Developing a Morphological analyser for the Bemba Langauge

Martin Namukombo

4th Year Project Report
Bsc Computer Science(Hons)
School of Informatics
University of Edinburgh

2024

Abstract

This project presents the development of a finite-state morphological analyser for the Bemba language, a Bantu language spoken primarily in Zambia. Bemba is a morphologically rich language with a complex system of noun classes, verbal morphology, and derivational processes. The lack of extensive computational resources for Bemba has hindered the progress of natural language processing (NLP) applications in the language. To address this gap, this study aims to develop a morphological analyser using finite-state transducers (FSTs) and evaluate its performance on morphological analysis and spelling correction tasks.

The analyser is built using the Foma toolkit and is based on linguistic rules and patterns derived from Bemba grammar books and linguistic literature. The methodology involves data preparation, including the creation of curated word lists and a morphological tag set, followed by the implementation of morphological rules using FSTs. The analyser handles a wide range of morphological phenomena, including noun class prefixes, verbal extensions, and derivational processes.

Evaluation is conducted on a representative test set curated from Bemba grammar books, focusing on noun and verb derivation. The analyser achieves high accuracy and coverage in morphological analysis, successfully generating analyses for all words in the test set. However, some over-generation is observed due to the lack of fine-grained constraints on subject-verb-object relations. The analyser also demonstrates effectiveness in spelling correction, recovering the correct spelling of words in the presence of various error types.

This work contributes to the development of computational resources for Bemba and lays the foundation for further NLP research and applications in the language. The morphological analyser can be integrated into various NLP tasks, such as machine translation, information retrieval, and text generation, to support the processing and understanding of Bemba text. For example, extracting lemmas from the morphological analysis can improve the effectiveness of information retrieval systems by reducing inflectional variations. Similarly, the analyser can be used for data augmentation in machine translation, generating additional training data by inflecting words based on the identified morphological patterns. This is particularly valuable for low-resource languages like Bemba, where parallel corpora are scarce.

Future work can focus on refining the analyser, incorporating more comprehensive verb classification, and developing ranking mechanisms for spelling correction. Collaboration with linguists and the Bemba-speaking community is crucial for further improving and validating the analyser. With continued efforts, this study paves the way for building more robust language technologies for Bemba and other under-resourced languages.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Martin Namukombo*)

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Adam Lopez, for his guidance and support throughout this project. His expertise in natural language processing has been invaluable in shaping the direction and outcome of this research.

I am also grateful to the Bemba language researchers, whose many publications provided insights during the development and evaluation of the morphological analyser. Their contributions have been essential in ensuring the linguistic accuracy and relevance of this work.

Finally, I would like to thank my family and friends for their unwavering support and encouragement throughout my academic journey. Their belief in me has been a constant source of motivation and inspiration.

Table of Contents

Chapter 1

Introduction

Bemba is a Bantu language spoken primarily in Zambia and neighboring countries, with an estimated 3.6 million speakers (Simons and Fennig, 2018). As a member of the Bantu language family, Bemba exhibits rich morphological complexity, characterized by an extensive system of noun classes, complex verbal morphology, and productive derivational processes (Kula, 2002). Despite the growing body of linguistic research on Bemba, computational resources for the language remain limited. To date, no comprehensive morphological analyser has been developed specifically for Bemba, hindering the progress of natural language processing (NLP) applications in the language.

This study aims to address this gap by developing a morphological analyser for Bemba using finite-state transducers (FSTs), a well-established formalism for modeling morphological processes (Beesley and Karttunen, 2003). The analyser will be built using the Foma toolkit (Hulden, 2009) and will be based on linguistic rules and patterns derived from Bemba grammar books and linguistic literature. The analyser will cover a wide range of morphological phenomena in Bemba, including noun class prefixes, verbal extensions, and derivational processes.

The development of the Bemba morphological analyser will involve several key stages. Firstly, data preparation will be carried out, which includes the creation of curated word lists based on available Bemba corpora and linguistic resources. A morphological tag set will also be developed to represent the grammatical categories and features of Bemba words. Next, the morphological rules will be implemented using FSTs, leveraging the expressive power of the formalism to handle complex morphological patterns and dependencies.

Evaluation of the analyser will be conducted on a representative test set curated from Bemba grammar books, focusing on noun and verb derivation. The analyser's performance will be assessed in terms of accuracy, coverage, and its ability to handle exceptions and irregularities. Additionally, the analyser will be evaluated on a spelling correction task, demonstrating its potential for handling noisy and erroneous input.

The contributions of this study are twofold. Firstly, it presents the development of a comprehensive morphological analyser for Bemba, a resource that has been lacking in the language's computational landscape. The analyser will serve as a valuable tool

for various NLP tasks, such as machine translation, information retrieval, and text generation, enabling the processing and understanding of Bemba text. For instance, extracting lemmas from the morphological analysis can enhance the performance of information retrieval systems by reducing inflectional variations and improving matching accuracy. Similarly, the analyser can be utilized for data augmentation in machine translation tasks, generating additional training data by inflecting words based on the identified morphological patterns. This is particularly beneficial for low-resource languages like Bemba, where parallel corpora are scarce.

Secondly, this study contributes to the broader field of computational morphology for Bantu languages, demonstrating the effectiveness of FSTs in modeling the complex morphological systems of these languages. The techniques and approaches employed in developing the Bemba morphological analyser can serve as a reference for researchers working on similar tasks for other Bantu languages.

The remainder of this report is structured as follows. Chapter 2 provides background information on morphological analysis, Bemba language, and finite-state transducers. Chapter 3 reviews related studies in Bemba linguistics and computational resources for Bantu languages. Chapter 4 describes the methodology employed in this study, including data preparation, morphological rule implementation, and evaluation. Chapter 5 presents the results of the morphological analysis and spelling correction evaluations, along with selected analysis outputs and inflection generation examples. Chapter 6 discusses the findings, limitations, and future directions. Finally, Chapter 7 concludes the report, summarizing the contributions and implications of this study.

1.1 Morphology

Morphology is the study of the internal structure of words and the rules governing the formation of words in a language (**?**). It deals with the identification and analysis of morphemes, the smallest meaningful units in a language, and how they combine to form words. Morphological analysis is a crucial component of natural language processing (NLP) systems, as it provides valuable information about the grammatical properties and relationships of words, such as their part of speech, tense, number, and agreement features (Beesley and Karttunen, 2003).

In the context of Bantu languages, including Bemba, morphology plays a particularly important role due to the rich morphological complexity exhibited by these languages (Nurse and Philippson, 2003). Bantu languages have an extensive system of noun classes, each marked by specific prefixes that govern agreement patterns throughout the sentence. Verbs in Bantu languages also display a complex morphological structure, with slots for subject and object markers, tense and aspect markers, and derivational suffixes.

Morphological analysis in Bantu languages typically involves two main tasks: segmentation and parsing. Segmentation is the process of dividing a word into its constituent morphemes, while parsing involves identifying the grammatical properties and relationships of these morphemes. Finite-state transducers (FSTs) have emerged as a powerful formalism for modeling and processing morphological phenomena in computational

linguistics (Beesley and Karttunen, 2003). FSTs provide a compact and efficient representation of morphological rules, allowing for the analysis and generation of word forms based on their underlying morphological structure.

1.2 Bemba Language

Bemba is a Bantu language spoken primarily in Zambia and neighboring countries, with an estimated 3.6 million speakers (Simons and Fennig, 2018). It belongs to the Bantu language family, which includes over 500 languages spoken across sub-Saharan Africa (Nurse and Philippson, 2003). Bemba exhibits typical Bantu language features, such as a complex system of noun classes, agglutinative morphology, and a subject-verb-object (SVO) word order.

Bemba has a rich morphological system, with nouns divided into 9 noun classes, each marked by a specific prefix. These noun classes govern agreement patterns throughout the sentence, with verbs, adjectives, and other modifiers agreeing with the noun class of their respective targets. Verbs in Bemba also have a complex morphological structure, with slots for subject and object markers, tense and aspect markers, and derivational suffixes.

The study of Bemba linguistics has a relatively long history, with several key works contributing to our understanding of the language's structure and morphology. One of the earliest comprehensive descriptions of Bemba grammar is Schoeffer's "A Grammar of the Bemba Language", published in 1907 (Schoeffer, 1907). This work provides a detailed account of Bemba's noun class system, verbal morphology, and syntactic structures, serving as a foundation for subsequent research on the language.

Despite the growing body of linguistic research on Bemba, computational resources for the language remain limited. To date, no comprehensive morphological analyser has been developed specifically for Bemba, hindering the progress of NLP applications in the language. This study aims to address this gap by developing a morphological analyser for Bemba using finite-state transducers (FSTs), leveraging insights from linguistic research and adapting techniques used for other Bantu languages.

1.3 Finite State Transducers

Finite-state transducers (FSTs) have emerged as a powerful and widely used formalism for modeling and processing morphological phenomena in computational linguistics (Beesley and Karttunen, 2003). FSTs provide a compact and efficient representation of morphological rules and transformations, making them well-suited for tasks such as morphological analysis and generation.

An FST is a mathematical model that represents a mapping between two sets of strings: an input set and an output set. It consists of a finite set of states and a set of transitions between those states, with each transition labeled by an input symbol and an output symbol. FSTs can be used to model both concatenative and non-concatenative morphological processes, making them highly expressive and flexible.

One of the key challenges in morphological analysis is dealing with dependencies between morphemes and enforcing context-dependent rules. FSTs, being paths through a network, lack an inherent mechanism to enforce such constraints. However, computational morphology packages like Foma (Hulden, 2009) provide a feature called ”flag diacritics” that allows for the insertion of special symbols to control the application of rules. Flag diacritics set, unset, or check flags during the traversal of the FST, enabling the modeling of complex morphological phenomena beyond simple concatenation.

In the context of Bemba morphological analysis, FSTs can be used to model the mapping between surface forms (i.e., the actual words in the language) and their underlying morphological representations. This mapping defines a unified framework for both analysis and generation tasks. The same FST can be used to analyse a word form into its constituent morphemes and to generate word forms from a given set of morphemes.

The development of the Bemba morphological analyser in this study leverages the expressive power of FSTs and the capabilities of the Foma toolkit (Hulden, 2009). By encoding morphological rules and patterns derived from Bemba grammar books and linguistic literature, the analyser aims to provide a comprehensive and accurate model of Bemba morphology. The use of flag diacritics allows for the handling of dependencies and context-sensitive rules, ensuring the generation of well-formed and grammatically correct word forms.

Chapter 2

Background

2.1 Morphological Analysis

2.1.1 Overview

Morphological analysis is a crucial component of natural language processing (NLP) systems, enabling the identification and interpretation of the internal structure of words (Beesley and Karttunen, 2003). By breaking down words into their constituent morphemes – the smallest meaningful units in a language – morphological analysers provide valuable information about a word's grammatical properties, such as its part of speech, tense, number, and agreement features (Roark and Sproat, 2007).

The importance of morphological analysis is particularly pronounced in morphologically rich languages, such as the Bantu language family, which includes Bemba. In these languages, a single word often encodes a wealth of grammatical information through the use of prefixes, suffixes, and other morphological processes (Nurse and Philippson, 2003). Consequently, effective morphological analysis is essential for a wide range of NLP tasks in these languages, including machine translation, information retrieval, and parsing (Muhirwe, 2007; Bender, 2022).

For instance, in machine translation, morphological analysis enables the accurate identification and mapping of morphological features between the source and target languages. By decomposing words into their constituent morphemes and capturing their grammatical properties, morphological analysers facilitate the generation of grammatically correct translations and improve the overall quality of the translated text.

Similarly, in information retrieval tasks, morphological analysis plays a crucial role in enhancing the effectiveness of search and matching algorithms. By extracting lemmas (base or dictionary forms) from inflected words, morphological analysers reduce the impact of inflectional variations and improve the recall of relevant documents. This is particularly important in morphologically rich languages like Bemba, where words can have multiple inflected forms that convey the same core meaning.

Moreover, morphological analysis is a fundamental step in various NLP pipelines, serving as a prerequisite for tasks such as part-of-speech tagging, parsing, and named entity

recognition. By providing accurate morphological information, analysers enable downstream components to make informed decisions and improve the overall performance of NLP systems.

Morphological analysis typically involves two main tasks: *analysis* and *generation*. Analysis typically encompasses some form of segmentation or parsing, while generation involves creating word forms from root forms or other morphological components (Hammarström and Borin, 2011).

2.1.1.1 Analysis

Analysis tasks include segmentation and parsing. Segmentation is the process of dividing a word into its constituent morphemes, while parsing involves identifying the grammatical properties and relationships of these morphemes.

For example, given the Bemba word "*taulachisenda*" (meaning "you have not carried it(yet)"), a morphological analyser would perform the following analysis tasks:

Table 2.1: Example of morphological analysis

Task	Input	Output	
Segmentation	taulachisenda	ta-u-la-chi-send-a	
Parsing	taulachisenda	ta-	(negation marker),
		-u-	(subject concord, 2nd person singular),
		-la-	(aspect marker, not yet completed),
		-chi-	(object concord, 1st person singular, class 4),
		-send-	(verb root, "carry")
		-a	(terminative)

To illustrate the process, each morpheme segment within a word is assigned a set of predefined symbols for functional identification. As a practical example, consider the analysis output for '*taulachisenda*' from the analyser developed in this study.

+Negate+SbjC+Sg2+SCls1++T4+ObjC+Sg1+SCls4+**send**+V+VFA

ta:	+Negate:	- Negation marker
u:	+SbjC+Sg2+SCls1:	- Subject Concord, 2nd person singular, Class 1
la:	+T4:	- Tense/Aspect marker (not yet completed)
chi:	+ObjC+Sg1+SCls4:	- Object Concord, 1st person singular, Class 4
send:	+V:	- Verb root (send/carry)
a:	+VFA:	- Verb Final Affix

In addition to parsing, the output isolates the dictionary form/lemma for each analysed word. This is crucial for linking inflected forms back to their root.

The extracted lemmas can be utilized in various NLP tasks to improve performance and efficiency. For example, in information retrieval, using lemmas instead of inflected forms can significantly enhance the recall of relevant documents by reducing the impact of morphological variations. By indexing and searching based on lemmas, the retrieval

system can match documents containing different inflected forms of the same word, thereby improving the coverage and effectiveness of the search results.

Similarly, in tasks such as text classification or clustering, using lemmas as features instead of inflected forms can help reduce the dimensionality of the feature space and capture the core semantic content of words. This can lead to more accurate and generalized models, as the lemmas abstract away the morphological variations and focus on the underlying meaning of the words.

2.1.1.2 Generation

Generation involves creating word forms from root forms or other morphological components. This task is particularly useful for spelling correction and text generation applications. The goal is to output all possible word forms based on a given root and a set of morphological rules, without relying on any sentential context. The wordform itself provides all the necessary information for generation while the analyser incorporates linguistic knowledge about morphology, including affixation rules, morphophonological alternations, and constraints on morpheme co-occurrence.

For example, using the Bemba verb root ”pyang” (meaning ”sweep”), the analyser can generate the infinitive form ”kupyanga” by prefixing ”ku-” and suffixing ”-a”. Other possible generations based on this root are shown in Table 2.2.

rr

Table 2.2: Example of morphological generation

Word Form	Morphological Composition	Meaning
kupyanga	ku-pyang-a	to sweep (infinitive)
bapyanga	ba-pyang-a	they sweep
chipyanga	chi-pyang-a	it sweeps
pyangila	pyang-il-a	sweep for (applicative)
pyangulula	pyang-ulul-a	reversive
pyangilila	pyang-ilil-a	completive
pyangana	pyang-an-a	reciprocal
pyangisha	pyang-ish-a	intensive
pyenge	py-e-ng-e	modified

By accurately performing these analysis and generation tasks, morphological analysers enable downstream NLP applications to exploit the rich grammatical information encoded in morphologically complex words and generate appropriate word forms for various purposes.

The generation capability of morphological analysers is particularly valuable in the context of low-resource languages like Bemba. In such languages, the availability of large-scale annotated corpora is often limited, which poses challenges for data-driven approaches to NLP tasks. However, by leveraging the linguistic knowledge encoded in the morphological analyser, it becomes possible to generate synthetic data for training and augmenting NLP models.

For instance, in machine translation, the morphological analyser can be used to generate additional parallel data by inflecting words in the source language and aligning them with their corresponding translations in the target language. This synthetic data augmentation approach can help alleviate the scarcity of parallel corpora and improve the performance of machine translation systems for low-resource languages.

Similarly, in text generation tasks such as language modeling or text summarization, the morphological analyser can be employed to generate grammatically correct and semantically coherent word forms based on the provided context. By incorporating the morphological rules and constraints encoded in the analyser, the generated text can exhibit better fluency and adhere to the linguistic properties of the language.

2.1.2 Approaches to Computational Morphology

Various computational approaches have been proposed for morphological analysis, ranging from rule-based methods to data-driven and hybrid techniques. Rule-based approaches rely on hand-crafted linguistic rules and lexicons to analyse words, while data-driven methods learn morphological patterns from annotated corpora without explicit linguistic annotation.(Hammarström and Borin, 2011)w.

One popular rule-based formalism for morphological analysis is finite-state transducers (FSTs) (Beesley and Karttunen, 2003). FSTs provide a compact and efficient representation of morphological rules, allowing for the modelling of complex morphological processes such as affixation, reduplication, and morphophonological alternations. Many successful morphological analysers for morphologically rich languages have been developed using FSTs, including analysers for Swahili (Pretorius and Bosch, 2005), Zulu (Pretorius and Bosch, 2005), and Turkish (Eryıit and Adalı, 2008). An advantage of a rule-based approach is the ability to handle non-concatenative morphological processes, such as infixation, reduplication, and stem alternations. An example of this in Bemba, is verb stem alternation when converting certain simple verbs, to their modified forms e.g

*-sok**o**shy**a*** (quarrel)	*-sok**we**shy**e***
-o- -a	*-we -e*

Tense modification (tense 7, 9, 10, and 11) is applied by altering the verb root (*sokosh*).

Data-driven approaches on the other hand, have gained prominence in recent years with the availability of large corpora and advances in machine learning techniques. These approaches often employ statistical or machine learning techniques to identify morpheme boundaries and learn morphological regularities (Hammarström and Borin, 2011; Creutz et al., 2007).

However, as pointed out by Hammarström and Borin (2011), the vast majority of unsupervised morphology learning research focuses on purely concatenative morphology, assuming that words are formed by simply concatenating morphemes together. This assumption may not hold for languages with non-concatenative morphological processes, limiting the effectiveness of these approaches in capturing the full range of morphological phenomena.

The survey by Hammarström and Borin (2011) concludes with a suggestion for more interaction between unsupervised learning approaches and linguistically informed

research on morphology. The authors emphasise the need to incorporate linguistic knowledge and typological insights into unsupervised morphology learning to develop more effective and generalisable models.

Similarly, Can and Manandhar (2014) highlight the challenges and limitations of current unsupervised morphology learning methods, noting that they often fail to generalise well beyond the specific languages or phenomena they were designed for. The authors suggest that future research should address non-concatenative morphology, stem alternation, morpheme clustering, and morphological transformation rule induction to improve the effectiveness and applicability of unsupervised approaches.

2.1.3 Designing for Accuracy

It is important to note that while unsupervised learning approaches have their merits in tasks related to morphological segmentation and reducing data sparsity in downstream NLP application, they may not be suitable for the full range of morphological analysis and generation required for Bemba.

Designing a morphological analyser for accuracy requires careful consideration of language specific factors and variables. In the case of Bemba, this would entail explicitely modeling morphological patterns, such as noun class agreement, verbal extensions, and derivational processes. Having a single model to capture these patterns allows for unified representation of word derivation. This opens doors to additional uses cases such as generating inflected forms given a root, as will be demonstrated.

2.1.4 Finite-State Transducers in Morphological Analysis

FSTs provide a compact and efficient representation of morphological rules and transformations, making them well-suited for tasks such as morphological analysis and generation.

An FST can be thought of as is a mathematical model that represents a mapping between two sets of strings: an input set and an output set. It consists of a finite set of states and a set of transitions between those states, with each transition labeled by an input symbol and an output symbol. FSTs can be used to model both concatenative and non-concatenative morphological processes, making them highly expressive and flexible.

Figure 2.1 illustrates a simple finite state transducer (FST) that pluralizes the nouns ”fox” and ”dog”. Assuming the task is to define a rudimentary machine for pluralization, this FST demonstrates the core concept: it accepts ”fox” and ”dog” as inputs and appends the appropriate suffixes to produce ”foxes” and ”dogs”.

Here, ε represents an empty string input that triggers the output of ”s” for ”dog” to form ”dogs” and ”es” for ”fox” to form ”foxes”. The ε transitions allow the FST to append these suffixes without reading any additional input characters. It should be noted that The example intentionally permits the direct transition from state q_0 to q_2, producing the string ’es’ in addition to the erroneous ’foxdox’. This simplification is intentional and is further discussed below.

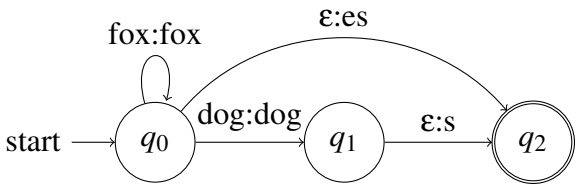

Figure 2.1: Example FST for English plural noun formation

In the context of morphological analysis, FSTs can be used to model the mapping between surface forms (i.e., the actual words in a language) and their underlying morphological representations. This mapping defines a unified framework for both analysis and generation tasks. The same FST can be used to analyse a word form into its constituent morphemes and to generate word forms from a given set of morphemes.

Several computational morphology packages and tools, such as foma (Hulden, 2009), OpenFST (Allauzen et al., 2007), and HFST (Lindén et al., 2011), provide implementations of finite-state transducers specifically designed for morphological processing. These packages offer high-level interfaces and optimized algorithms for constructing, manipulating, and applying FSTs to morphological analysis and generation tasks.

One of the key challenges in morphological analysis is dealing with dependencies between morphemes and enforcing context-dependent rules. FSTs, being paths through a network, lack an inherent mechanism to enforce such constraints. However, computational morphology packages like foma (Hulden, 2009) provide a feature called ”flag diacritics” that allows for the insertion of special symbols to control the application of rules. Flag diacritics set, unset, or check flags during the traversal of the FST, enabling the modelling of complex morphological phenomena beyond simple concatenation. Figure 2.2 illustrates the use of flag diacritics in the English plural noun formation example.

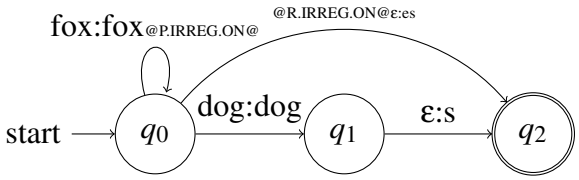

Figure 2.2: FST example with flag diacritics for English plural noun formation

The flag @P.IRREG.ON@ places a positive set on the feature IRREG, while @P.IRREG.ON@ specifies a requirement of this flag(IRREG) for a path to be considered. Any transition directly from q_0 to q_2 will not have this flag and hence, will be rejected.

This is not directly defined in the structure of the FST, but the foma provides an interface to specify these rules while managing any dependencies during compilation.

2.1.5 Foma

The morphological analyser presented in this study is implemented using Foma (Hulden, 2009), a finite-state toolkit designed for morphological processing. Foma provides a

high-level programming language and a set of utilities for constructing, manipulating, and applying finite-state transducers (FSTs) to morphological analysis and generation tasks.

Foma represents lexicons as a set of continuation classes. A continuation class specifies the possible next states or actions to be taken based on the current input symbol. The basic structure of a lexicon entry in Foma's XFST format is as follows:

```
LEXICON Root
output1   ContinuationClass1;
output1   ContinuationClass2;
...
```

In this format, each line represents a lexicon entry, with the output on the left-hand side and the corresponding continuation class on the right-hand side, separated by spaces or tabs. The semicolon (;) marks the end of each entry.

This defines a path starting from the Root, producing 'output1', then continuing to 'ContinuationClass1'. This can further be augmented to accept inputs. For example:

```
LEXICON Verbs
input1:output1   ContinuationClass1;
input2:output2   ContinuationClass2;
...
```

In this case, "input1" and "input2" represent the input symbols (e.g., verb stems), while "output1" and "output2" represent the corresponding output symbols (e.g., morphological tags or surface forms).

The special symbol "#" is used to represent the end of a word in Foma. It is typically used in the final continuation class to indicate that the word formation process is complete. For example:

```
LEXICON VerbFinal
+Verb+Present:a   #;
+Verb+Past:ile   #;
...
```

In this example, the "VerbFinal" continuation class specifies the possible verb endings, such as the present tense marker "a" and the past tense marker "ile". The "#" symbol indicates that the word formation is complete after appending these endings.

One of the key features of Foma is its support for flag diacritics, which are special symbols used to enforce long-distance dependencies and control the application of rules based on certain conditions. Flag diacritics allow for the modelling of complex morphological phenomena that go beyond simple concatenation, such as agreement, reduplication, and non-concatenative processes.

Flag diacritics in Foma are represented using the "@" symbol followed by a flag type and a flag value. The most common flag types are:

`"@P.Flag.Value@": Positive setting of a flag`

```
"@N.Flag.Value@": Negative setting of a flag
"@R.Flag.Value@": Require a flag to have a specific value
"@D.Flag.Value@": Disallow a flag to have a specific value
"@C.Flag@": Clear a flag
```

As was shown in 2.2, these flags allow us to impose restrictions in which paths get to reach a word boundary.

2.1.6 Evaluation

2.1.6.1 Evaluation of Morphological analysers

Evaluating the performance of morphological analysers is crucial for assessing their effectiveness and guiding further improvements. The choice of evaluation strategies and metrics largely depends on the approach used to develop the analyser, such as learned approaches (supervised and unsupervised) or rule-based methods.

Learned approaches to morphological analysis, including both supervised and unsupervised methods, rely on training data to learn the mapping between word forms and their morphological analyses. Evaluation will typically involve measuring the accuracy of the predicted analyses against a held-out test set.

The accuracy of a learned morphological analyser can be calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correctly predicted analyses}}{\text{Total number of test instances}} \tag{2.1}$$

In addition to accuracy, precision, recall, and F1-score are commonly used metrics for evaluating learned morphological analysers (Cotterell et al., 2016). These metrics provide a more nuanced view of the analyser's performance, taking into account the trade-off between correctly identified morphemes and false positives/negatives.

Precision measures the proportion of correctly predicted morphemes among all predicted morphemes, while recall measures the proportion of correctly predicted morphemes among all true morphemes in the test set. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the analyser's performance.

$$\text{Precision} = \frac{\text{Number of correctly predicted morphemes}}{\text{Total number of predicted morphemes}} \tag{2.2}$$

$$\text{Recall} = \frac{\text{Number of correctly predicted morphemes}}{\text{Total number of true morphemes}} \tag{2.3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

Given rule-based approaches to morphological analysis rely on hand-crafted linguistic rules and lexicons to analyse and generate word forms, evaluation is not as straight

forward. As this is not a learning problem, it relies mostly on linguistic expertise or an assessment of the analysers' capacity in handling various morphological phenomena.

A common evaluation strategy is to measure the coverage on a representative corpus of the language (Karttunen, 2003) and calculating the percentage of words that receive a valid morphological analysis. High coverage indicates that the analyser can handle a wide range of word forms and morphological patterns.

In the context of this study, the focus will be on the models ability to handle the language's specific morphological patterns when measured against a representative sample of words from existing literature. This approach could also be extended to include precision in cases where the analyser over-generates words, and recall if it under-generates possible word forms.

2.1.6.2 Error Analysis

Regardless of the approach used, error analysis and qualitative evaluation are essential for understanding the strengths and limitations of morphological analysers. This is particularly useful in the case of a rule based approach as it may point to novel linguistic phenomena, or erroneous application of rules, providing insights into challenges and areas for improvement (Stymne, 2008).

In summary, the evaluation of morphological analysers should be multidimensional, taking into account the specific approach used and the goals of the analysis. Supervised learning approaches can be evaluated using accuracy, precision, recall, and F1-score metrics, while unsupervised approaches may rely on linguistic plausibility, downstream task performance, and quantitative measures like boundary prediction accuracy. Rule-based approaches can be evaluated in terms of coverage, ambiguity handling, and precision/recall. Qualitative evaluation, error analysis, and comparative evaluation provide additional insights into the performance and limitations of morphological analysers.

2.2 Bemba Morphology

2.2.1 Overview

Bemba is a Bantu language spoken primarily in Zambia and neighbouring countries, with an estimated 3.6 million speakers (Simons and Fennig, 2018). As a member of the Bantu language family, it exhibits rich morphological complexity, characterised by an extensive system of noun classes, complex verbal morphology, and productive derivational processes (Kula, 2002). It uses the Latin alphabet and has a relatively simple consonant inventory, with a few notable features, such as the presence of prenasalised consonants and the absence of voicing contrast for stops (Hamann and Kula, 2015). Bemba's syllable structure is predominantly open, with consonant-vowel (CV) syllables being the most common.

To clarify what counts as a syllable in Bemba, let's consider the examples mentioned above. The verb stem 'konka' (follow) has two syllables: 'ko' and 'nka', with the root

vowel being 'o'. This follows the CVCCV structure, where the syllable boundary is between the vowel and the following consonant.

Similarly, the verb root 'konk' also has two syllables: 'ko' and 'nk', following the CVCC structure. The syllable boundary is placed between the root vowel 'o' and the following consonant 'k', resulting in the syllables 'ko' and 'nk'.

In Bemba, syllables are typically of the form CV (consonant-vowel) or CVC (consonant-vowel-consonant). The syllable boundary is usually placed after the vowel, separating it from the following consonant. This syllabification pattern is important when analyzing verb morphology, particularly in the context of deriving modified verb forms based on the number of syllables and the root vowel.

Bemba follows the typical grammatical structure of Bantu languages, with a subject-verb-object (SVO) word order and a complex system of agreement between nouns and their modifiers, as well as between subjects and verbs (Nurse and Philippson, 2003).

2.2.2 Word Classes

Nouns
Nouns in Bemba are divided into 9 noun classes, each marked by a specific prefix. These noun classes govern agreement patterns throughout the sentence, with verbs, adjectives, and other modifiers agreeing with the noun class of their respective targets. Table 2.3 shows these classes and their corresponding plural prefixes.

Class	Plural Prefix	Example
1	ba	ba-ntu (people)
2	mi	mi-mana (rivers)
3	n	n-pili (becomes mpili after vowel fusion)
4	fi	fi-ntu (things)
5	ma	ma-ni (eggs)
6	tu	tu-ana (becomes twana after vowel fusion)
7	no plural, singular bu	bu-lungu (beads)
8	ku	ku-lya (eating)
9	ku, mu, pa	mostly locative nouns kumyesu (our house)

Table 2.3: Noun Prefixes in Bemba

For example, the noun "puna" (chair) belongs to class 4, and its corresponding plural subject concord is "fi-", as in:

"i**fi**puna **fi**li pano" (the chairs are here).

Likewise, the class 1 noun, "ntu"(person) with conord "ba-":

"a**ba**ntu **ba**li pano" (the people are here).

Verbs

Verbs in Bemba consist of a root and multiple affixes, which convey information about the verb's subject, tense, aspect, mood, and object. The verbal template includes slots for subject and object markers, tense and aspect markers, and derivational suffixes. Tense indicates the time of the action (past, present, future). Aspect describes the flow of time within the action (completed, ongoing, habitual). Mood conveys the speaker's attitude towards the action (indicative, imperative, subjunctive). Table 2.4 illustrates the verbal template with an example.

Subject	**Tense**	**Object**	**Root**	**Extension**	**Final Vowel**
ba-	-ka-	-mu-	-bomb-	-el-	-a

Table 2.4: Verbal template in Bemba with the example "bakamubombela" (they will work for him/her)

Bemba verbs are marked for tense, aspect, and mood. In total, there are 18 tense, aspect and mood markers. A short selection is included in Table 2.5.

Tense Markers	**Aspect Suffixes**	**Mood Indicators**
-a-: immediate past	-aci: perfective	-a: indicative mood
-ali-: near past	-lee-: imperfective	-e: subjunctive mood
-ile: far past	-laa-: iterative	-eni: imperative mood

Table 2.5: Bemba Verbal Markers

Bemba also features an extensive system of derivational morphology, with suffixes that can change the meaning or valency of verb roots. These suffixes include:

Suffix	**Meaning**	**Example**
-il-/-el-	applicative	-bomb-el-a (work for/at)
-iw-/-ew-	passive	-lub-iw-a (be lost)
-an-	reciprocal	-mon-an-a (see each other)
-ish-/-esh-	causative	-ly-esh-a (cause to eat)
-ul-/-ol-	reversive	-kak-ul-a (untie)
-ulul-/-olol-	repetitive	-kak-ulul-a (untie repeatedly)
-ik-/-ek-	stative (neuter)	-lub-ik-a (be lost)

Table 2.6: Derivational suffixes

These derivational suffixes can be applied to various verb roots to create new verb forms with different meanings or valency and allow for the expression of a wide range of semantic and syntactic relations. For example:

- -fik- (arrive) + -il- (applicative) = -fik-il-a (arrive for/at)
- -send- (carry) + -esh- (causative) = -send-esh-a (cause to carry)
- -kos- (pull) + -an- (reciprocal) = -kos-an-a (pull each other)
- -lek- (leave) + -ik- (stative) = -lek-ek-a (be left)

2.2.3 Word Derivation

Noun Derivation Bemba employs several strategies for deriving nouns from other word classes, particularly from verbs and other nouns. **Verb-to-Noun Derivation:** Nouns can be derived from verbs by adding noun class prefixes to the verb root. The most common noun classes used for this purpose are:

- Class 1 (mu-): agent nouns
  - -bomb- (work) + mu- = mu-bomfi (worker)
- Class 4 (ci-): instrument nouns
  - -lemb- (write) + ci- = ci-lembo (pen)

Noun-to-Noun Derivation: Nouns can also be derived from other nouns by adding noun class prefixes or suffixes. For example:

- Class 6 (ka-): diminutive nouns
  - mu-ntu (person) + ka- = ka-ntu (small person)
- Class 6 (tu-): plural diminutive nouns
  - ci-puna(chair) + tu- = tu-puna (small chair)
- -ina (mother) + possessive concord = -ina-ine (my mother), -ina-ko (your mother)

Verb Derivation

As mentioned earlier, Bemba has an extensive system of verbal derivation, using suffixes to modify the meaning or valency of verb roots.

Verb-to-Verb Derivation: Derivational suffixes can be added to verb roots to create new verb forms with different semantic or syntactic properties. The most common derivational suffixes are:

- -il-/-el-: applicative
  - -lub- (lose) + -il- = -lub-il-a (lose for/at)
- -iw-/-ew-: passive
  - -send- (carry) + -iw- = -send-ew-a (be carried)
- -an-: reciprocal
  - -mon- (see) + -an- = -mon-an-a (see each other)
- -ish-/-esh-: causative
  - -ly- (eat) + -esh- = -ly-esh-a (cause to eat)

Multiple derivational suffixes can be combined to create more complex verb forms, such as:

- -fum- (come from) + -il- (applicative) + -an- (reciprocal) = -fum-il-an-a (come from each other)

- -lands- (speak) + -esh- (causative) + -iw- (passive) = -lands-esh-iw-a (be caused to speak)

These derivational processes contribute to the richness and productivity of Bemba's morphological system, allowing for the creation of new words to express a wide range of meanings.

2.3 Related Studies

2.3.1 Bemba Linguistics

The study of Bemba linguistics has a relatively long history, with several key works contributing to our understanding of the language's structure and morphology. One of the earliest comprehensive descriptions of Bemba grammar is Schoeffer's "A Grammar of the Bemba Language", published in 1907 (Schoeffer, 1907). This work provides a detailed account of Bemba's noun class system, verbal morphology, and syntactic structures, serving as a foundation for subsequent research on the language.

In more recent years, several studies have focused on specific aspects of Bemba morphology and its implications for linguistic theory. Kula (2002) provides an in-depth analysis of Bemba's verbal derivation system, exploring the semantic and syntactic properties of extensions such as the causative, applicative, and passive. This work highlights the complex interactions between derivational suffixes and their effects on argument structure and thematic roles.

Other studies have examined Bemba's noun class system and its role in agreement and concord phenomena. Spitulnik's (1987) investigation of the semantic attributes of Bemba noun classes reveals the intricate conceptual networks that underlie the language's nominal classification system. By exploring the semantic motivations behind noun class assignment, this work sheds light on the cognitive and cultural factors that shape Bemba's grammatical structure.

Comparative studies have also been conducted, situating Bemba within the broader context of Bantu languages. Marten's (2006) analysis of Bantu verbal morphology includes data from Bemba, highlighting the language's similarities and differences with other Bantu languages in terms of tense-aspect marking and derivational processes. Such comparative work is crucial for understanding the typological characteristics of Bemba and its place within the Bantu language family.

2.3.2 Computational Resources

Despite the growing body of linguistic research on Bemba, computational resources for the language remain limited. To date, no comprehensive morphological analyser has been developed specifically for Bemba, hindering the progress of NLP applications in the language.

However, several computational tools and resources have been created for other Bantu languages, which can serve as valuable references and starting points for the devel-

opment of a Bemba morphological analyser. Notable examples include the Zulu morphological analyser developed by Pretorius and Bosch (**?**), which employs finite-state transducers to model Zulu's complex verbal morphology. This analyser has been successfully applied to various NLP tasks, such as part-of-speech tagging and machine translation (Pretorius and Bosch, 2009).

Similarly, Elwell (2008) presents a morphological analyser for Swahili, another Bantu language, using the Xerox finite-state tools. This analyser covers a wide range of Swahili's morphological phenomena, including noun class agreement, verbal inflection, and derivation. The techniques and strategies employed in the development of these analysers can inform the design and implementation of a Bemba morphological analyser, taking into account the language-specific characteristics and challenges.

In addition to morphological analysers, parallel corpora and lexical resources have been developed for some Bantu languages. The Helsinki Corpus of Swahili (Hurskainen, 2004), for example, provides a large collection of annotated Swahili texts, which has been used to train and evaluate NLP tools such as part-of-speech taggers and parsers. While no comparable corpus exists for Bemba, the methodologies used in the creation and annotation of such resources can guide the development of similar tools for Bemba.

Recent initiatives, such as the FLORES-200 (Team et al., 2022) and BigC (Sikasote et al., 2023), have made significant strides in creating parallel corpora for Bemba. The FLORES-200 dataset provides a collection of translated sentences in over 200 languages, enabling the development of machine translation systems and cross-lingual NLP tools. The BigC project similarly provides a large collection of parallel sentences, Bemba included.

The lack of extensive computational resources for Bemba underscores the importance of the current study, which aims to address this gap by developing a morphological analyser for the language. By leveraging insights from linguistic research and adapting techniques used for other Bantu languages, this work seeks to lay the foundation for further NLP research and applications in Bemba.

Chapter 3

Methodology

3.1 Data Preparation

3.1.1 Data Sources and Preprocessing

The development of the Bemba morphological analyser relied primarily on Schoeffer's "A Grammar of the Bemba Language" (Schoeffer, 1907) as a key reference for understanding the language's grammatical rules and patterns. The Big C and Flores datasets were used to extract a list of unique Bemba words, while the Zambian Ministry of Educations' orthography book **?** provided additional insights into Bemba's writing system and linguistic structure.

The data preprocessing stage involved several key steps. First, a comprehensive list of Bemba words was compiled by applying space tokenization on the Big C and Flores datasets and extracting the unique words. The words were then lowercased and resulted in a total of 66,000 unqiue tokens. No additional filtering was applied.

Next, word stems were identified and extracted from the word list using pattern matching techniques, such as regular expressions. This was not largely a manual process and did not involve any form of automation. For example, regular expressions were used to match and extract verb stems by identifying common verbal prefixes and suffixes, such as subject and object concords, tense markers, and derivational suffixes. Similarly, noun stems were extracted by matching and removing noun class prefixes. The extracted stems were then stored in separate files based on their word class (e.g., nouns, verbs, adjectives). This was an iterative approach as there is a lot of ambiguity regarding classifying nouns by class. Schoeffer's book was extensively used as source to guide this process.

During the preprocessing stage, a process of morpheme segmentation and reversal was applied to handle cases where morphemes had undergone fusion. Fusion occurs when two or more morphemes combine and undergo phonological changes, resulting in a form that differs from the simple concatenation of the individual morphemes. For example, the word "twana" (children) is the result of the fusion of the prefix "tu-" and the stem "-ana". To handle such cases, regular expressions were used to identify and

separate the fused morphemes, and the original underlying morphemes were recorded separately.

Example entries from the preprocessed data:

- Verbs:
  - -belenga (read) Segmentation: ”-belenga” (stem)
  - fyaleenda (walk) Segmentation: ”-enda” (stem)
  - ukutula (pierce/puncture) Segmentation: ”-tula” (stem)
- Affixes:
  - ba- (plural prefix for class 1 nouns)
  - -ile (past tense suffix)
  - -ish- (causative suffix)

The preprocessed data served as the foundation for building the morphological analyser, providing the necessary lexical and grammatical information to construct the FST and implement the analysis and generation rules.

3.1.2 Curated Word Lists

The development of the Bemba morphological analyser relied on carefully curated word lists for different grammatical categories. These word lists were compiled from various sources, including grammar books, dictionaries, and linguistic literature. The purpose of these curated word lists was to ensure comprehensive coverage of Bemba's morphological patterns and to provide a reliable basis for evaluating the analyser's performance.

The curated token lists included the following grammatical categories and their respective quantities:

It is important to note that some grammatical categories, such as adjectives, had lower quantities in the curated word lists. This is because adjectives in Bemba are highly dependent on the noun class system, with each adjective having different forms corresponding to different noun classes. As a result, the coverage of adjectives in the analyser was primarily achieved through the implementation of the noun class agreement rules rather than exhaustive listing of all possible adjective forms.

3.1.3 Morphological Tag Set

To represent the morphological properties of Bemba words, a tag set was developed based on the language's grammatical categories and features. The tag set includes the following categories:

```
Multichar_Symbols

+SbjC +ObjC
```

Grammatical Category	**Quantity**
Adjectives	14
Numeric Adjectives	6
Numeric Concords	5
Adverbs (with affixes)	3
Adverbs (basic)	34
Conjunctions	16
Interjections	14
Nouns (Class 1)	25
Nouns (Class 2)	45
Nouns (Class 3)	0
Nouns (Class 4)	27
Nouns (Class 5)	122
Nouns (Class 6)	66
Nouns (Class 7)	7
Nouns (Class 8)	0
Nouns (Class 9)	0
Prepositions	13
Pronouns (Demonstrative)	84
Pronouns (Interrogative)	9
Pronouns (Personal)	4
Pronouns (Possessive)	6
Pronouns (Relative)	20
Verb Stems	266

Table 3.1: Quantity of words in curated word lists by grammatical category

```
+Sg1 +Sg2 +Sg3
+Pl1 +Pl2 +Pl3
+SCls1 +SCls2 +SCls3
+OCls1 +OCls2 +OCls3
+SCls2 +SCls3 +SCls4 +SCls5 +SCls6 +SCls7 +SCls8 +SCls9
+OCls2 +OCls3 +OCls4 +OCls5 +OCls6 +OCls7 +OCls8 +OCls9
+Negate
+Pres +Fut +Pers +Pot +OC +Lat +Pst +Cons +Inf
+Stat +VF-A +VF-E +VF-ENI
+T1 +T2 +T3 +T4 +T5 +T6 +T7 +T8 +T9 +T10 +T11 +T12 +T13 +T14 +T15 +T16 +T17 +T18
+MOD +COMPL +APPL +REV +CAUS +FREQ +REFL +INTE +RECIP
@P.Negate.ON@
@D.Negate.ON@
@C.Negate@
@P.AddedPrefix.ON@
@D.AddedPrefix.ON@
@C.AddedPrefix@
```

Table 3.2 provides a description of each tag in the morphological tag set.

Table 3.2: Morphological tag set for Bemba

Tag	Description	Tag	Description
+SbjC	Subject Concord	+Pl1, +Pl2, +Pl3	Plural (1st, 2nd, 3rd person)
+ObjC	Object Concord	+SClsX	Subject Concord (Class X)
+Sg1, +Sg2, +Sg3	Singular (1st, 2nd, 3rd person)	+OClsX	Object Concord (Class X)
+Negate	Negation	+Pres, +Fut, +Pst	Present, Future, Past Tense
+Pers, +Pot, +OC	Persistive, Potential, Object Concord	+Lat, +Cons, +Inf	Latent, Consecutive, Infinitive
+VF-A, +VF-E, +VF-ENI	Final Vowel (-a, -e, -eni)	+TX	Tense (X = 1 to 18)
+MOD, +COMPL, +APPL	Modified, Completive, Applicative	+REV, +CAUS, +FREQ	Reversive, Causative, Frequentative
+REFL, +INTE	Reflexive, Intensive	+RECIP	Reciprocal

Each tag consists of an abbreviation for the grammatical category followed by the specific feature or class. For example, "+SCls1+OCls2+T3[VERB]+FREQ" would represent a verb with a subject concord of class 1, an object concord of class 2, tense 3, and a frequentative suffix.

The morphological tag set serves as a standardized representation of the morphological properties of Bemba words. It allows for the consistent annotation and analysis of words in the morphological analyser, enabling the accurate identification and generation of word forms based on their underlying morphological structure.

The morphological tag set also includes special tags for handling flag diacritics, which are used to control the application of morphological rules and enforce long-distance dependencies. Tags like @P.Negate.ON@, @D.Negate.ON@, @C.Negate@, @P.AddedPrefix.ON@, @D.AddedPrefix.ON@, and @C.AddedPrefix@ are used in conjunction with flag diacritics to set, unset, or check specific flags during the morphological analysis process.

3.2 Building the Morphological analyser

3.2.1 Grammar Overview

The Bemba morphological analyser was implemented primarily using finite-state transducers (FSTs), a well-established formalism for modelling morphological processes (Beesley and Karttunen, 2003).

Figure 3.1 presents a high-level design of the FST, illustrating the different lexicons and their interactions.

The FST is compiled by providing a lexicon script, which encodes all the transformation rules represented as a set of continuation classes. This script is generated in the Python programming language, and defines each lexicon, including the root lexicon. The root serves as the entry point and defines continuations to the tense-rewrite lexicon (entry point for verb derivation) and the noun lexicon, as shown below:

```
LEXICON Root
TENSE-REWRITE ;
NOUN-LEXICON ;
```

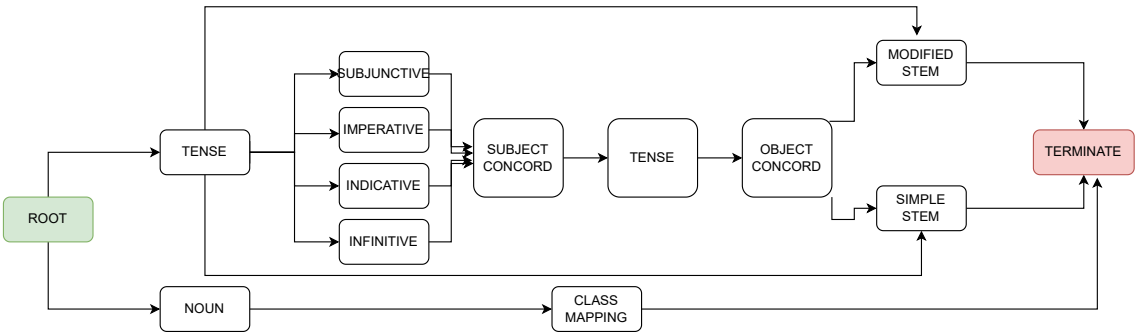

Figure 3.1: High-level design of the Bemba morphological analyser FST

```
LEXICON SUBJUNC-IN
SUBJUNC-SC-IN ;
SUBJUNC-SC-IN-FUT ;
shina  SUBJUNC-SC-IN-FUT-2 ;
ka SUBJUNC-SC-IN-FUT-2 ;

ka SUBJUNC-SC-IN ;
na SUBJUNC-SC-IN ;
```

The Root lexicon represents the starting point of the morphological analysis. It defines the continuation classes that determine the possible paths for word formation. In this example, the Root lexicon specifies two continuation classes: TENSE-REWRITE and NOUN-LEXICON.

The TENSE-REWRITE continuation class serves as the entry point for verb derivation. It handles the rewriting of tense markers and directs the analysis to the appropriate verbal inflection paths. The NOUN-LEXICON continuation class, on the other hand, leads to the analysis of noun forms, including the handling of noun class prefixes and agreement patterns.

Each path in the Root lexicon then branches off to additional derivation paths. A detailed overview of this derivation graph is shown in Figure A.1.

The derivation graph illustrates the various paths and continuations that are followed during the morphological analysis process. It shows how the analysis proceeds from the Root lexicon to different sub-lexicons based on the morphological properties of the input word.

For example, the SUBJUNC-IN lexicon handles subjunctive forms and includes several continuation classes for different subjunctive patterns. The SUBJUNC-SC-IN and SUBJUNC-SC-IN-FUT continuation classes represent different paths for subjunctive subject concords, while ‘shina‘ and ‘ka‘ denote continuation classes that handle subjunctive markers places before the subject marker.

The morphological rules and constraints are encoded within these sub-lexicons, specifying the allowed combinations of morphemes, agreement patterns, and morphophonological alternations.

3.2.2 Morphological Rule Implementation

The implementation of morphological rules in the Bemba analyser follows the finite-state transducer (FST) formalism. Each morphological rule is modeled as an individual FST, specifying the input and output symbols, as well as any necessary constraints and transformations. These individual rule FSTs are then combined using FST operations like concatenation and union to create a comprehensive model of Bemba morphology.

To illustrate the process, let's consider a simplified example of verb morphology in Bemba. We can broadly categorize verb derivation into two forms: simple (without stem alternation) and complex (with stem alternation). The simple case takes the root form of the verb in its unmodified form and appends relevant prefixes and a final 'a' as the terminative. The rewrite structure for this form would be:

```
<neg> <subj conc> <tense> <obj conc> <simple stem> a
```

For example, the stem 'konk' (follow), coupled with the following components:

- Negative prefix: ta-
- Subject concord: ba- (they)
- Tense marker: -le- (T1, present, imperfect, continuous)
- Object concord: -chi- (object, usually large object/thing)
- Terminative: -a

Generates:

```
ta-ba-le-chi-konk-a (They are not following it.)
(tabalechikonka)
```

Table 3.3 shows the FST network for this simple verb form.

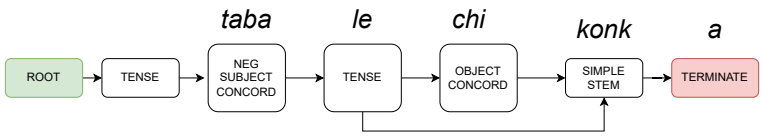

Table 3.3: FST network for the simple verb form

However, this represents only a simple verb form network. As outlined in the grammar section, the simple stem is just one of nine different verb forms (modified, applied, completive, etc.).

Complex forms necessitate rewrites to be applied directly to the verb stem, in addition to a different set of terminatives. For the stem root 'fik'(arrive), Table **??** illustrates how to derive its modified forms.

Given that the verb stem 'konk' has two syllables and the root vowel 'o', the modified form would be 'konkele'. This modified form inherently carries constraints on which tense aspects can co-occur with it and thus, the FST network would skip the tense marker. This now takes a new branch that goes on to append the modified prefix '-ele'.

Table 3.4 shows the updated FST network with the modified form.

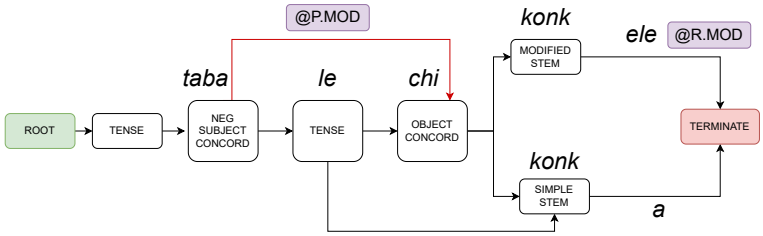

Table 3.4: Updated FST network with the modified form

An obvious inefficiency with this approach is that it does not generalise to all verb stems. It requires classifying verbs by syllable count and verb root, in addition to placing skip connections as shown above.

Foma addresses this issue by using flag diacritics to enforce the fact that a derivation goes to a modified stem, without the need for manually handcrafting skip connections. This allows for a network that prevents having an additional skip connection before the tense marker, while still enforcing the syllable count and root vowel markers to prevent block certain suffixes. This indidiated in the diagram with the placement of ‘@P.MOD‘ ‘@R.MOD‘. There is not requirement on what these flags are called and in our case ‘MOD‘ could be an arbitrary string. The `red` connection places a positive marker on our ‘MOD‘ flag. To complement this, we place a ‘require‘ flag on the modified stem output. When this connection gets to the ‘MODIFIED STEM‘ block, its allowed through because it has the required positive flag. Connection not going through the red connection will be blocked from reaching a final state as they do not possess the required flag.

So far, we have covered generating modified verb stems, and I need not dive into the other forms before illustrating some challenges with this approach. We have classified verbs based on root vowel and syllable count. While this is sufficient for the most basic verbs, further exceptions start to apply based on the syllables immediately following the root vowel. Specifically, this covers two cases:

Case 1: Verbs with two syllables and a final syllable ’-shya’. The rewrite in this case changes ’-shya’ to ’-shishye’. For example:

```
root        stem        modified
teshy       teshya      tesheshye
(hear)      (hear)      (heard, distant past of today)
```

Case 2: If the verb has more than 2 syllables and ends in: -ba, -ka, -ma, -mya, -na, -sa, -shya, -ta.

Our rewrite table changes to:

This is rewrite by considering a pair of the root vowel and final vowel, with the right handside representing their respective replacements. For example, the verb root ’longa’ (gather/pack) becomes ’longele’(gathered/packed), while ’sokoshya’ (quarrel) becomes ’sokweshye’(quarelled).

Verb-stem	Modified-stem
a -a	e -e
e -a	-e
i -a	i -e
o -a	we -e
u -u	wi -e

Table 3.5: Updated rewrite forms for verbs with specific final syllables

Our prior formulation does not consider cases where we may have to apply changes to the verb root. Considering the derivation table defined above, a possible extension to this could be preserving a modified verb diacritic but routing it to another lexicon that further splits these rewrites based on the syllable after the root vowel. That is, using the observation that verbs of 2 syllables, with root [’o’, ’e’] and final vowel ’a’, the completion stemming from this goes to a lexicon that appends ’ene’ while verbs with root [’a’, ’i’, ’u’] appends ’ile’.

This could look like:

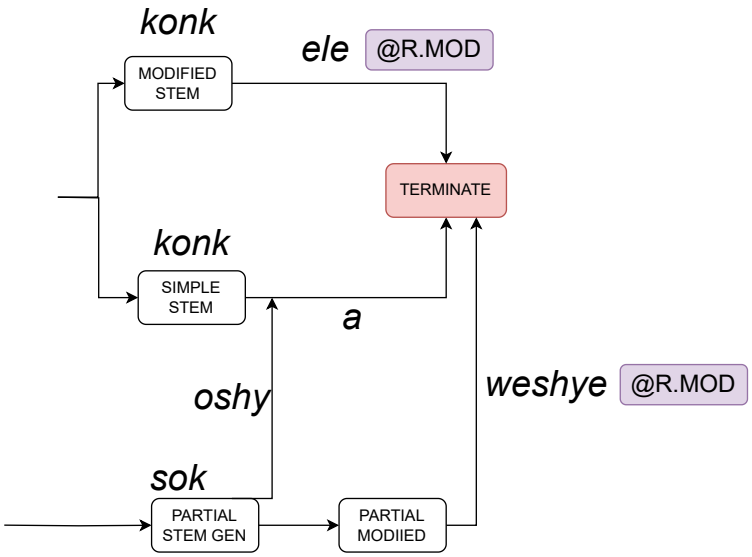

Table 3.6: Updated verb rewrite based on final syllable-based splitting

This splits the verb stem ‘sokoshy‘ into ‘sok‘ and ‘oshy‘, the later allows us to apply regular affixation to the verb root ‘sokoshy‘ while the branch enables us to append ‘weshye‘ generating the modified form ‘sokweshye‘. At this point, the complexity of having so many diacritics is becoming clear. In fact, getting the grammar to this point already showed slow compilation despite not adding additional rewrite rules. The main challenge at this point is that we are not simply fixing the verb root somewhere in this derivation path but rather performing a rewrite directly to this verb root. For example, unlike converting ’longa’ to its modified form ’longele’ where simply append the suffix ’ele’ to the root ’long’, converting the stem ’sokoshya’ to its modified form ’sokweshye’ requires replacing ’o’ with ’we’. The paths that generate this stem will simply spit out the entire string without the rewrite performed. To enable this rewrite, we either need

to split this verb into two segments that would allow us to generate the segment 'sok' and then branch from this to a variant that generates the modified form and another that generates the unmodified version. The modified form is conditioned on specific tense markers (this form only occurs with tense 7, 9, 10, and 11), so we would need more additional diacritics to track which tense is paired with the modified form and prevent paths that would generate invalid forms.

This idea could be abstracted to other verb forms. For example, consider the verb stem 'fika' (arrive):

- fika: simple
- fikila: applied (conveyed using preposition in English, e.g., to, for, in, against, with)
- fikilila: completive (completion or finality)
- fikana: reciprocal (mutual, combined action, interaction)
- fishya: causal (causing to, allowing to take place/be done)
- fikishya: intensive (add intensity, energy, or emphasis)
- fikulula: reversive (often corresponds to the English prefix un-)
- fikaula: frequentative (repetition of action)
- ifika: reflexive (against/for oneself)

So far, this only presents a simplified representation of how this network was initially constructed and leaves out details on stem forms that insert prefixes at the start of a verb (e.g., reflexive form) and other exceptions to these rewrites. For example, 'lala' (sleep), despite having 2 syllables and an a-a pattern, does not become 'lalile' but becomes 'lele'.

An obvious problem with this approach is that it adds to the already complex network of dependencies, given the addition of more diacritics. Further, we would need a sub-lexicon of forms to handle invalid derivations (consider the case of 'lala') that prevent the derivation of this form without affecting all the other derivations on which it does not break any rules.

I initially experimented with ditching the idea of splitting verbs, but instead relied on using an additional rewrite transducer that tried to post-process an unmodified output into a correct form. That is, assume we are dealing with concatenative morphology and clump all the morphemes together. After generating intermediate output, we can now begin to apply some post-processing that collapses these forms based on what would occur in the non-concatenative case. For the case of 'sokoshya', we would not output 'so' and then create 2 branches, one that generates the true form (so-¿koshya) and another that generates the modified form (so-kweshya). Instead, we would generate the full form while still outputting a string representing its modified form, i.e.:

```
'sokoshya' + 'we-e' -> 'sokoshya+we-e'
```

We then use a rewrite transducer that accepts 'oshya+we-e' to output 'weshye'. In this way, we have the path that generates the unmodified form 'sokoshya' while we keep an intermediary 'sokoshya+we-e' that gets rewritten to the true form 'sokweshye'.

This approach works to some extent, but still requires the use of diacritics to prevent invalid derivations from occurring. Consider the verb stem 'pwishya' (finish); deriving its applied form can be done by doubling the final syllable '-ishya', making it 'pwishishya' (finish for). This can be done by creating a rewrite rule that takes 'ishya+ishya' and outputs 'ishishya'. The completive forms are derived from the applied form, therefore, to generate its completive form 'pwishishishya' (finish completely), we define another rewrite transducer that takes 'ishya+ishya+ishya' to generate 'ishishishya', the completive form.

This broadly covers the main considerations and challenges with implementing this while relying solely on the use of lexicon and continuation classes. I eventually settled on defining a set of sub-lexicons for the different verb forms. Trying to rewrite these directly in the FST seemed rather complicated and made the eventual network almost impossible to compile. The Python-based approach followed a similar approach to what has been discussed earlier, except for not directly performing these rewrites in place. Firstly, this gets rid of the trouble of trying to ground the different stem types by syllable count, final syllable type, and root vowel but instead relies on regular expressions that can generate a stem lexicon for each of the 9 different stem forms. While this means having to replicate different forms of the same grammar, it simplifies the network in terms of size and readability. The final structure reflects what's represented at the start of this chapter.

3.2.3 Handling Exceptions and Irregularities

In a purely FST-based approach, addressing exceptions and irregularities would involve creating separate sub-lexicons for irregular forms and using flag diacritics to control the application of rules. This can lead to a more complex FST structure and may require manual intervention to identify and handle specific exceptions.

However, by generating lexicons in Python, we can leverage the flexibility and expressiveness of the programming language to handle exceptions more efficiently. Instead of relying solely on FST mechanisms, we can maintain a fixed mapping of verb stems and their desired outputs, effectively capturing the irregularities and exceptions.

For example, consider the verb 'lala' (sleep), which does not follow the regular modification rule of becoming 'lalile' but instead becomes 'lele'. In a Python-based approach, we can define a dictionary that maps the irregular verb stem to its corresponding modified form:

```
irregular_verbs = {
'lala': 'lele',
'fwala': 'fwele'
}
```

During the lexicon generation process, we can check if a verb stem exists in the

irregular_verbs dictionary. If it does, we can directly assign the modified form specified in the dictionary, bypassing the regular modification rules. This approach allows us to handle exceptions and irregularities in a more concise and maintainable manner.

Table 3.7 provides some examples of exceptions and irregularities encountered in Bemba morphology and how they were handled using the Python-based approach.

Word	**Regular Form**	**Actual Form** heightlala (sleep)
-lalile	-lele fwala (dress)	-fwalile
-fwele twala (carry)	-twalile	-twele temwa (love) passive
-temwile	-temwene height	

Table 3.7: Examples of exceptions and irregularities in verb morphology

In addition to handling these cases, it's worth noting a simplifying assumption I had made in regard to dealing with diacriticed characters. Contemporary written Bemba does not extensively use diacriticed characters (special characters present in certain language alphabets, denoting deviations in pronunciation). Strict grammar books often include these to draw distinctions between certain words and phrases, especially when these distinctions are drawn by variations in stress during pronunciation, yet would only be apparent when diacritics are used to indicate this. For example:

1. kola (catch, usually in reference to some living object — e.g., hunting)
2. koóla (claw at something)
3. koola (cough)

In spoken form, 'oó' in 2 often has a rising tone, while in 3, it's flat. These tones, shown with diacritics, help distinguish tense markers. Tense affixes, usually brief, rely mostly on tone. I chose to omit diacritics to reflect the language's modern use, reducing detail and sometimes making it unclear which derivation is meant without analysing the output.

3.3 Online Web Interface

As a point of interaction, a simple web interface has been implemented to provide access to the analysis and word matching functionality of the model. Figure 3.2 shows a screenshot of the web interface.

While this was not a priority, it was added as a user-friendly way to interact with the underlying morphological analyser and explore the generated analyses.

The interface consists of a text input field where users can enter a Bemba word. Upon submitting the word, the system processes the input using the morphological analyser and displays the resulting analyses. The analyses include the segmentation of the word into its constituent morphemes, along with the corresponding morphological tags and grammatical information. The web interface serves as a valuable tool for demonstrating the capabilities of the Bemba morphological analyser and making it accessible to a wider audience. it is currently hosted on https://bemmorph.martinnn.com

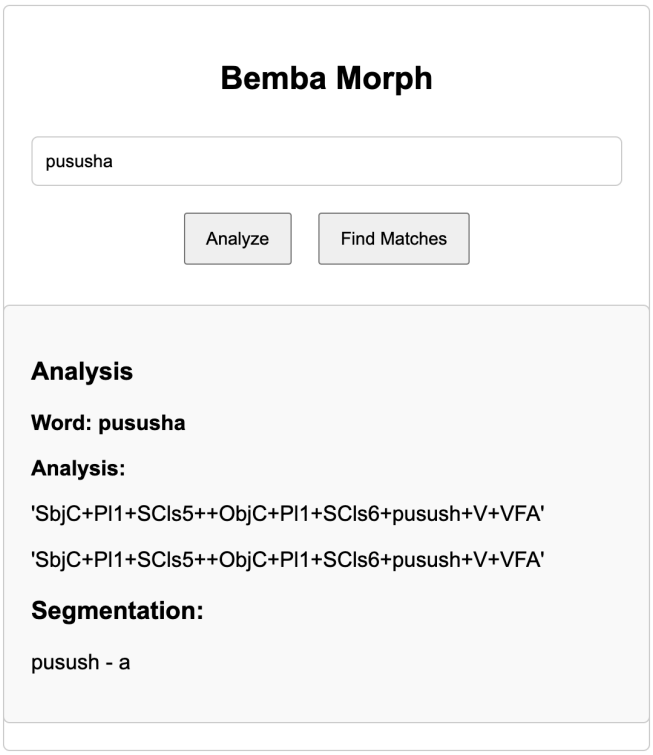

Figure 3.2: Web interface for the Bemba morphological analyser

Chapter 4

Results and Analysis

This chapter presents the results and analysis of the evaluation conducted on the Bemba morphological analyser. The evaluation specifically covered verbal and noun derivation and relied on an excerpt of words taken from the grammar book by Schoeffer (1907).

4.1 Scope and Coverage

When presenting a prototype of any kind, one of the first questions that comes to mind is: In what sense is this prototype not yet functionally complete, or, in what sense is it still a subset of the final product? This question regarding the Bemba analyzer prototype can be addressed by considering two aspects: its scope in terms of word categories and their morphological structure included, and its lexical coverage as reflected by the number of different noun stems and verb roots present in its embedded lexicon.

In terms of word categories, the current prototype focuses primarily on verbs and nouns, which form the core of Bemba morphology. The analyser covers a wide range of verbal and nominal morphological patterns, including tense, aspect, mood, subject and object concords, and derivational processes. However, other word categories such as adjectives, adverbs, and pronouns are not yet fully implemented in the prototype.

Regarding lexical coverage, the prototype's embedded lexicon contains a substantial number of noun stems and verb roots, allowing it to analyze and generate a significant portion of Bemba words. However, it is important to note that the lexicon is not exhaustive and may not include all possible noun stems and verb roots found in the language. Expanding the lexicon to achieve comprehensive coverage would be an important step towards a more complete morphological analyser.

4.2 Evaluation Methodology

An evaluation was conducted to assess the accuracy and correctness of the analyser when compared to the expected 'true' parses. As there is currently no standard dataset available for this task, a custom dataset was curated to perform the evaluation. The evaluation dataset was based on recorded information about what true derivations would

look like, as described in the grammar book by Schoeffer (1907). The authors provide extensive examples of morphological segments and how they are derived, serving as a reliable reference for constructing test cases.

For instance, consider rule 121 from the grammar book:

> TENSE 16. Prefix -le. 121. A near but less immediate future than Tenses 14, 15, and only with this force in intransitive verbs and verbs of state or condition (cf. T. 1, § 105): - e.g. nde-isa, for n-le-isa, I will shortly, or presently, come, I am coming soon.

From this rule, we can formulate a test case using the verb derivation template. Given a verb stem that uses this tense, we would expect a valid parse to be of the form:

```
<SC><OC> - le - <VROOT> <TERMINATIVE>
```

As expected, the parse for the verb 'kalefwaya' ("it is searching for/wants") generates the following analysis:

```
SbjC+Sg1+SCls6++T16fway+V+VFA
```

In the case of segmentation, the examples used in the book conveniently add hyphens to denote morphological boundaries. 'kalefwaya' would be represented as 'ka-le-fwaya'. This segmentation can be obtained from the analysis output by replacing the morphological tags with their respective outputs. SbjC+Sg1+SCls6+ generates 'ka', +T16 generates 'le', +V denotes the presence of the verb root (which can be removed as we already have the verb root), and finally, +VFA is replaced with 'a' and concatenated to the root, forming: ka-le-fwaya.

Having a model that reflects the analysis as described in the grammar book forms a valid baseline for evaluation.

4.3 Word Inflection

The evaluation process focused on two main aspects: morphological analysis and spelling correction.

To evaluate the analyser's performance in morphological analysis, a representative test set was curated based on the grammar book "A Grammar of the Bemba Language" by Schoeffer (1907). The test set consisted of carefully selected words covering various noun and verb types, considering factors such as syllable count, root vowel, and final syllable.

The test set composition included:

The choice of verb types in the test set was based on the morphological patterns observed in Bemba verbs. Verbs in Bemba can be broadly categorized based on their syllable count, root vowel, and final syllable. These factors play a crucial role in determining the morphological behavior of verbs, particularly in the formation of different verb forms such as the modified, applied, causative, and reciprocal forms.

Category	**Quantity**
Nouns	30
Verbs with 2 syllables ending in -a	42
Verbs with 2 syllables ending in -shya	12
Verbs with 3 or more syllables ending in -ba, -ka, -ma, -mya, -na, -sa, -shya, or -ta	8

Table 4.1: Quantity of grammatical categories in the test set

To elaborate, it is helpful to draw a comparison with English syntax. In English, we understand that the basic sentence structure often follows the pattern of concatenating a Noun Phrase (NP) and a Verb Phrase (VP) (**?**). For example:

- Noun Phrase (NP): ”The quick brown fox”
- Verb Phrase (VP): ”jumped over the lazy dog.”
- Combined into a sentence: ”The quick brown fox jumped over the lazy dog.”

By stripping the adjectives (quick, brown, lazy), this can be simplified to:

- Noun Phrase (NP): ”The fox”
- Verb Phrase (VP): ”jumped over the dog”
- Simplified sentence: ”The fox jumped over the dog.”

In contrast, a Bemba verb represents a condensed version of the information conveyed in an English sentence. It incorporates the subject, object, and various grammatical markers within a single word. The verb ”yachiitomboka,” for instance, can be broken down as follows:

- Subject concord: i- (class 3)
- Tense marker: -achi- (past)
- Object concord: -i- (class 3)
- Verb root: -tombok- (jump)
- Final vowel: -a

The complete verb ”yachiitomboka”[1] translates to ”it jumped over it.” The specific subject and object are not explicitly stated within the verb but are indicated by the subject and object concords. In a full sentence, the nouns representing the subject and object would be placed before and after the verb, respectively: ”inkusa yachiitomboka imbwa” (the fox jumped over the dog).

This comparison highlights the morphological complexity of Bemba verbs, where a single word encapsulates information that is typically distributed across multiple words in English. The verb ”yachiitomboka” not only conveys the action (jumping) but also specifies the participants (subject and object) and the tense.

[1]The final representation after phonological rules have been applied is ”yachiitomboka.”

Considering the density of information encoded within Bemba verbs, the morphological analyser must be capable of handling every verb form and its corresponding morphological patterns. The test set is meant to capture a representative sample of verb and noun forms, and provide a basis for evaluating how well it replicates the expected morphological patterns.

The analyser achieved full coverage on the curated test set, successfully generating morphological analyses for all the selected words. However, it is important to note that the test set primarily focused on noun and verbal derivation, and not all adjectives were included in the evaluation due to their dependency on noun classes.

4.4 Morphological Analysis

The evaluation of the morphological analyser's performance in morphological analysis yielded promising results. The analyser demonstrated high accuracy and coverage, successfully generating analyses for all the words in the curated test set, which included nouns and verbs with various syllable counts, root vowels, and final syllables. It is worth noting the entire collection of examples used in the grammar book comprised of only 48 derivations, and the analyser could accurately parse each of them. Additional words were included from my curated word list.

4.2 presents a selection of analysis outputs for different word categories.

Word Category	**Input Word**	**Analysis Output**
Verb (2 syllables, -a)	panga	pang+V+VFA
Verb (2 syllables, -shya)	eshya	eshy+V+VFA
Verb (3+ syllables, -ma)	longana	longan+VFA

Table 4.2: Examples of morphological analysis outputs

The analysis outputs demonstrate the analyser's ability to correctly identify the morphological components of words, such as noun class prefixes, verb roots, and derivational suffixes. For instance, the noun "umuntu" (person) is analyzed as consisting of the prefix "u-" (Class 1) and the noun stem "ntu". Similarly, the verb "longan" (become long) is analyzed as the verb root "long-" followed by the causative suffix "-am-" and the final vowel "-a".

These examples highlight the analyser's effectiveness in handling the morphological patterns and rules specific to Bemba nouns and verbs. The accurate segmentation and identification of morphological components enable downstream NLP tasks to leverage the rich grammatical information encoded in Bemba words.

However, it is essential to acknowledge the presence of over-generation in some cases. Over-generation occurs when the analyser produces valid morphological forms that may not be considered grammatically correct or semantically meaningful in the context of the Bemba language.

For example, consider the verb "enda" (walk). The analyser generated forms like "ba-la-enda" (they will walk) and "ba-la-chi-enda" (they will walk it). While "ba-la-enda"

is a valid form, ”ba-la-chi-enda” is semantically incongruous because the verb ”enda” is intransitive and cannot take a direct object.

Another example of over-generation is the verb ”lala” (sleep). The analyser generated forms like ”ba-la-lala” (they will sleep) and ”ba-la-mu-lala” (they will sleep him/her). Again, ”ba-la-lala” is a valid form, but ”ba-la-mu-lala” is semantically incorrect because ”lala” is an intransitive verb that does not take a direct object.

The issue of over-generation can be attributed to the lack of verb categorization and the absence of constraints on the combination of verbs with subject and object concords. The current approach generates all possible derivations for a given verb stem, without considering the semantic and syntactic restrictions imposed by the verb’s argument structure and the compatibility between the subject, verb, and object.

Despite the presence of over-generation, the analyser’s overall performance in morphological analysis remains highly accurate and demonstrates its effectiveness in handling the complex morphological patterns of Bemba nouns and verbs.

4.5 Selected Analysis Outputs

To provide a more concrete understanding of the morphological analyser’s performance, this section presents a selection of analysis outputs for different word categories and inflection patterns. These examples are drawn from the test set and demonstrate the analyser’s ability to generate accurate morphological analyses and replicate the expected patterns described in the grammar book.

4.5.1 Verb Inflection

Table 4.3 showcases a selection of verb inflection examples, covering different tenses, aspects, and moods. The input words are presented along with their corresponding morphological analyses and segmentations.

It is important to note that while the current study contributes a large list of morphemes and curated word lists, as shown in the previous tables, the analysis primarily focused on covering the verb and noun examples presented in the grammar book. The complete word lists have been developed as a resource pool for adding more rules and expanding the analyser’s coverage in future iterations.

The test cases used in the evaluation were extracted from the examples provided in the grammar book, which serves as a representative sample of word inflections and morphological patterns in Bemba. Although the grammar book does not include explicit tagsets, it uses hyphens to denote morpheme boundaries, facilitating the comparison between the analyser’s outputs and the expected segmentations.

The evaluation methodology employed in this study, which involves assessing the analyser’s ability to generate analyses that match the patterns described in the grammar book, provides a solid foundation for validating its performance and accuracy. The analyser’s success in replicating the expected inflections and segmentations demonstrates its effectiveness in modeling Bemba morphology based on linguistic rules and patterns.

Input Word	**Morphological Analysis**	**Segmentation**
nlepyanga	+SbjC+Sg1+SCls1++T1+pyang+V+VFA	n-le-pyang-a
balesenda	+SbjC+Pl2+SCls2++T1+send+V+VFA	ba-le-send-a
twaisa	+SbjC+Pl1+SCls2++T2+is+V+VFA	twa-is-a
aisa	+SbjC+Sg1+SCls1++T2+is+V+VFA	a-is-a
tulalwala	+SbjC+Pl1+SCls2++T4+lwal+V+VFA	tu-la-lwal-a
tuapyanga	+SbjC+Pl1+SCls2++T5+pyang+V+VFA	tu-a-pyang-a
natupyanga	+SbjC+Pl1+SCls2++T6+pyang+V+VFA	na-tu-pyang-a
aishile	+SbjC+Sg1+SCls1++T9+ishil+V+VFE	a-ishil-e
aishile	+SbjC+Sg1+SCls1++T10+ishil+V+VFE	a-ishil-e
alitemwa	+SbjC+Sg1+SCls1++T11+temw+V+VFE	ali-temw-e
alibika	+SbjC+Sg1+SCls1++T12+bik+V+VFA	a-li-bik-a
tuleya	+SbjC+Pl1+SCls2++T13+y+V+VFA	tu-le-y-a
tukasoma	+SbjC+Pl1+SCls2++T17+som+V+VFA	tu-ka-som-a

Table 4.3: Examples of verb inflection analysis outputs

Furthermore, the evaluation showcases the analyser's potential for generating inflected word forms from root forms, as demonstrated in Table 4.3. By accurately producing various verb forms based on the provided verb roots, the analyser exhibits its capability for morphological generation, which can be valuable for tasks such as data augmentation and language learning.

Chapter 5

Discussion

5.1 Morphological Analysis

The analyser demonstrated high accuracy and coverage in the morphological analysis evaluation. It successfully generated morphological analyses for all the words in the curated test set, which included nouns and verbs with various syllable counts, root vowels, and final syllables. It should be reiterated that the collection extracted from the grammar book only spanned 48 analyses. Additional words were included, all of which a valid analysis was

However, it is essential to acknowledge the presence of over-generation in some cases. Over-generation occurs when the analyser produces valid morphological forms that may not be considered grammatically correct or semantically meaningful in the context of the Bemba language.

For example, consider the verb ”enda” (walk). The analyser generated forms like ”ba-la-enda” (they will walk) and ”ba-la-chi-enda” (they will walk it). While ”ba-la-enda” is a valid form, ”ba-la-chi-enda” is semantically incongruous because the verb ”enda” is intransitive and cannot take a direct object.

Another example of over-generation is the verb ”lala” (sleep). The analyser generated forms like ”ba-la-lala” (they will sleep) and ”ba-la-mu-lala” (they will sleep him/her). Again, ”ba-la-lala” is a valid form, but ”ba-la-mu-lala” is semantically incorrect because ”lala” is an intransitive verb that does not take a direct object.

The issue of over-generation can be attributed to the lack of verb categorization and the absence of constraints on the combination of verbs with subject and object concords. The current approach generates all possible derivations for a given verb stem, without considering the semantic and syntactic restrictions imposed by the verb’s argument structure and the compatibility between the subject, verb, and object.

Chapter 6

Conclusion

6.1 Summary of Contributions

This study presents the development of a morphological analyser for Bemba, a Bantu language spoken primarily in Zambia, using finite-state transducers (FSTs). The analyser aims to address the lack of comprehensive computational resources for Bemba and to support various natural language processing (NLP) tasks in the language.

The main contributions of this work are as follows:

The development of a morphological analyser for Bemba based on linguistic rules and patterns derived from Bemba grammar books and linguistic literature. The analyser covers analysis of nouns and verbs for a given root lexicon.

The creation of curated word lists and a morphological tag set to represent the grammatical categories and features of Bemba words. These resources serve as the foundation for building the analyser and evaluating its performance. The implementation of morphological rules using finite-state transducers (FSTs) and the Foma toolkit, leveraging the expressive power of the formalism to handle complex morphological patterns and dependencies.

The evaluation of the analyser on a representative test set curated from Bemba grammar books, focusing on noun and verb derivation. The analyser demonstrates full coverage in morphological analysis, successfully generating analyses for all words in the test set.

The development of a web interface to provide access to the analysis and word matching functionality of the morphological analyser, making it accessible to a wider audience for language learning, linguistic research, and the development of language technologies for Bemba.

The identification of areas for improvement and future research, such as addressing over-generation through more fine-grained constraints and verb categorization.

This study contributes to the broader field of computational morphology for Bantu languages, demonstrating the effectiveness of FSTs in modeling the complex morphological systems of these languages. The Bemba morphological analyser developed

in this work serves as a valuable resource for various NLP tasks, such as machine translation, information retrieval, and text generation, enabling the processing and understanding of Bemba text.

6.2 Final Remarks

The development of the Bemba morphological analyser using finite-state transducers (FSTs) has been a significant step towards addressing the lack of comprehensive computational resources for the language. The analyser has demonstrated its ability to handle a wide range of morphological phenomena in Bemba and to generate accurate analyses for noun and verb derivation.

The evaluation results highlight the analyser's high accuracy and coverage in morphological analysis, successfully generating analyses for all words in the test set. These achievements underscore the potential of the analyser as a valuable tool for NLP tasks and linguistic research in Bemba.

However, it is important to acknowledge the limitations and areas for future improvement. The current analyser primarily focuses on noun and verb morphology, and further work is needed to expand its coverage to other word categories, such as adjectives, adverbs, and pronouns. Additionally, the issue of over-generation in some cases highlights the need for more fine-grained constraints on subject-verb-object relations and verb categorization. Paramount to the issues discussed above is the need for a gold standard test set. It was challenging to form a comparable standard of performance if no such benchmark exists. More work may be needed in developing a test.

Future research could explore the integration of the morphological analyser with other NLP tasks, such as part-of-speech tagging, parsing, and machine translation, to enhance the processing and understanding of Bemba text. The development of parallel corpora and lexical resources for Bemba would also greatly benefit the advancement of NLP applications in the language.

The web interface developed as part of this study serves as a valuable tool for making the morphological analyser accessible to a wider audience. However, there is scope for further enhancements, such as the inclusion of additional features like visualizations of morphological structures, integration with other linguistic resources or corpora, and the ability to export analysis results for further processing or research purposes.

In conclusion, this study has laid the foundation for the development of computational resources for Bemba and has demonstrated the effectiveness of finite-state transducers in modeling the language's complex morphological system. The Bemba morphological analyser serves as a valuable resource for NLP tasks and opens up possibilities for further research and applications in the language. With continued efforts and collaboration, we can work towards building more comprehensive and robust language technologies for Bemba and other under-resourced languages.

The contributions of this work extend beyond the development of the morphological analyser itself. The creation of curated word lists and a morphological tag set for Bemba provides a solid foundation for future research and resource development in the

language. These resources can be utilized by other researchers and developers working on Bemba language technologies, promoting standardization and interoperability across different tools and applications.

Furthermore, the methodology and techniques employed in this study can be adapted and applied to the development of morphological analysers for other Bantu languages. The use of finite-state transducers, the incorporation of linguistic knowledge, and the evaluation strategies described in this work can guide similar efforts in the field of Bantu language processing.

The web interface developed as part of this study serves as a starting point for making the morphological analyser accessible to a wider audience. By providing a user-friendly platform for interacting with the analyser, the web interface facilitates the exploration and analysis of Bemba words, promoting language learning, linguistic research, and the development of language technologies. Future enhancements to the web interface, such as the integration of additional features and resources, can further enhance its usability and impact.

It is important to recognize the significance of this work in the context of preserving and promoting linguistic diversity. Bemba, like many other under-resourced languages, faces challenges in terms of digital presence and computational resources. The development of the Bemba morphological analyser contributes to the efforts in language documentation, preservation, and revitalization. By providing tools and resources for processing and understanding Bemba text, this work supports the maintenance and growth of the language in the digital age.

Moreover, the insights gained from the development and evaluation of the Bemba morphological analyser can inform broader discussions on the role of computational linguistics in supporting linguistic diversity and language vitality. The challenges encountered and the solutions proposed in this study can contribute to the ongoing discourse on best practices and strategies for developing language technologies for under-resourced languages.

Bibliography

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. Openfst: A general and efficient weighted finite-state transducer library. pages 11–23, 2007.

Kenneth R Beesley and Lauri Karttunen. *Finite-state morphology*. CSLI publications, 2003.

Emily M Bender. *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. Springer Nature, 2022.

Burcu Can and Suresh Manandhar. Methods and algorithms for unsupervised learning of morphology. In *Computational Linguistics and Intelligent Text Processing*, pages 177–205. Springer, 2014.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2002. URL https://www.aclweb.org/anthology/W16-2002.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1):3, 2007.

Robert Elwell. A morphological analyzer for Swahili. http://aflat.org/files/swahili.pdf, 2008. Accessed: 2023-03-28.

Gülşen Eryıit and Ekin Adalı. Affix stripping morphological analyzer for Turkish. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 33–36, 2008.

Silke Hamann and Nancy C Kula. Laryngeal features in Bemba. *Journal of African Languages and Linguistics*, 36(1):1–25, 2015.

Harald Hammarström and Lars Borin. Unsupervised learning of morphology. In *Computational Linguistics*, volume 37, pages 309–350. MIT Press, 2011.

Mans Hulden. Foma: a finite-state compiler and library. pages 29–32, 2009.

Arvi Hurskainen. Helsinki Corpus of Swahili. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1175–1178. 2004.

Lauri Karttunen. Finite-state technology. *The Oxford Handbook of Computational Linguistics*, 2003. doi: 10.1093/oxfordhb/9780199276349.013.0023.

Nancy C Kula. Morphology and the lexicon of Bemba. In *Proceedings of the 3rd World Congress of African Linguistics*, pages 255–270, 2002.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. Hfst—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer, 2011.

Lutz Marten. A comparative analysis of Bantu verbal morphology. In *Bantu Grammar: Description and Theory*, pages 25–50. Mouton de Gruyter, 2006.

Jackson Muhirwe. Computational analysis of Kinyarwanda morphology: The morphological alternations. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1959–1962, 2007.

Derek Nurse and Gérard Philippson. *The Bantu languages*. Routledge, 2003.

Laurette Pretorius and Sonja E Bosch. Computational morphology of the nguni languages: a case study. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 189–196, 2005.

Laurette Pretorius and Sonja E Bosch. Finite-state morphology of the nguni language cluster: modelling and implementation issues. In *Proceedings of the 10th International Conference on Finite-State Methods and Natural Language Processing*, pages 72–81, 2009.

Brian Roark and Richard Sproat. *Computational Approaches to Morphology and Syntax*. Oxford University Press, 2007.

A Schoeffer. *A Grammar of the Bemba Language as Spoken in North-East Rhodesia*. Macmillan, 1907.

Claytone Sikasote, Eunice Mukonde, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. BIG-C: a multimodal multi-purpose dataset for Bemba. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2062–2078, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.115. URL https://aclanthology.org/2023.acl-long.115.

Gary F Simons and Charles D Fennig. Bemba. https://www.ethnologue.com/language/bem, 2018. Accessed: 2023-03-28.

Debra Spitulnik. Semantic superstructuring and infrastructuring: Nominal class struggle in ChiBemba. *Chicago Linguistic Society*, 23:397–414, 1987.

Sara Stymne. Evaluating morphological analysis for machine translation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*

(LREC'08), Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/808_paper.pdf.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

Appendix A

First appendix

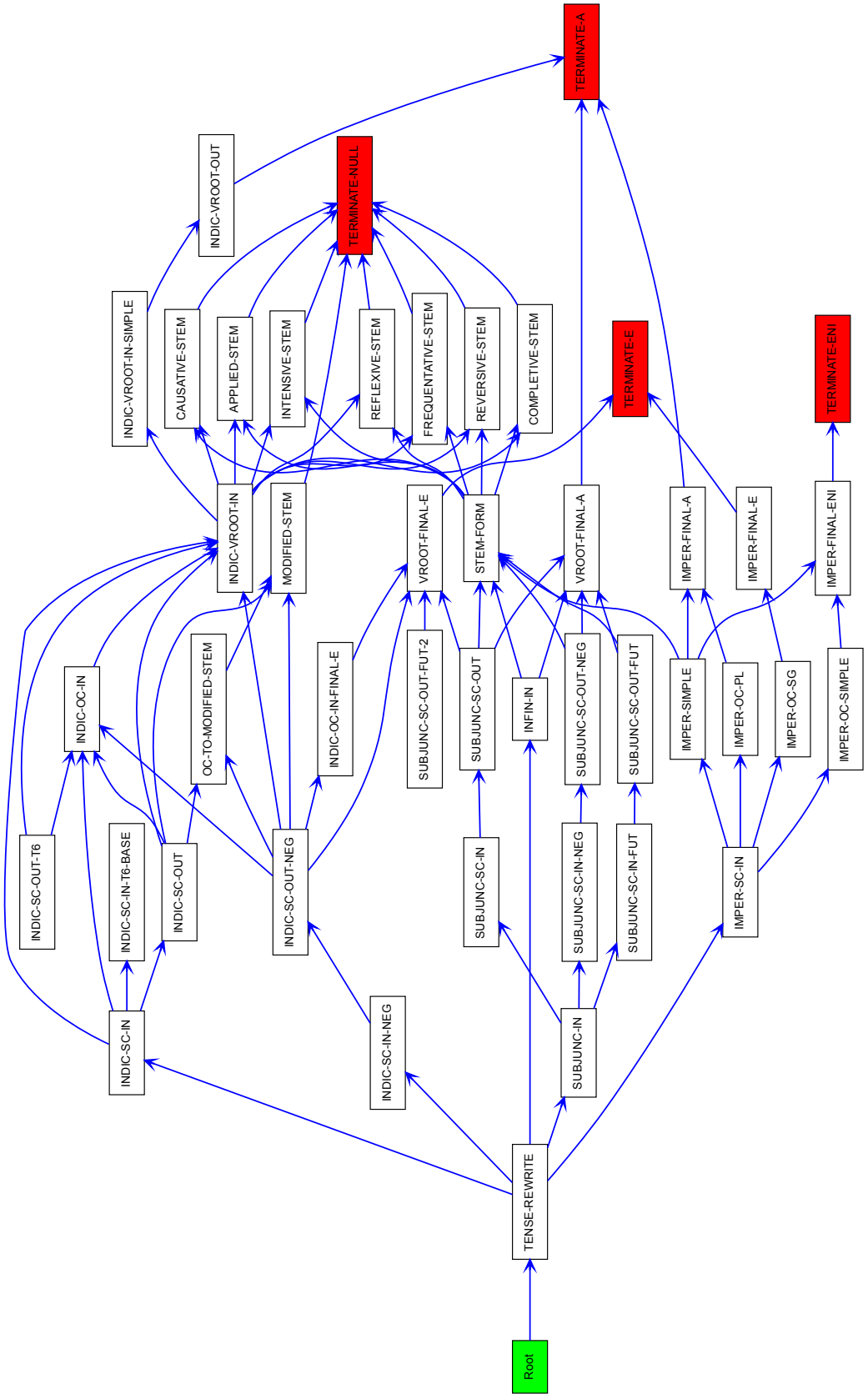

Figure A.1: Derivation graph of the Bemba morphological analyser FST