A. Isard, V. Rieser, and D. Gkatzia. The Association for Computer Linguistics, 2016, pp. 84–88.

[18] N. N. Canonici. *The grammatical structure of Zulu*. Department of Zulu Language and Literature, University of Natal-Durban, 1987.

[19] S. Choi, S. Cha, and C. C. Tappert. "A survey of binary similarity and distance measures". In: *Journal of Systemics, Cybernetics and Informatics* 8.1 (2010), pp. 43–48.

[20] R. Dale. "An introduction to natural language generation". `http://comp.mq.edu.au/~rdale/teaching/esslli/`. Presentation at European Summer School in Logic, Language and Information. 1995.

[21] R. Dale and E. Reiter. "Building natural language generation systems". In: *Cambridge University Press* (2000).

[22] V. Dallmeier, C. Lindig, and A. Zeller. "Lightweight Defect Localization for Java". In: *ECOOP 2005 - Object-Oriented Programming, 19th European Conference, Glasgow, UK, July 25-29, 2005, Proceedings*. Ed. by A. P. Black. Vol. 3586. Lecture Notes in Computer Science. Springer, 2005, pp. 528–550.

[23] S. A. Davey. "The moods and tense of the verb in Xhosa". MA thesis. Department of Bantu Languages, University of South Africa, 1973.

[24] K. van Deemter, M. Theune, and E. Krahmer. "Real versus Template-Based Natural Language Generation: A False Opposition?" In: *Computational Linguistics* 31.1 (2005), pp. 15–24.

[25] L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302.

[26] C. M. Doke. *Textbook of Zulu grammar*. 2rd. Longmans Southern Africa, 1931.

[27] C. M. Doke. *Textbook of Zulu grammar*. 6th. Maskew Miller Longman, 1992.

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

[29] S. R. S. Dokkara, S. V. Penumathsa, and S. G. Sripada. "Verb Morphological Generator for Telugu". In: *Indian Journal of Science and Technology* 10.13 (2017).

[30] S. R. S. Dokkara, S. V. Penumathsa, and S. G. Sripada. "A Simple Surface Realization Engine for Telugu". In: *ENLG 2015 - Proceedings of the 15th European Workshop on Natural Language Generation, 10-11 September 2015, University of Brighton, Brighton, UK*. Ed. by A. Belz, A. Gatt, F. Portet, and M. Purver. The Association for Computer Linguistics, 2015, pp. 1–8.

[31] H. E. Driver and A. L. Kroeber. *Quantitative Expression of Cultural Relationships*. University of California Publications in American Archaeology and Ethnology. University of California Press, 1932.

[32] A. Gatt and E. Krahmer. "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation". In: *ArXiv e-prints* (Mar. 2017). arXiv: 1703.09902.

[33] A. Gatt and E. Reiter. "SimpleNLG: A Realisation Engine for Practical Applications". In: *ENLG 2009 - Proceedings of the 12th European Workshop on Natural Language Generation, March 30-31, 2009, Athens, Greece*. Ed. by E. Krahmer and M. Theune. The Association for Computer Linguistics, 2009, pp. 90–93.

[34] A. Gatt and E. Reiter. "SimpleNLG: A Realisation Engine for Practical Applications". In: *Proceedings of the 12th European Workshop on Natural Language Generation*. ENLG '09. Athens, Greece: Association for Computational Linguistics, 2009, pp. 90–93.

[35] K. W. Getao and E. K. Miriti. "Computational Modelling in Bantu Language". In: *Special topics in computing and ICT research : Advances in Systems Modelling and ICT Applications* (2006).

[36] D. Gkatzia, O. Lemon, and V. Rieser. "Natural Language Generation enhances human decision-making with uncertain information". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics, 2016.

[37] H. R. Glahn. "Computer-Produced Worded Forecasts". In: *Bulletin of the American Meteorological Society* 51.12 (1970), pp. 1126–1131.

[38] E. Goldberg, N. Driedger, and R. I. Kittredge. "Using natural-language processing to produce weather forecasts". In: *IEEE Expert* 9.2 (1994), pp. 45–53.

[39] E. Goldberg, R. Kittredge, and A. Polguere. "Computer generation of marine weather forecast text". In: *Journal of atmospheric and oceanic technology* 5.4 (1988), pp. 473–483.

[40] Government of South Africa. *National Language Policy Framework Final Draft*. http://www.gov.za/sites/www.gov.za/files/langpolicyfinal_0.pdf. Accessed: 24 January 2017. 2012.

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

[42] A. S. Grover, G. B. Van Huyssteen, and M. W. Pretorius. "South African human language technologies audit". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation: LREC 2010*. Ed. by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias. European language resources distribution agency, May 2010.

[43] A. S. Grover, G. B. Van Huyssteen, and M. W. Pretorius. "The South African human language technology audit". In: *Language resources and evaluation* 45.3 (2011), pp. 271–288.

[44] B. Gyawali. "Surface Realisation from Knowledge Bases". PhD thesis. Universite de Lorraine, France, Jan. 2016.

[45] P. Jaccard. "The Distribution of the Flora in the Alpine Zone". In: *The New Phytologist* 11.2 (1912), pp. 37–50.

[46] G. Jäger and J. Rogers. "Formal language theory: refining the Chomsky hierarchy". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367.1598 (2012), pp. 1956–1970.

[47] S. Kahane. "What Is a Natural Language and How to Describe It? Meaning-Text Approaches in Contrast with Generative Approaches (Invited Talk)". In: *Computational Linguistics and Intelligent Text Processing, Second International Conference, CICLing 2001, Mexico-City, Mexico, February 18-24, 2001, Proceedings*. Ed. by A. F. Gelbukh. Vol. 2004. Lecture Notes in Computer Science. Springer, 2001, pp. 1–17.

[48] M. Kay. "Nonconcatenative Finite-State Morphology". In: *EACL 1989, 3rd Conference of the European Chapter of the Association for Computational Linguistics, April 1-3, 1987, University of Copenhagen, Copenhagen, Denmark*. Ed. by B. Maegaard. The Association for Computer Linguistics, 1987, pp. 2–10.

[49] C. M. Keet. "An assessment of orthographic similarity measures for several African languages". In: *CoRR* abs/1608.03065 (2016).

[50] C. M. Keet and L. Khumalo. "Basics for a Grammar Engine to Verbalize Logical Theories in isiZulu". In: *Rules on the Web. From Theory to Applications - 8th International Symposium, RuleML 2014, Co-located with the 21st European Conference on Artificial Intelligence, ECAI 2014, Prague, Czech Republic, August 18-20*. 2014, pp. 216–225.

[51] C. M. Keet and L. Khumalo. "Toward Verbalizing Ontologies in isiZulu". In: *Controlled Natural Language - 4th International Workshop, CNL 2014, Galway, Ireland, August 20-22, 2014. Proceedings*. Ed. by B. Davis, K. Kaljurand, and T. Kuhn. Vol. 8625. Lecture Notes in Computer Science. Springer, 2014, pp. 78–89.

[52] C. M. Keet and L. Khumalo. "Grammar rules for the isiZulu complex verb". In: *Southern African Linguistics and Applied Language Studies* 35.2 (2017), pp. 183–200.

[53] C. M. Keet and L. Khumalo. "Toward a knowledge-to-text controlled natural language of isiZulu". In: *Language Resources and Evaluation* 51.1 (2017), pp. 131–157.

[54] S. M. Kerpedjiev. "Automatic Generation of Multimodal Weather Reports from Datasets". In: *Proceedings of the Third Conference on Applied Natural Language Processing*. ANLC '92. Trento, Italy: Association for Computational Linguistics, 1992, pp. 48–55.

[55] L. Khumalo. "An analysis of the Ndebele passive construction". PhD thesis. University of Oslo, Norway, 2007.

[56] S. Kim, H. Alani, W. Hall, P. H. Lewis, D. E. Millard, N. R. Shadbolt, and M. J. Weal. "Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web". In: *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation & Knowledge Markup, Lyon, July 22-26, 2002*. 2002.

[57] R. Kittredge, A. Polguère, and E. Goldberg. "Synthesizing Weather Forecasts from Formated Data". In: *Proceedings of the 11th Coference on Computational Linguistics*. COLING '86. Bonn, Germany: Association for Computational Linguistics, 1986, pp. 563–565.

[58] M. Koleva and D. Klakow. "Towards Adaptation of NLP Tools for Closely-Related Bantu Languages: Building a Part-of-Speech Tagger for Zulu". PhD thesis. Master's thesis, Saarland University, Germany, 2013.

[59] S. F. Liang, R. Stevens, D. Scott, and A. Rector. "Automatic Verbalisation of SNOMED Classes Using OntoVerbal". In: *Artificial Intelligence in Medicine: 13th Conference on Artificial Intelligence in Medicine, AIME 2011, Bled, Slovenia, July 2-6, 2011. Proceedings*. Ed. by M. Peleg, N. Lavrač, and C. Combi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 338–342.

[60] S. F. Liang, R. Stevens, D. Scott, and A. L. Rector. "OntoVerbal: a Protégé plugin for verbalising ontology classes". In: *Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series, Graz, Austria, July 21-25, 2012*. Ed. by R. Cornet and R. Stevens. Vol. 897. CEUR Workshop Proceedings. CEUR-WS.org, 2012.

[61] J. Maho. *A comparative study of Bantu noun classes*. Acta Universitatis Gothoburgunsis, 1999.

[62] A. Mazzei, C. Battaglino, and C. Bosco. "SimpleNLG-IT: adapting SimpleNLG to Italian". In: *INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK*. Ed. by A. Isard, V. Rieser, and D. Gkatzia. The Association for Computer Linguistics, 2016, pp. 184–192.

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

[64] J. McLaren. *A Xhosa Grammar, revised and re-written in the new orthography, edited by G. H. Welsh*. Longmans, Green and Company, 1944.

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

[66] A. E. Meeussen. "Bantu grammatical reconstructions". In: *Africana linguistica* 3.1 (1967), pp. 79–121.

[67] R. Mitkov. "Generating public weather reports". In: *Proceedings of Current Issues in Computational Linguistics. Penang, Malaysia* (1991).

[68] F. S. M. Mncube. *Xhosa manual*. Juta and Company, 1955.

[69] H. L. Nabe, P. W. Dreyer, and G. L. Kakana. *Xhosa Dictionary: English, Xhosa, Afrikaans*. Educum Publishers, 1976.

[70] W. Ng'ang'a. "Swahili Inflectional Morphology for the Grammatical Framework". In: *Proceedings of the Conference on Human Language Technology for Development*. HLTD 2011. Athens, Greece: Center for Language Engineering, Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, 2009, pp. 100–105.

[71] W. Ng'ang'a. "Building Swahili Resource Grammars for the Grammatical Framework". In: *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday*. Ed. by D. Santos, K. Lindén, and W. Ng'ang'a. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 215–226.

[72] D. Nurse. *Tense and aspect in Bantu*. Oxford University Press, 2008.

[73] D. Nurse and G. Philippson. *The Bantu languages*. Routledge, 2003.

[74] R. de Oliveira and S. Sripada. "Adapting SimpleNLG for Brazilian Portuguese realisation". In: *INLG 2014 - Proceedings of the Eighth International Natural Language Generation Conference, Including Proceedings of the INLG and SIGDIAL 2014 Joint Session, 19-21 June 2014, Philadelphia, PA, USA*. Ed. by M. Mitchell, K. F. McCoy, D. McDonald, and A. Cahill. The Association for Computer Linguistics, 2014, pp. 93–94.

[75] M. C. Peel, B. L. Finlayson, and T. A. McMahon. "Updated world map of the Köppen-Geiger climate classification". In: *Hydrology and earth system sciences discussions* 4.2 (2007), pp. 439–473.

[76] C. S. Peirce. "The numerical measure of the success of predictions". In: *Science* ns-4.93 (1884), pp. 453–454.

[77] A. M. Peters. "A computer oriented generative grammar of the Xhosa verb". PhD thesis. University of Wisconsin, USA, 1966.

[78] D. Posel and J. Zeller. "Home language and English language ability in South Africa: Insights from new data". In: *Southern African Linguistics and Applied Language Studies* 29.2 (2011), pp. 115–126.

[79] L. Pretorius and S. Bosch. "Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele". In: *Language Technology for Normalisation of Less-Resourced Languages* (2012), p. 73.

[80] L. Pretorius and S. E. Bosch. "Finite-State Computational Morphology: An Analyzer Prototype For Zulu". In: *Machine Translation* 18.3 (2003), pp. 195–216.

[81] L. Pretorius and S. E. Bosch. "Containing overgeneration in Zulu computational morphology". In: *Southern African Linguistics and Applied Language Studies* 26.2 (2008), pp. 209–216.

[82] L. Pretorius and S. E. Bosch. "Finite State Morphology of the Nguni Language Cluster: Modelling and Implementation Issues". In: *Finite-State Methods and Natural Language Processing, 8th International Workshop, FSMNLP 2009, Pretoria, South Africa, July 21-24, 2009, Revised Selected Papers*. Ed. by A. Yli-Jyräs, A. Kornai, J. Sakarovitch, and B. W. Watson. Vol. 6062. Lecture Notes in Computer Science. Springer, 2009, pp. 123–130.

[83] L. Pretorius, L. Marais, and A. Berg. "A GF miniature resource grammar for Tswana: modelling the proper verb". In: *Language Resources and Evaluation* 51.1 (2017), pp. 159–189.

[84] L. Pretorius, B. Viljoen, R. Pretorius, and A. Berg. "Towards a computational morphological analysis of Setswana compounds". In: *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies* 29.1 (2008), pp. 1–20.

[85] L. Pretorius, B. Viljoen, R. Pretorius, and A. Berg. "A finite state approach to Setswana verb morphology". In: *Finite-State Methods and Natural Language Processing* (2010), pp. 131–138.

[86] R. Pretorius, A. Berg, L. Pretorius, and B. Viljoen. "Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography". In: *Proceedings of the First Workshop on Language Technologies for African Languages*. Association for Computational Linguistics. 2009, pp. 66–73.

[87] R. Pretorius, B. Viljoen, and L. Pretorius. "A finite-state morphological analysis of Tswana nouns". In: *South African Journal of African languages* 25.1 (2005), pp. 48–58.

[88] A. Ranta. "Grammatical Framework". In: *Journal of Functional Programming* 14.2 (2004), pp. 145–189.

[89] A. Ranta. *Grammatical Framework Tutorial*. http://www.grammaticalframework.org/doc/tutorial/gf-tutorial.html#toc4. Accessed: 11 June 2017. 2010.

[90] A. Ranta, A. El Dada, and J. Khegai. "The GF resource grammar library". In: *Linguistic Issues in Language Technology* 2.2 (2009), pp. 1–63.

[91] E. Reiter. "NLG vs. Templates". In: *eprint arXiv:cmp-lg/9504013*. Apr. 1995.

[92] E. Reiter. "An Architecture for Data-to-text Systems". In: *Proceedings of the Eleventh European Workshop on Natural Language Generation*. ENLG '07. Germany: Association for Computational Linguistics, 2007, pp. 97–104.

[93] E. Reiter. *You Need to Understand your Corpora! The Weathergov Example*. https://ehudreiter.com/2017/05/09/weathergov. Accessed: 2017-12-28. May 2017.

[94] E. Reiter and R. Dale. "Building applied natural language generation systems". In: *Natural Language Engineering* 3.01 (1997), pp. 57–87.

[95] E. Reiter and S. Sripada. "Should Corpora Texts Be Gold Standards for NLG?" In: *Proceedings of the 2nd International Conference on Natural Language Generation*. 2002, pp. 97–104.

[96] E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. "Choosing words in computer-generated weather forecasts". In: *Artif. Intell.* 167.1-2 (2005), pp. 137–169.

[97] D. J. Rogers, T. T. Tanimoto, et al. "A computer program for classifying plants." In: *Science (Washington)* 132 (1960), pp. 1115–18.

[98] M. E. Ross. "Designing and Using Program Evaluation as a Tool for Reform". In: *Journal of Research on Leadership Education* 5.12 (2010), pp. 481–500.

[99] A. P. Rovai. "A practical framework for evaluating online distance education programs". In: *The Internet and Higher Education* 6.2 (2003), pp. 109–124.

[100] R. Rubinoff. "Integrating text planning and linguistic choice by annotating linguistic structures". In: *Aspects of Automated Natural Language Generation: 6th International Workshop on Natural Language Generation Trente, Italy, April 5–7 1992 Proceedings*. Ed. by R. Dale, E. Hovy, D. Rösner, and O. Stock. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 45–56.

[101] D. P. Ruth and M. R. Peroutka. "The interactive computer worded forecast". In: *Preprints, 9th International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*. Anaheim, CA, USA: American Meteorological Society, 1993, pp. 321–326.

[102] F. Schilder, B. Howald, and R. Kondadadi. "Gennext: A consolidated domain adaptable NLG system". In: *Proceedings of the 14th European Workshop on Natural Language Generation*. 2013, pp. 178–182.

[103] P. Schonstein. *Xhosa: a cultural grammar for beginners*. African Sun Press, 1994.

[104] M. wa Selepe. "My Name". In: *Explorings: A Collection of Poems for the Young People of Southern Africa*. Ed. by R. Malan. New Africa Books, 1988, p. 11.

[105] M. Setati, J. Adler, Y. Reed, and A. Bapoo. "Incomplete journeys: Code-switching and other language practices in mathematics, science and English language classrooms in South Africa". In: *Language and education* 16.2 (2002), pp. 128–149.

[106] S. M. Shieber. "Evidence Against the Context-Freeness of Natural Language". In: *The Formal Complexity of Natural Language*. Ed. by W. J. Savitch, E. Bach, W. Marsh, and G. Safran-Naveh. Dordrecht: Springer Netherlands, 1987, pp. 320–334.

[107] G. Sibanda. "Vowel processes in Nguni: Resolving the problem of unacceptable VV sequences". In: *Selected Proceedings of the 38th Annual Conference on African Linguistics, Gainesville, Florida, March 22-25, 2007*. Ed. by M. Matondo, F. M. Laughlin, and E. Potsdam. Cascadilla Proceedings Project, Somerville, Massachusetts, USA, 2007, pp. 38–55.

[108] B. Sigurd, C. Willners, M. Eeg-Olofsson, and C. Johansson. "Deep Comprehension, Generation and Translation of Weather Forecasts (Weathra)". In: *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*. COLING '92. Nantes, France: Association for Computational Linguistics, 1992, pp. 749–755.

[109] M. H. Smith, R. Garigliano, and R. G. Morgan. "Generation in the LOLITA System: An Engineering Approach". In: *Proceedings of the Seventh International Workshop on Natural Language Generation*. INLG '94. Kennebunkport, Maine: Association for Computational Linguistics, 1994, pp. 241–244.

[110] M. Smouse, S. Gxilishe, J. de Villiers, and P. de Villiers. "Children's acquisition of subject markers in isiXhosa". In: *Pronouns and Clitics in Early Acquisition. Mouton DeGruyter, Berlin/New York* (2012), pp. 209–236.

[111] T. Sorgenfrei. "Molluscan assemblages from the marine middle Miocene of South Jutland and their environments". In: *Danmarks geologiske undersoegelse* 2.79 (1959), pp. 403–408.

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

[113] S. Spiegler, B. Golénia, K. Shalonova, P. A. Flach, and R. C. F. Tucker. "Learning the morphology of Zulu with different degrees of supervision". In: *2008 IEEE Spoken Language Technology Workshop, SLT 2008, Goa, India, December 15-19, 2008*. Ed. by A. Das and S. Bangalore. IEEE, 2008, pp. 9–12.

[114] S. Spiegler, A. van der Spuy, and P. A. Flach. "Ukwabelana - An open-source morphological Zulu corpus". In: *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*. Ed. by C. Huang and D. Jurafsky. Tsinghua University Press, 2010, pp. 1020–1028.

[115] R. E. Sripada S.G. and I. Davy. "SUMTIME-MOUSAM: Configurable Marine Weather Forecast Generator". In: *Expert Update* 6.3 (2003), pp. 4–10.

[116] S. G. Sripada, E. Reiter, I. Davy, and K. Nilssen. "Lessons from Deploying NLG Technology for Marine Weather Forecast Text Generation". In: *Proceedings of the 16th European Conference on Artificial Intelligence*. ECAI'04. Valencia, Spain: IOS Press, 2004, pp. 760–764.

[117] S. G. Sripada, E. Reiter, J. Hunter, J. Yu, and I. P. Davy. "Modelling the Task of Summarising Time Series Data Using KA Techniques". In: *Applications and Innovations in Intelligent Systems IX: Proceedings of ES2001, the Twenty-first SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 2001*. Ed. by A. Macintosh, M. Moulton, and A. Preece. London: Springer London, 2002, pp. 183–196.

[118] S. Sripada, N. Burnett, R. Turner, J. Mastin, and D. Evans. "A Case Study: NLG meeting Weather Industry Demand for Quality and Quantity of Textual Weather Forecasts". In: *Proceedings of the 8th International Natural Language Generation Conference (INLG)*. Philadelphia, Pennsylvania, U.S.A: Association for Computational Linguistics, 2014, pp. 1–5.

[119] Statistics South Africa. *2011 Census : Census in brief*. http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf. Accessed: 03 May 2016. 2012.

[120] Statistics South Africa. *Census 2011 Provincial Profile: Eastern Cape*. http://www.statssa.gov.za/publications/Report-03-01-71/Report-03-01-712011.pdf. Accessed: 20 February 2017. 2014.

[121] Statistics South Africa. *Census 2011 Provincial profile: KwaZulu-Natal*. http://www.statssa.gov.za/publications/Report-03-01-74/Report-03-01-742011.pdf. Accessed: 20 February 2017. 2014.

[122] A. J. Stent. *Aspects of Natural Language Generation*. Tech. rep. 701. Rochester, NY, USA: University of Rochester, 1998.

[123] P. Taljaard and S. Bosch. *Handbook of isiZulu*. JL van Schaik (Pty) Ltd, 1988.

[124] K. T. Taraldsen. "The nanosyntax of Nguni noun class prefixes and concords". In: *Lingua* 120.6 (2010), pp. 1522–1548.

[125] M. Theune. *Natural Language Generation for dialogue: system survey*. Tech. rep. TR-CTIT-03-22. Enschede, the Netherlands: University of Twente, Centre for Telematics and Information Technology (CTIT), May 2003.

[126] R. Todeschini, V. Consonni, H. Xiang, J. D. Holliday, M. Buscema, and P. Willett. "Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets". In: *Journal of Chemical Information and Modeling* 52 (2012), pp. 2884–2901.

[127] R. Turner, S. Sripada, E. Reiter, and I. Davy. "Using Spatial Reference Frames to Generate Grounded Textual Summaries of Georeferenced Data". In: *INLG 2008 - Proceedings of the Fifth International Natural Language Generation Conference, June 12-14, 2008, Salt Fork, Ohio, USA*. Ed. by M. White, C. Nakatsu, and D. McDonald. The Association for Computer Linguistics, 2008.

[128] *University of KwaZulu-Natal launches two books and innovative isiZulu language technologies*. https://zu.oxforddictionaries.com/explore/kwazulu-natal-isizulu-technologies. Accessed: 11-July-2017.

[129] *Use mail merge to personalize letters for bulk mailings*. https://support.office.com/en-us/article/Use-mail-merge-to-personalize-letters-for-bulk-mailings-d7686bb1-3077-4af3-926b-8c825e9505a3. Accessed: 15-August-2017.

[130] P. Vaudry and G. Lapalme. "Adapting SimpleNLG for Bilingual English-French Realisation". In: *ENLG 2013 - Proceedings of the 14th European Workshop on Natural Language Generation, August 8-9, 2013, Sofia, Bulgaria*. Ed. by A. Gatt and H. Saggion. The Association for Computer Linguistics, 2013, pp. 183–187.

[131] M. J. Warrens. "Bounds of Resemblance Measures for Binary (Presence/Absence) Variables". In: *Journal of Classification* 25.2 (2008), pp. 195–208.

[132] A. Werner. *Introductory sketch of the Bantu languages*. School of oriental studies, London institution. K. Paul, Trench, Trubner & Company, Limited, 1919.

[133] K. Winkler, T. Kuhn, and M. Volk. "Evaluating the fully automatic multi-language translation of the Swiss avalanche bulletin". In: *Proceedings of the 4th International Workshop, CNL 2014, Galway, Ireland, August 20-22, 2014*. 2014, pp. 44–54.

[134] S. Wintner. "Formal Language Theory". In: *The Handbook of Computational Linguistics and Natural Language Processing*. Ed. by A. Clark, C. Fox, and S. Lappin. Wiley-Blackwell, 2010, pp. 9–42.

[135] K. Wong and M. H. Kim. "Privacy-preserving similarity coefficients for binary data". In: *Computers & Mathematics with Applications* 65.9 (2013). Advanced Information Security, pp. 1280–1290.

[136] T. Yao, D. Zhang, and Q. Wang. "MLWFA: A Multilingual Weather Forecast Text Generation System". In: *The 38th Annual Meeting of the Association for Computational Linguistics*. ACL 2000. Software Demonstration. Hong Kong: Department of Computer Science, Engineering, The Hong Kong University of Science, and Technology, 2000.

[137] H. Zhang, H. Wu, J. Gao, Y. Zhao, and Z. Lv. "Meteorological bulletin automatic generation based on spatio-temporal reasoning". In: *2011 International Conference on Machine Learning and Cybernetics*. Vol. 4. July 2011, pp. 1927–1931.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. The template used for the thesis was updated from a template that was created and released under the permissive MIT (X11) license. The original template can be found online at github.com/suchow/. The updated template can be found at github.com/AdeebNqo/