



Investigating supervised surface segmentation of isiZulu text using synthetic data generation

SN Mkhwanazi



orcid.org 0009-0003-3050-4744

Dissertation accepted in fulfilment of the requirements for the degree *Master of Science in Computer Science* at the North-West University

Supervisor: Dr L Marais
Co-supervisor: Prof R Goede

Graduation: July 2025

A white rectangular box used to redact information, likely a student number or contact details.

PLAGIARISM DECLARATION

I Sthembiso Nkosentsha Mkhwanazi
(full name and surname and student number) hereby declare that this assignment / paper / project / portfolio is my own work.

I further declare that:

1. the text and bibliography reflect the sources I have consulted, and
2. where I have made reproductions of any literary or graphic work(s) from someone else, I have obtained the necessary prior written approval of the relevant author(s)/publisher(s)/creator(s) of such works and/or, where applicable, from the Dramatic, Artistic and Literary Rights Organisation (DALRO).
3. sections with no source referrals are my own ideas, arguments and/or conclusions.

Signature:  Student number: _____

Date: 19 March 2025

Acknowledgements

I begin by thanking God for His grace in my life, which has enabled me not only to *will* but also to *do* this work (Philippians 2:13). Throughout it all, I remain deeply grateful for the incredible people I’ve had the privilege to meet - serendipitously - many of whom have contributed significantly to this study in ways that words may never fully capture. I will, however, attempt to acknowledge a few by name.

First and foremost, I wish to express my sincere gratitude to Dr Laurette Marais, my main supervisor, for her support throughout this journey. Her guidance, constructive feedback, and steady encouragement have been invaluable in the development and completion of this dissertation, which I’m glad to present after a long and challenging process.

My heartfelt thanks also go to Prof. Roelien Goede, who kindly assumed the role of co-supervisor for this study. Her expert advice, supportive feedback, and readiness to assist despite her demanding schedule have been invaluable throughout this journey.

I am deeply grateful to my family for their enduring love and unwavering support, especially during moments of doubt, even when they did not fully understand the “why” behind this path. To my mother and my siblings *Xolile*, *Sandile*, and *Akhona*, I just wanna say, “*ngiyabonga ngakho konke*”.

To my colleagues at the Council for Scientific and Industrial Research (CSIR), thank you for your collaboration, encouragement, and shared pursuit of ‘Touching lives through innovation’, and for being EPIC. Your support has made this process even more meaningful.

I also wish to acknowledge the intellectually stimulating environments of the seminars where this work was preliminarily presented, including DHASA, AI Expo Africa, IndabaX South Africa, and Hundzula Retreat. These platforms not only allowed me to share my work but also provided valuable discussions and feedback that refined this study.

Finally, I extend my deepest gratitude to the CSIR as an organisation for financially sponsoring this study, and to the National Integrated Cyber Infrastructure System (NICIS) and its Centre for High Performance Computing (CHPC) for providing the infrastructure that enabled the technical execution of this research.

Abstract

IsiZulu, one of South Africa’s most widely spoken languages, is classified as a low-resource language, especially regarding digital tools. As part of the Nguni family, isiZulu exhibits complex morphology and conjunctive orthography. These features result in data sparsity, as a single root or stem may appear in numerous morphological variants, complicating language modelling. This underscores the importance of morphological segmentation, a natural language processing NLP task that decomposes words into their smallest meaningful units (morphemes). Rule-based methods yield high accuracy in low-resource contexts but typically lack robustness and are costly to develop. Conversely, machine learning approaches require large, high-quality datasets, often unavailable for low-resource languages. To address these challenges, this study employs a hybrid approach using a rule-based system, the isiZulu Resource Grammar (ZRG), to generate synthetic datasets with varying segmentation granularities. These datasets underwent data augmentation through syntactic tree manipulation, significantly increasing their size and diversity. Subsequently, this data trained supervised machine learning models: Conditional Random Fields (CRF), Long Short-Term Memory (LSTM), and Transformer-based models—for morphological segmentation. The effectiveness of these models was assessed intrinsically, using precision, recall, F1 score, BLEU, and chrF, and extrinsically, by evaluating their impact on Neural Machine Translation (NMT) quality for isiZulu-English translation. Intrinsic evaluation showed that the Transformer model consistently outperformed the CRF and LSTM models, achieving segmentation accuracy above 0.9 across all metrics and granularity styles. Additionally, the hybrid approach demonstrated superior robustness, effectively handling out-of-vocabulary (OOV) words and performing segmentation 30 times faster than ZRG alone. Extrinsic evaluation confirmed that segmentation improved translation quality, with Segmenter Two achieving the highest BLEU score (0.235), representing a 25.0% improvement over the unsegmented baseline (0.188). These findings highlight the effectiveness of integrating rule-based and machine learning approaches for morphological segmentation, offering a scalable solution for processing low-resource languages with complex morphologies such as isiZulu in NLP applications.

Keywords: Agglutinative Languages; Morphological Segmentation; isiZulu; Supervised Segmenter Learning; Rule-Based Segmenter.

Contents

1	Introduction	1
1.1	Introduction to the Study	1
1.2	Background and Contextualisation	2
1.3	Problem Statement	6
1.4	Research Questions, Objectives and Hypotheses	7
1.4.1	Research Questions	7
1.4.2	Research Objectives	7
1.5	Research Methodology	8
1.5.1	Hypotheses	8
1.6	Ethical Considerations	9
1.6.1	Data Collection and Use	9
1.6.2	Privacy and Cultural Sensitivity	10
1.6.3	Transparency and Reproducibility	10
1.7	Dissertation Structure	10
1.8	Chapter Summary	11
2	Research Methodology	13
2.1	Introduction	13
2.2	The Concept of Research	14

2.3	Research Paradigms	15
2.3.1	Ontological Assumptions	15
2.3.2	Epistemological Assumptions	16
2.3.3	Axiological Assumptions	16
2.3.4	Methodology	16
2.4	Research Philosophy	18
2.4.1	Positivism	18
2.4.2	Interpretivism	19
2.4.3	Pragmatism	20
2.4.4	Critical Social Theory	20
2.4.5	Critical Realism	21
2.5	Positioning the Present Study	22
2.6	Hypothetico-Deductive Methodology	23
2.6.1	Application in this Research	25
2.7	Chapter Summary	30
3	Morphological Segmentation	31
3.1	Introduction	31
3.2	Background and Linguistic Characteristics of isiZulu	31
3.2.1	Background and History of isiZulu	31
3.2.2	Linguistic Characteristics of isiZulu	34
3.2.3	Morphological Characteristics: Overview	38
3.3	Morphological Segmentation: Overview	43
3.3.1	Introduction to Morphological Segmentation	44
3.3.2	Canonical Segmentation	44

3.3.3	Surface Segmentation	44
3.4	Approaches to Morphological Segmentation	45
3.4.1	Rule-Based Approaches	45
3.4.2	Statistical Machine Learning Approaches	52
3.4.3	Deep Learning Approaches	63
3.5	Morphological Segmentation Metrics	74
3.5.1	Intrinsic evaluation	74
3.5.2	Extrinsic Evaluation	79
3.6	Chapter Summary	81
4	Data Preparation	82
4.1	Ukwabelana Dataset	83
4.1.1	Dataset Composition	83
4.1.2	Development and Annotation Process	84
4.1.3	Limitations and Challenges	84
4.2	National Centre for Human Language Technology Text Corpora	84
4.2.1	Dataset Composition	85
4.2.2	Development and Annotation Process	85
4.2.3	Limitations and Challenges	86
4.3	Reflection on ZRG's Limitations and Advantages	86
4.3.1	Limitations of the ZRG	87
4.3.2	Strengths of the ZRG	89
4.4	Data Acquisition	91
4.4.1	Data Pre-Processing	92
4.4.2	Data Organisation	93

4.5	Data Parsing	93
4.5.1	Dataset Preparation and Batching	94
4.5.2	Parallel Processing with Docker	94
4.5.3	Runtime Parsing and Output Management	94
4.5.4	Error Handling and Output Consolidation	95
4.5.5	Parsing Output Example	95
4.5.6	Processing Trees	98
4.5.7	Example of Application	98
4.6	Data Augmentation	101
4.6.1	Workflow for Data Augmentation	101
4.6.2	Augmentation Techniques	102
4.7	ZRG Linearisation	108
4.7.1	Different Linearisation Strategies	108
4.7.2	Application of Linearisation Strategies	109
4.8	Final Dataset Output	111
4.9	Chapter Summary	112
5	Design of Models	113
5.1	Introduction	113
5.2	Models Selection	114
5.2.1	Criteria for Model Selection	114
5.2.2	Models Considered	115
5.3	Models Design and Training	118
5.3.1	Transformer Models	118
5.3.2	Long Short-Term Memory Models	132

5.3.3	Conditional Random Field Models	134
5.4	Chapter Summary	139
6	Experiment Evaluation	141
6.1	Model Performance Analysis: Training and Validation	142
6.1.1	Training Loss	142
6.1.2	Validation Loss	143
6.2	Intrinsic Evaluation	145
6.2.1	Model Selection for Downstream Evaluation	148
6.2.2	Data Augmentation Impact	149
6.2.3	Investigating the Robustness of the Transformer Models	150
6.2.4	Investigating Efficiency of the Transformer Model	152
6.3	Extrinsic Evaluation	155
6.3.1	Neural Machine Translation Model	155
6.3.2	Neural Machine Learning Evaluation	158
6.4	Chapter Summary	165
7	Findings, Conclusion, and Future Research	167
7.1	Introduction	167
7.2	Summary of the Chapters	168
7.3	Findings Regarding Objectives and Research Questions	170
7.3.1	Summary of Achievements in Relation to the Objectives	172
7.4	Evaluating the Hypotheses	174
7.4.1	First Hypothesis: Performance of the Hybrid System	174
7.4.2	Second Hypothesis	177
7.5	Answering the Research Questions	179

7.5.1	Research Question 1	179
7.5.2	Research Question 2	180
7.6	Challenges and Limitations	181
7.6.1	Challenges	181
7.6.2	Limitations	182
7.7	Conclusion	183
A	Appendix	209
A.1	Models Hyper-Parameters	209
A.1.1	Hyper-Parameters for Unaugmented Models	209
A.1.2	Hyper-Parameters for English-to-IsiZulu Translation Models	209

List of Figures

2.1	Summary of the research methodology.	29
3.1	Distribution of isiZulu Speakers Across South African Provinces, Based on Statistics 2022 Data	32
3.2	Official South African Southern Bantu languages Guthrie group.	33
3.3	A simple diagram of the encoder-decoder.	64
4.1	Parse abstract syntax tree generated by ZRG with default segmentation. . .	97
4.2	The first found subtrees from the bigger tree.	99
4.3	The second found subtrees from the bigger tree.	100
4.4	The abstract syntax tree for “ <i>amantombazane amahle amathathu ahlala ekhaya nomama wabo</i> ”.	105
4.5	The abstract syntax tree for “ <i>ubaba kade enezimoto ezimbili ezimbi ezishayelwa abafana bakhe abadlala ibhola</i> ”.	106
4.6	The abstract syntax tree for “ <i>abafana bakabhuti banezinja ezinenkani</i> ”. . . .	110
5.1	An adapted vanilla Transformer architecture.	130
6.1	Transformer: training loss over epochs for different Segmenters.	143
6.2	LSTM: training loss over epochs for different segmenters.	144
6.3	Transformer: validation loss over epochs for different segmenters.	145
6.4	LSTM: validation loss over epochs for different segmenters.	146

6.5	Training loss across epochs: Translation models with varying segmentation strategies.	159
6.6	Validation loss across epochs: Translation models with varying segmentation strategies.	160
6.7	Validation loss across epochs: for the Unsegmented model.	161

List of Tables

3.1	Distribution of languages spoken in South Africa.	34
3.2	Noun classes in isiZulu showing singular and plural examples.	42
3.3	Contingency table: confusion matrix.	75
4.1	Manipulation of terms and resulting sentences.	107
4.2	Comparison of linearisation strategies for the sentence “ <i>abafana bakabhuti banezinja ezinenkani</i> ”.	110
5.1	Summary of data splitting for training, validation, and testing.	126
5.2	Transformer model architecture parameters.	126
5.3	Hyper-parameter search space configuration.	127
5.4	Final model configurations for segmenters across different linearisation strategies.	131
5.5	LSTM model default configuration.	133
5.6	Hyper-parameter search space.	134
5.7	LSTM optimal hyper-parameters for Segmenters One, Two, and Three. . .	134
5.8	CRF model base parameters.	137
5.9	Grid search parameter ranges for GridSearch.	137
5.10	Optimal CRF parameters across different segmentation schemes.	139
6.1	Morphological segmentation examples: input words with target and Segmenter One predicted segmentations.	146

6.2	Model performance comparison across different segmentation approaches (BLEU and chrF scores).	147
6.3	Comparison of precision, recall, and F1 scores across different architectures and segmentation styles.	148
6.4	Comparison of key metrics scores between Transformer models trained on unaugmented data and augmented data.	149
6.5	Segmentation results from different segmenters.	151
6.6	Gold standard segmentations and failure analysis.	151
6.7	Evaluation metrics for different segmenters.	152
6.8	Execution times for segmenters using ZRG and Transformer models.	153
6.9	Optimal hyper-parameters for translation models across different segmentation approaches.	158
6.10	IsiZulu–English translation metrics for Segmentation models (test set).	162
6.11	Translation quality metrics for different Segmentation models in English-isiZulu translation.	162
6.12	Translation scores for Segmentation models on updated FLORES (isiZulu–English).	163
6.13	Translation quality metrics scores for different segmentation styles using the updated FLORES dataset (English-isiZulu).	165
A.1	Model configurations for The Unaugmented Segmentation Models	209
A.2	Model Configurations for English-to-IsiZulu Translation With and Without Segmentation	209

Table of abbreviations

A table containing a list of abbreviations that will be used throughout text.

ZRG	isiZulu Resource Grammar
NMT	Neural Machine Translation
NLLB	No Language Left Behind
FLORES	Few-shot Learning for Zero-shot Evaluation in Translation
CRF	Conditional Random Fields
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Units
BLEU	Bilingual Evaluation Understudy
chrF	Character F-score
MEMM	Maximum Entropy Markov Model
OOV	Out-of-Vocabulary
POS	Part-of-Speech
GF	Grammatical Framework
AST	Abstract Syntax Tree
ML	Machine Learning
HPO	Hyper-Parameter Optimisation
GPU	Graphics Processing Unit
PGF	Portable Grammar Format
AI	Artificial Intelligence
NLP	Natural Language Processing
RNN	Recurrent Neural Network
FFN	Feed-Forward Network
Lin A	Fine-grained Segmentation Strategy
Lin B	Moderate Segmentation Strategy
Lin C	Coarse-grained Segmentation Strategy
HPO	Hyper-Parameter Optimisation

Chapter 1

Introduction

1.1 Introduction to the Study

Morphological segmentation is a Natural Language Processing (NLP) task that has shown much potential to improve the automatic processing of morphologically complex languages such as isiZulu (Creutz, 2006; Creutz & Lagus, 2007; Mager *et al.*, 2022; Tukeyev *et al.*, 2020). Approaches to segmentation be divided into two broad categories: rule-based systems and data-driven (machine learning) techniques, which can be further divided into unsupervised, supervised, and semi-supervised methods. While fully supervised methods typically achieve high performance, they rely on substantial amounts of annotated data resources that are often unavailable for low-resourced languages like isiZulu. As a result, researchers often have to rely on rule-based, unsupervised, or semi-supervised methods, which require fewer annotations. However, these methods present notable limitations, including lower accuracy, scalability challenges, and a heavy dependence on linguistic expertise. These constraints hinder their integration into NLP pipelines that require high-precision segmentation, emphasising the need for more effective supervised approaches tailored to low-resource languages.

The present study aims to present insights into the use of an existing rule-based system to generate synthetic data for investigating supervised morphological surface segmentation of isiZulu at different granularity levels. The goal is to leverage the strengths of both rule-based systems and machine learning approaches to overcome the weaknesses of each approach.

Due to the lack of substantial annotated data for morphological segmentation, this study uses isiZulu Resource Grammar (ZRG), which is a rule-based grammar developed using the Grammatical Framework (GF) programming language, to generate morphologically surface-segmented isiZulu text with three different segmentation styles. Additionally, the

study explores data augmentation as a strategy to expand the dataset further. ZRG performs a deep morphosyntactic analysis of the text, which, while linguistically robust, is computationally slow. Furthermore, as a rule-based system, it is brittle when encountering words outside its predefined lexicon, limiting its generalisability.

To enhance segmentation beyond the limitations of rule-based methods, the research adopts a machine learning-based approach to morphological surface segmentation, leveraging supervised learning techniques. These methods are preferred due to their adaptability and superior performance, especially in real-world applications where flexibility and robustness are critical.

1.2 Background and Contextualisation

IsiZulu is one of South Africa’s 12 official languages. It is a member of the Nguni language group, which forms part of the larger Niger-Congo language family (Mesham *et al.*, 2021). Despite being widely spoken, isiZulu, along with other Nguni languages (isiXhosa, Siswati, and isiNdebele), is considered a low-resourced language. Indeed, isiZulu is the most widely spoken language in South Africa, with approximately 15 million native speakers, accounting for 24% of the country’s population (StatsSA, 2022b).

One of the notable characteristics of Nguni languages is their agglutinating morphology and conjunctive orthography, as compared to disjunctive orthography, which is commonly associated with other South African languages, more specifically the Sotho languages. While Sotho languages share the agglutinating morphology with isiZulu and other Nguni languages, they have disjunctive orthography, where morphemes are written with spaces in between (Bosch & Pretorius, 2002; Taljard & Bosch, 2006).

IsiZulu is considered a low-resourced language due to limited access to linguistic resources such as machine-readable texts. This scarcity of resources has significantly hindered the development of technological tools that are necessary for ensuring the long-term digital preservation and vitality of isiZulu and similar languages (Loubser & Puttkammer, 2020). This challenge is particularly pronounced for resource-scarce agglutinative languages, where a single root or stem can combine with numerous morphemes to generate hundreds of unique word forms. For Nguni languages like isiZulu, the use of conjunctive orthography further complicates this issue. The result is data sparsity, as even large corpora may fail to encompass the full range of root and stem variations. This data sparsity creates significant challenges for language modelling approaches that rely on token identities. Consequently, these challenges have motivated the present research into morphological segmentation as a potential solution.

Morphological segmentation refers to the process of decomposing an orthographic word (its written form) into its constituent morphemes, which are the smallest meaning-bearing units of language (Creutz & Lagus, 2007). This technique has been found to be useful in NLP applications such as machine translation (MT) (Mager *et al.*, 2022; Tukeyev *et al.*, 2020) and automatic speech recognition (ASR) (Creutz, 2006), mainly when working with low-resource languages that feature complex morphological structures. This entails addressing the combinatorial explosion inherent in agglutinative morphology, where modifications in morpheme attachment to a root or stem can completely alter a word’s meaning. Such complexity often results in corpora containing limited variations of stems, leading to language models that struggle with out-of-vocabulary (OOV) issues when morphological segmentation is not applied. By breaking words into smaller, meaningful units, morphological segmentation facilitates better generalisation and improves performance in language modelling tasks for morphologically rich, resource-scarce languages.

The effort towards morphological segmentation has been divided into two goals: canonical segmentation and surface segmentation (Cotterell *et al.*, 2016a). Canonical segmentation entails segmenting a word into its underlying morphemes, with the morphemes represented in their canonical or lemmatised forms. The concatenation of these morphemes does not necessarily result in the orthographic word, whereas with surface segmentation, the word is divided into morpheme-based substrings (morphs) that can be combined to form the orthographic word (Cotterell *et al.*, 2016a).

There are two common techniques that are used for morphological segmentation: rule-based and machine learning-based (ML) approaches (Anand Kumar *et al.*, 2010; Goldsmith, 2001). Rule-based segmentation relies on predefined rules based on expert knowledge as well as the morphological structure of the language (Eiselen & Puttkammer, 2014a). The rule-based methods are considered to be time-consuming and expensive to develop, and since they use rules to segment texts, they tend to be fragile when encountering new words not within the lexicon (Villegas-Ch *et al.*, 2024). Despite these drawbacks, rule-based methods remain helpful when developed more especially in languages with well-documented morphological rules, serving as baselines or complementary tools when using hybrid approaches (Grönroos *et al.*, 2016; Zbib *et al.*, 2012).

Machine learning-based segmentation, on the other hand, uses algorithms to learn from data and identify morphemes based on stochastic and probabilistic methods (Goldsmith, 2001). Unlike rule-based systems, these techniques do not rely on explicitly defined rules; instead, they are heavily dependent on the availability of sufficient training data, which may compromise their reliability if inadequate. The primary machine learning methods are supervised, unsupervised, semi-supervised, and reinforcement learning (Murphy, 2012a), with the first three being the most commonly used in morphological segmentation.

The data set used in these machine learning-based approaches for morphological segmentation can either be unsegmented or segmented by having predefined morphological boundaries before being used to train the model. The former approach corresponds with unsupervised segmentation, where the model learns segmentation patterns directly from unannotated data, while the latter corresponds with supervised segmentation, where the model is trained on data with predefined segmentation (Eskander *et al.*, 2022). The third approach, known as semi-supervised, combines these approaches. All these techniques require a substantial amount of data to learn the segmentation pattern accurately, and their inherent adaptability enables them to perform well even on new words. This flexibility renders them suitable for deployment in real-world applications where not all words are present in the training set. However, obtaining sufficient annotated or unannotated data is a significant challenge, particularly for low-resource languages.

Morphological segmentation plays a critical role in NLP, particularly for morphologically complex and low-resource languages like isiZulu. The choice between surface and canonical morphological segmentation is influenced by several factors, including the specific linguistic characteristics of the language in question, the intended application of the segmentation, and the availability of annotated data (Occhipinti, 2024; Rice *et al.*, 2024).

When comparing surface segmentation with canonical segmentation, surface segmentation presents distinct advantages, particularly for language modelling. Canonical segmentation decomposes a word into its underlying morphemes in their canonical (or lemmatised) forms (Cotterell *et al.*, 2016a). While valuable for morphological analysis, these resulting morphemes often do not correspond directly with the actual orthographic forms of the language. This discrepancy can be problematic for language modelling, because the model may struggle to map canonical forms back to their surface realisations. In contrast, surface segmentation divides words into morpheme-based substrings that can be reassembled into the original orthographic word. This method captures the surface forms of words, including phonological variations that occur when morphemes combine.

A fundamental challenge in terms of surface segmentation is determining the optimal segmentation granularity; that is the level of detail or fineness in which words are segmented into their morphemic components. When surface segmentation is used, a word may have multiple valid segmentations depending on the granularity level. However, determining the optimal segmentation granularity for training language models remains an open research question (Ataman & Federico, 2018; Meyer & Buys, 2022; Salesky *et al.*, 2020). If segmentation is too aggressive, language models may have to learn too much compositional semantics to accurately capture the meaning, while if it is too conservative, the data sparsity problem may not be sufficiently addressed.

One way that has been explored to address this paradigm is to use methods such as Byte-pair Encoding (BPE), where the hyper-parameter fixing granularity must be optimised

(Salesky *et al.*, 2020). While BPE and other subword tokenisation techniques attempt to address this issue, there is no consensus on the best granularity level for morphologically rich languages (Salesky *et al.*, 2020).

Most morphological (canonical or surface) segmentation work in the literature follows the unsupervised approach, which is especially common for low-resource languages like isiZulu that lack sufficient annotated data (Bergmanis & Goldwater, 2017; Creutz *et al.*, 2007a; Goldsmith, 2001; Hammarström & Borin, 2011; Mzamo *et al.*, 2019a). This technique is attractive because it can be applied to any language with adequate digital textual data without needing expert annotations (Can & Manandhar, 2014; Ruokolainen *et al.*, 2016). However, despite its popularity, the unsupervised approach tends to achieve lower accuracy in morphological segmentation due to the lack of linguistic guidance (Wang *et al.*, 2016b). This limitation reduces their effectiveness in downstream applications, where precise segmentation may have a significant effect on performance.

An alternative to unsupervised morphological segmentation is supervised segmentation, which relies on annotated datasets (Moeng *et al.*, 2022; Pranjić *et al.*, 2024; Puttkammer & Du Toit, 2021). Supervised machine learning approaches are often preferred due to their higher accuracy and their ability to incorporate linguistic knowledge (Belth, 2024). Once trained, these models can efficiently segment words, generalising well to unseen data, making them more robust and less brittle than rule-based approaches. However, a major drawback of supervised segmentation is the need for large amounts of annotated data, which is costly and labour-intensive to produce, especially for low-resource languages. To mitigate this, researchers have explored semi-supervised approaches, which leverage a small set of annotated data alongside unlabelled data to improve performance while reducing annotation effort (Ruokolainen *et al.*, 2014; Yusupujang, 2018).

In contrast, rule-based morphological segmentation relies on predefined linguistic rules crafted by experts to reflect a language’s morphological structure (Eiselen & Puttkammer, 2014a; Villegas-Ch *et al.*, 2024). While this approach can yield high accuracy in well-defined linguistic contexts, it has several drawbacks. Rule-based systems are labour-intensive, expensive to develop, and often struggle with adaptability (Dasgupta & Ng, 2007; Villegas-Ch *et al.*, 2024; Westhelle *et al.*, 2022). Their performance also deteriorates when encountering new or unpredictable words, limiting their applicability in dynamic linguistic environments. Scalability is another significant challenge; for languages with complex morphologies, the large number of required rules can be overwhelming. Furthermore, morphophonological alternations introduce additional complexities, making it difficult to capture all variations through static rules (Joseph & Anto, 2015; Selvam & Natarajan, 2009).

The choice of segmentation approach depends on available linguistic resources and intended applications. Rule-based methods are effective when comprehensive linguistic

knowledge is available, whereas unsupervised approaches are preferable when annotated data is scarce. Supervised learning, despite its reliance on annotated data, offers superior accuracy and generalisation when adequate training data is available. In data-sparse settings, semi-supervised techniques provide a viable compromise, leveraging small annotated corpora alongside large raw text corpora for enhanced learning.

To date, supervised surface segmentation for isiZulu remains underexplored, despite its potential to improve downstream NLP tasks. A primary limitation in this regard is the lack of annotated surface segmented data required for training such models. To address this gap, the present study leverages the ZRG (Marais & Pretorius, 2023), a rule-based system implemented using the GF programming language.

The ZRG allows for the morphosyntactic analysis of isiZulu by generating syntax trees through parsing, which involves a deep linguistic analysis of the text. While this process can be slow, it exposes lexical categories (e.g., tense, polarity, and agreement markers) that can be systematically manipulated to generate synthetic data, a process referred to as data augmentation.

Additionally, the ZRG supports a linearisation process, which converts syntax trees into human-readable text. During linearisation, the grammar employs a binding token, facilitating the construction of surface forms and allowing for the specification of different segmentation granularities. This method allows for flexibility regarding segmentation strategies, enabling the exploration of optimal strategies based on linguistic approaches to isiZulu subword modelling.

It is essential to note that ZRG, being a rule-based system, suffers from the aforementioned limitations of rule-based approaches, such as being less robust when encountering words with the defined lexicon, hence failing to parse it and generate its syntax tree. It is also very slow, making it a less preferred strategy to be used for downstream tasks, where speed and accuracy are important factors. Moreover, ZRG was not originally implemented as a surface segmentation tool.

1.3 Problem Statement

To enable the interplay between rule-based and machine-learning approaches in investigating the morphological surface segmentation of isiZulu, this work leverages the strengths of the ZRG. The ZRG provides broad coverage of isiZulu, supports flexible segmentation styles, and enables the generation of surface-segmented data at different granularities in a linguistically informed manner. This approach eliminates the need for extensive system engineering or expert curation, which can be costly and time-consuming. However, recognising the limitations of rule-based systems, such as their lack of robustness and slower

processing speeds, this study proposes training a supervised machine learning model using the ZRG-generated annotated data. Supervised models are well-suited for generalising to unseen words, offering greater robustness and efficiency in segmentation, hence serving better as a pipeline to be used for investigating morphological segmentation in downstream tasks where fast and accurate segmentation is crucial.

1.4 Research Questions, Objectives and Hypotheses

The discussion in the previous section leads to the two research questions to be addressed in this study.

1.4.1 Research Questions

This research seeks to address the following questions:

1. How can a hybrid approach to supervised surface segmentation be pursued that leverages the strengths of a rule-based approach to overcome the weaknesses of a machine learning approach in the context of isiZulu?
2. How can such a hybrid approach be used to obtain insights into supervised surface segmentation strategies for isiZulu?

1.4.2 Research Objectives

In order to answer the research questions, the following primary objective will be pursued.

To investigate supervised surface segmentation of isiZulu text using synthetic data generation

In order to reach this primary objective, the following sub-objectives have been identified.

- *Sub-objective 1:* To utilise the ZRG to parse a large corpus of isiZulu text to obtain a treebank.
- *Sub-objective 2:* To utilise the ZRG to augment the treebank to obtain additional synthetic data.
- *Sub-objective 3:* To investigate the suitability of different machine learning models for the supervised segmentation task.

- *Sub-objective 4*: To train segmentation models on different linearisations of the treebanks.
- *Sub-objective 5*: To evaluate segmentation models intrinsically and extrinsically according to suitable metrics.

1.5 Research Methodology

This study employs the hypothetico-deductive methodology to investigate the morphological surface segmentation of the isiZulu text. Central to this methodology is the use of a rule-based GF grammar, ZRG, to generate a large corpus of surface-segmented isiZulu text across different granularities. This ensures linguistically informed and accurate segmentation.

To further mitigate the matter of data sparsity, data augmentation techniques are applied to enhance the diversity of the generated corpus. These datasets are used to train supervised segmentation models, which are subsequently evaluated both intrinsically (using metrics such as F1-score, BLEU, and chrF) and extrinsically through their impact on neural machine translation (NMT) performance.

The approach integrates comprehensive data preparation, segmentation modelling, and systematic evaluation to address linguistic and technological challenges inherent to low-resource, morphologically complex languages like isiZulu. By employing this hypothesis-driven methodology, the research ensures a structured and systematic exploration of the role of morphological segmentation in enhancing downstream NLP tasks such as NMT. Detailed discussions of each phase, including corpus generation, segmentation modelling, and evaluation strategies, are presented in subsequent chapters. The hypotheses governing this method are outlined below.

1.5.1 Hypotheses

In line with the hypothetico-deductive methodology, the following hypotheses are investigated.

1. In developing a hybrid approach to supervised surface segmentation of isiZulu, using a machine learning approach as a foundation will ensure robustness and efficiency, while a combination of synthetically generated and automatically annotated data would address the requirement of machine learning approaches for large amounts of training data.

These sub-hypotheses have been developed:

- (a) The hybrid approach will exceed the rule-based approach in robustness;
 - (b) The hybrid approach will exceed the rule-based approach in efficiency; and
 - (c) The hybrid approach will reach a sufficient¹ level of accuracy in comparison to the rule-based approach.
2. The hybrid approach will result in one or more segmenters that improve performance in a downstream task such as machine translation.

1.6 Ethical Considerations

This study adheres to ethical principles to ensure responsible research practices in the development and evaluation of morphological segmentation. Ethical concerns are particularly relevant during data collection, processing, and model evaluation, given the potential social and linguistic implications of working with a low-resource language.

IsiZulu, like many other low-resource languages, lacks large publicly available datasets that are both diverse and high quality. This poses ethical challenges during data collection, as the scarcity of digital data resources increases the risk of bias, inaccuracies, and misrepresentations that could negatively impact the digital development of the language.

To mitigate these risks, careful and ethically responsible data collection is essential to ensure fair representation, linguistic integrity, and minimal bias in the development of computational models.

1.6.1 Data Collection and Use

The data used in this study was sourced from publicly available corpora and other commonly used datasets in NLP research. However, these datasets may have been originally collected through web scraping or other automated means, making them susceptible to biases present on the internet or other sources. Consequently, it is possible that inherent biases from the original data sources could be reflected in the dataset used for training the models in this study.

In addition to that, special care has been taken to ensure that the collection and processing of isiZulu text respect copyright regulations and do not violate any intellectual property rights. The study follows best practices for dataset attribution, and proper citations are provided for all corpora used.

¹The notion of sufficiency is discussed in detail in Chapter 7.

1.6.2 Privacy and Cultural Sensitivity

Since the datasets primarily consist of linguistic data, they contain minimal personally identifiable information (PII), making privacy risks relatively low. However, ethical considerations go beyond privacy concerns and extend to the linguistic and cultural sensitivity associated with working with indigenous languages such as isiZulu.

This research does not seek to alter the natural structure of isiZulu, but rather aims to enhance its digital representation in NLP applications. This is made possible through the ZRG, which has been developed with a focus on linguistic transparency and integrity (Marais & Pretorius, 2023).

1.6.3 Transparency and Reproducibility

To promote open science and ensure that the findings contribute to the broader NLP research community, this study adheres to transparency and reproducibility standards by:

1. Clearly documenting data sources, pre-processing steps, and model training configurations. All related resources will be made publicly available and accessible in the GitHub repository².
2. Providing implementation details to enable reproducibility.
3. Sharing insights into the strengths and limitations of the proposed methods.

By addressing these ethical considerations, this study aims to contribute responsibly to NLP research while upholding principles of fairness, linguistic inclusivity, and social responsibility. Additionally, this research proposal was submitted for ethical approval to North-West University, where it was reviewed and approved by the relevant Research Ethics Committee. The assigned ethics approval number is NWU-01328-23-A9.

1.7 Dissertation Structure

This study is made up of the following chapters:

Chapter 1 – Introduction: This chapter provides an overview of the research, including the problem statement and the research objectives, and summarises the research methodology. It sets out the research context and outlines the thesis’s structure.

²<https://github.com/Sthesha/supervised-surface-segmentation>

Chapter 2 – Research Methodology: This chapter outlines the hypothetico-deductive methodology guiding this study. It establishes the research paradigms and provides a structured methodology for the investigation of morphological surface segmentation for isiZulu texts using supervised machine learning approaches to gain insight into different segmentation strategies.

Chapter 3 – Morphological Segmentation: This chapter provides an overview of morphological segmentation, covering issues pertaining to linguistic background, rule-based approaches with examples, machine learning techniques, as well as metrics and benchmarking used for evaluating segmentation.

Chapter 4 – Data Preparation: This chapter outlines the process of preparing the data for training the supervised surface segmentation models. It includes the generation of surface segmented data using rule-based GF grammar, data augmentation techniques, and the creation of a representative dataset for isiZulu.

Chapter 5– Models Design: This chapter outlines the design and development of the supervised surface segmentation models, detailing the selection of learning algorithms, model architecture, parameter optimisation, and the training process. It also provides the rationale for design choices and their alignment with the research objectives.

Chapter 6 – Experimental Evaluation: This chapter presents an evaluation of the developed supervised segmentation models, focusing on both intrinsic metrics (e.g., F1-score, BLEU, chrF) to assess segmentation quality and extrinsic metrics to evaluate their impact on downstream NMT performance. It includes comparative analyses across segmentation granularities and a discussion of the results in the context of addressing data sparsity in isiZulu.

Chapter 7 - Findings, Conclusion and Future Research: This chapter provides a concise summary of the preceding chapters, the results of the formulated hypotheses, and based on the key findings, it addresses the outlined research questions. In addition, it reflects on the challenges and limitations that have been identified and suggests future research directions. Finally, it provides the study’s overall conclusions.

1.8 Chapter Summary

This chapter provided an introduction, background, and overview of the study. These were followed by its research objectives, focusing on addressing the challenges posed by data sparsity and morphological complexity in isiZulu, a low-resource and morphologically rich language. Through a systematic exploration of the research aims, the study emphasises the potential of supervised surface segmentation models in enhancing the performance of NLP applications.

The objectives provide a clear roadmap for the study, beginning with the generation of a large, linguistically valid, and representative corpus of surface-segmented isiZulu text using a rule-based grammar, known as ZRG. The study also proposed the integration of data augmentation techniques to address data sparsity and the identification of the most effective supervised learning algorithms for this investigation which also underline the study's focus on the methodological rigour approach. This methodology is grounded in the hypothetico-deductive approach, which formulates several hypotheses that are aligned with the study objectives and tests them accordingly.

Finally, the inclusion of an extrinsic evaluation objective, where the impact of segmentation granularity is measured through NMT performance, emphasises the practical relevance and broader implications of this research. By addressing these objectives, this study aims to contribute meaningfully to the field of NLP for low-resource languages, setting the stage for future advancements in morphological segmentation and its applications.

Chapter 2

Research Methodology

2.1 Introduction

The present chapter sets out to present a methodology for developing a supervised surface segmentation model for isiZulu using synthetic data generated using the ZRG and evaluating its effectiveness in addressing the data sparsity issue in a downstream task, specifically a machine translation system. Building upon the foundational concepts introduced in the previous chapter, this chapter elaborates on the systematic approach adopted to achieve the research objectives.

This chapter explores the research philosophies and paradigms that inform the study and outlines the specific methodological framework implemented. The focus of this study is to investigate the surface segmentation of isiZulu text with different levels of granularity using a supervised machine learning approach developed using synthetic data. By leveraging the *hypothetico-deductive method* as the guiding approach, this study enables hypothesis formulation and empirical testing, ensuring that the chosen methodologies are well suited to tackle the challenges of agglutinative languages such as isiZulu.

By embedding the research within established philosophical and paradigmatic frameworks, this chapter not only establishes a foundation for a methodological approach, but also emphasises the significance of morphological segmentation in terms of addressing data sparsity for low-resource languages with complex morphologies, such as isiZulu, in the context of language modelling.

This chapter is structured as follows: Section 2.2 provides a foundational perusal of research as a concept, leading to Section 2.3 and Section 2.4, which discuss different research paradigms and philosophies. These are presented as assumptions and beliefs that guide a researcher’s approach to conducting a study. Section 2.5 positions the present study within the introduced assumptions and beliefs, which also influence the chosen method-

ology, namely the *hypothetico-deductive methodology*, which is discussed in Section 2.6. Finally, Section 2.7 concludes the chapter with a summary.

2.2 The Concept of Research

The term “research” has been used interchangeably in the literature, encompassing various interpretations. According to the Cambridge Dictionary (n.d.), research is defined as, “a detailed study of a subject, especially in order to discover (new) information or reach a (new) understanding”. Khaldi (2017) describes research as the systematic acquisition of knowledge through meticulous and structured investigation. Similarly, Dane (1990) emphasises research as a process aimed at asking and answering questions about the world.

As a process, research enables the adoption, refinement, or rejection of certain knowledge based on evidence and analysis. The aims of the research are extensive, including exploring new phenomena, explaining causes and relationships, evaluating interventions, predicting outcomes, understanding complex systems, and developing practical solutions to real-world problems (Collis & Hussey, 2014; Garg, 2016). In addition, research plays a crucial role in contributing towards theory development and testing, providing a foundation for scientific progress and application (Whetten, 1989).

According to Goundar (2012), three key conditions must be met when undertaking research. Firstly, the research process must be governed by clearly defined methodologies, such as qualitative or quantitative approaches, and influenced by the standards and practices inherent to the researcher’s discipline. Secondly, the researcher should employ tools, methods, and techniques that have been rigorously tested for reliability and validity. Reliability pertains to the consistency and repeatability of the chosen methods, while validity ensures that these methods are chosen correctly and accurately measure the intended phenomena. Lastly, the researcher must conduct the research objectively and without bias, thus striving to eliminate bias or vested interests that may compromise the impartiality and credibility of the outcomes. Throughout the research process, conclusions should be reached based on the evidence gathered, avoiding the incorporation of personal biases or interests unrelated to the study’s objectives (Goundar, 2012).

Having established the foundational understanding of research as a systematic process for acquiring and refining knowledge, it is essential to explore the broader philosophical and paradigmatic frameworks that underpin and guide research practices. These frameworks, including the research paradigms and research philosophies, serve as the theoretical lenses through which researchers view the world, formulate questions, and select appropriate methods. By exploring these paradigms and philosophies, one can better understand how they influence the design, execution, and interpretation of this study.

2.3 Research Paradigms

It is salient to consider the concept of research paradigms, because these guide the acquisition of knowledge and scientific discoveries through their underlying principles and assumptions (Park *et al.*, 2020). The term “paradigm”, often referred to as a worldview, encompasses the fundamental collection of beliefs, theories, philosophical assumptions, and ideas that researchers hold. These paradigms shape the design and execution of their research (Mafuwane *et al.*, 2011). A research paradigm provides a lens through which researchers interpret the world and evaluate the methodological aspects of their study, ultimately influencing the choices for data collection and analysis strategies.

Conducting scientific research requires careful consideration of the research paradigm, as this paradigm establishes specific assumptions about how the world operates. Indeed, Myers (2002) emphasises the importance of adopting an appropriate research paradigm to ensure validity, thus allowing researchers to build on their philosophical stance. Furthermore, Clavier *et al.* (2012) highlight that paradigms serve as critical reference points for others to understand the researcher’s underlying assumptions, which have a significant bearing on the study’s design and outcomes. These paradigms are commonly categorised into four key domains: ontological, epistemological, axiological, and methodological (Creswell & Creswell, 2017).

2.3.1 Ontological Assumptions

Ontological assumptions refer to the nature of reality, whether it is regarded as objective, subjective, or socially constructed, and encompass the aspects to be investigated or encountered during the research process (Alele & Malau-Aduli, 2023). According to Guba and Lincoln (1994:108), ontological assumptions aim to address the question: What is the form and nature of reality, and what can be known about it? Simply put, these assumptions focus on understanding the fundamental nature of the phenomena being researched. The nature of reality implies different ontological perspectives, which, in turn, influence the research approach. Assumptions about reality are often categorised along a spectrum, ranging from an objective reality that exists independently of human perception to a subjective reality that is shaped by individual experiences and interpretations (Ahmed, 2008). Ontological assumptions play a critical role in structuring a researcher’s thinking about the topic under investigation. Such assumptions are salient in guiding the formulation of research questions and determining how those questions are addressed (Kivunja & Kuyini, 2017).

2.3.2 Epistemological Assumptions

Epistemology is a branch of philosophy that focusses on the study of knowledge and beliefs, describing how knowledge about reality is acquired, conceptualised, and applied (Hatch, 2018). According to Guba and Lincoln (1994:108), this assumption addresses the question: What is the nature of the relationship between the knower or would-be knower and what can be known? The response to this question is often influenced by the ontological assumptions underpinning the research.

Epistemology seeks to explore questions such as: How is knowledge produced? What standards distinguish good knowledge from bad knowledge? And how should reality be defined or represented (Hatch, 2018)? Epistemology also pertains to how one can communicate this knowledge to others (Burrell & Morgan, 2019).

2.3.3 Axiological Assumptions

Axiology pertains to values and ethics, reflecting the role of a researcher's personal values and biases in the research process, including actions taken after the research is completed (Saunders, 2009). It influences the entire research process, and is critical for ensuring the credibility and integrity of the study. As noted in Pretorius (2024), axiology encourages researchers to reflect on how their own values, beliefs, and biases shape the design, execution, and interpretation of their studies.

Furthermore, axiology assumptions are interconnected with ontological and epistemological assumptions. Understanding the nature of reality (ontology) enables researchers to assess the truth value of knowledge, while determining the knowability of this reality (epistemology) informs the science of truth. Once these foundations have been established, defining the correct science of values (axiology) becomes more straightforward (Engle, 2008).

2.3.4 Methodology

This dimension of the research undertaking is concerned with the general strategy or action plan that guides the choice and use of specific methods in the context of a particular research paradigm (Wahyuni, 2012). It refers to a system of *methods*, procedures, or principles employed to achieve specific objectives. In research, methodology can be defined as systematic procedures or strategies that researchers use to describe, explain, and predict phenomena or to carry out a research project (Aguiar, 2024). According to Guba and Lincoln (1994), a methodology seeks to determine the steps that an inquirer must take to uncover what they believe can be known.

A research methodology helps to clarify several aspects of a research project, such as why it was conducted, how and why the hypothesis was formulated, and the methods or techniques chosen to address the research problem. This includes details about the data used, how it was collected, and related questions. A closely related concept and often used interchangeably is *research methods*, which refer to the specific procedures employed in conducting research (Howell, 2012). As Goundar (2012:45) aptly state, research methodology, “refers to more than a simple set of methods; rather it refers to the rationale and the philosophical assumptions that underlie a particular study relative to the scientific method”, while research methods constitute the execution phase of both scientific and non-scientific research.

Similarly to other paradigmatic assumptions, the specific methodological choices that a researcher adopts are influenced by their responses to ontological, epistemological, and axiological questions. This interconnectedness ensures that all these paradigms work together, enabling researchers to conduct systematic investigations that produce meaningful and reliable knowledge.

In terms of methodologies, there are various methodologies that can be employed in different types of research, and the term is generally considered to include research design, data collection, and data analysis (Goundar, 2012). Research methodologies are commonly categorised into two main types: quantitative and qualitative methodologies (Onwuegbuzie & Leech, 2005). These are briefly discussed below.

1. The quantitative methodology emphasises numerical attributes, relying on objective measurements and statistical analysis to explore the relationships between variables and describe the causes of change (Kornuta & Germaine, 2019). Quantitative research involves gathering numerical (quantitative) data, which is systematically analysed using statistical methods to address specific research questions or hypotheses. Researchers who use quantitative methods often propose hypotheses, collect numerical data, and use statistical evidence to support or refute these hypotheses, enabling broader generalisations based on a scientific approach (Rana *et al.*, 2023).
2. In contrast to quantitative methods, qualitative research methodologies utilise non-numerical data, such as observations, textual analysis, and interviews, to answer open-ended questions such as “how” and “why”, which makes it ideal for conducting research in investigating non-linear phenomena such as experiences, perspectives, and behaviours that can be too complex to be captured using quantitative methods (Goundar, 2012; Tenny *et al.*, 2017). Rather than relying on statistical analyses or pre-formulated hypotheses, qualitative studies often involve open-ended inquiry, allowing hypotheses or theories to emerge naturally during the research process (Kornuta & Germaine, 2019). The researcher plays an active role in interpreting data, relying heavily on subjective insights and descriptive observations. While qualita-

tive approaches inherently involve subjective interpretations by the researcher, these interpretations are grounded in rigorous and structured analysis of collected data, rather than mere personal opinions or ungrounded assumptions.

Understanding and reflecting upon these beliefs is essential, as they underpin the methodological choices and strategies employed in any research project. By clarifying these perspectives, researchers can ensure that their study aligns with its philosophical foundations, enhancing its rigour and relevance. Again, these philosophical assumptions are part of the research philosophy, which includes various possibilities.

2.4 Research Philosophy

The term *research philosophy* is an evolving concept that lacks a unified definition among scholars and is often used interchangeably with research paradigms. According to Žukauskas *et al.* (2018), a research philosophy represents the system of thought that a researcher adopts to produce new and reliable knowledge about their research object. Similarly, Saunders (2009) describes it as the development of knowledge and its nature, emphasising that it embodies critical assumptions about a researcher's perspective of the world, which, in turn, influences their research strategy and methods.

This study focuses on a subset of the research philosophies most widely discussed in the social sciences: positivism, interpretivism, pragmatism, and critical realism, as highlighted by Žukauskas *et al.* (2018) and Ryan (2018). These were selected due to their relevance to the study's objectives and their widespread application in social science research

2.4.1 Positivism

The positivist philosophy asserts that the social world can be studied and understood objectively (Žukauskas *et al.*, 2018), much like the natural world. This philosophical stance emphasises the use of scientific methods to produce credible data and facts, which are considered to be independent of human interpretation or bias. Positivist studies primarily rely on quantitative data, utilising experiments and established theories to formulate hypotheses that can be rigorously tested, confirmed, or rejected (Saunders, 2009).

A key characteristic of positivist research is its focus on objectivity and detachment, aligning with its axiological assumption. Researchers who adhere to this philosophy strive to minimise personal values or subjective influences that could impact research outcomes (Irshaidat, 2022). The ontological assumption upheld by positivism asserts the existence of a rigid, objective reality which can be understood through appropriate instruments, data collection, and empirical analysis. Epistemologically, positivism assumes that knowledge

about this reality is acquired through rigorous observational strategies, ensuring that it remains objective and unaltered (Ringberg & Reihlen, 2008).

By emphasising empirical evidence and measurable phenomena, positivist approaches seek to arrive at conclusions that are replicable and universally valid. This makes it particularly suitable for research that involves systematic observation, controlled experimentation, and hypothesis-driven inquiries. In this study, a positivist approach is adopted to validate claims about the performance of supervised machine learning approaches, when these are trained on a combination of synthetically generated and automatically annotated data, to ascertain whether they enhance the robustness and efficiency of morphological segmentation models for isiZulu by addressing the data sparsity challenge inherent in low-resource, morphologically rich languages.

Having established the foundational understanding of research as a systematic process of acquiring and refining knowledge, it is essential to look at the broader philosophical and paradigmatic frameworks that underpin and guide research practices (Alele & Malau-Aduli, 2023). These frameworks, encompassing research paradigms and philosophies, serve as the theoretical lenses through which researchers view the world, formulate questions, and select appropriate methods.

2.4.2 Interpretivism

Another prominent research philosophy is interpretivism or constructivism, which emerged as a critique of positivism (Yong *et al.*, 2021). Interpretivism posits that the social world can be understood in multiple ways by different individuals, highlighting the crucial role of researchers in observing and interpreting this world (Ali, 2023). It is grounded in constructivist ontology, which asserts that reality is not discovered but constructed through human interactions and social processes (Goldkuhl, 2012).

Researchers conducting interpretivist studies are often required to adopt an empathetic viewpoint, allowing them to better understand the perspectives of the research subjects. This reflects a fundamental epistemological assumption of interpretivism, namely that knowledge is subjective and arises from interactions between the researcher and the participants (Saunders, 2009). The associated axiological perspective acknowledges the researcher’s participation in the data generation process. However, it also emphasises the importance of consciously listening to the participants and evaluating their input without contaminating it with personal biases or values, ensuring the reliability of the findings (Luyt *et al.*, 2012).

This approach is typically used to study people and their social interactions, contrasting with positivist approaches that often focus on natural sciences and non-human objects such as computers (Saunders, 2009). However, it also emphasises the importance of consciously

listening to the participants and evaluating their input without contaminating it with personal biases or values, ensuring the reliability of the findings.

2.4.3 Pragmatism

The two approaches discussed above, positivism and interpretivism, represent opposite ends of the research philosophy spectrum, each with clear advantages and limitations. However, many research studies do not fall entirely into one of these categories, as they often involve exploring both rigid and subjective realities. This calls for a more flexible approach that combines elements from multiple philosophies, where pragmatism as a research philosophy comes in.

Pragmatism asserts the possibility of working with diverse assumptions, including ontological and epistemological perspectives from both positivist and interpretivist stances (Saunders, 2009). As noted by Alghamdi and Li (2013), pragmatism is not restricted to any one system of philosophy or reality. Instead, it allows researchers the flexibility to select methods and techniques most suitable for addressing their research questions, whether these are quantitative or qualitative, and regardless of the type of data being analysed.

This philosophical stance emphasises the practical outcomes of research and supports the use of mixed-methods approaches, combining the rigour of scientific inquiry with the depth of interpretive understanding. Pragmatism, therefore, serves as a valuable framework for studies that require methodological diversity and adaptability to address complex research problems effectively.

2.4.4 Critical Social Theory

Critical Social Theory (CST) refers to a broad range of theoretical approaches that critique existing social conditions and power structures with the goal of emancipating humans and the planet from the hostile consequences of modernity (Celikates & Flynn, 2023; Manners, 2020). The CST is rooted in the Frankfurt School's "Critical Theory" that emerged in the 1930s as interdisciplinary research approach that combines philosophy and social science with the emancipatory objectives of various social and political movements. The CST examines how different dimensions of domination and oppression, such as economic, racial, gendered, and political, shape society and seeks not just to understand these dynamics but to challenge and change them (Browne, 2000).

The ontological assumption of CST can be described as historical realism or critical realism, however, in a different sense from what is presented by Bhaskar (2013). This means that CST does believe in a reality (particularly social reality) that is "real" and has objec-

tive consequences; this “reality” is not fixed or immutable; it is shaped by historical and social forces, especially by relations of power and oppression (Scotland, 2012). Guba and Lincoln (1994:109) characterise the ontology of this research philosophy as “historical realism – a reality that is shaped by social, political, cultural, economic, ethnic, and gender values, crystallised over time”. In simpler terms, what we take to be “reality” (especially in the social world) is the product of history and power dynamics – for example, racial categories or gender roles are real in their consequences, but are historically constructed rather than being natural phenomena.

In terms of epistemological assumptions, CST is described as transactional or subjectivist, meaning that the inquirer and the object of inquiry are not independent, but interactively linked. As a result, the values of the investigator inevitably influence the findings (Guba & Lincoln, 1994). Within this framework, the researcher’s values serve as primary resources driving the research philosophy. Rather than attempting to remain a neutral observer, the researcher explicitly uses their values to uncover and interpret social truths. According to Cohen (as cited in Scotland, 2012:35), “what counts as knowledge is determined by the social and positional power of advocates of that knowledge”.

Concerning axiological assumptions, CST places values at its core, explicitly addressing normative questions such as “What is intrinsically worthwhile?” (Scotland, 2012:13). This inherently normative orientation emphasises respect for cultural norms and societal values (Kivunja & Kuyini, 2017; Scotland, 2012). Central to CST are values of emancipation and social justice, which form the foundational components of its research paradigm.

Methodologically, CST favours approaches that promote critical insight, active participation, and social change. Since the ultimate goal of CST is not merely to study society but to transform it, the methodology employed is often described as dialogic (Kivunja & Kuyini, 2017). Ideology critique represents a core methodological practice, enabling critical analysis and interrogation of underlying values and assumptions to expose injustice (Scotland, 2012). Furthermore, CST frequently incorporates action research as a means of actively seeking to change social realities (Dieronitou, 2014).

2.4.5 Critical Realism

Critical realism (CR) is a relatively recent and comprehensive philosophy of science developed by the English philosopher Roy Bhaskar as an alternative to the prevailing philosophical paradigms of positivism, interpretivism, and pragmatism (Lawani, 2021). CR comprises two primary components: transcendental realism and critical naturalism, addressing the philosophy of science and social science, respectively (Bhaskar, 2013; Bhaskar, 2014). Transcendental realism is focused on the ontology of the natural world, asserting that real mechanisms exist beyond empirical observation and hermeneutic interpretation.

Critical naturalism extends this realist ontology to the social realm, suggesting that underlying social structures and mechanisms also exist beyond immediate observation and interpretation (Bhaskar, 2013; Zhang, 2023). Ultimately, critical realism asserts that an objective reality exists (as in positivism), but our knowledge of this reality is inherently partial and shaped by theoretical frameworks. Consequently, critical realism emphasises the necessity of exploring deeper structures and causal mechanisms that operate or exist beyond direct empirical observation.

Similarly to pragmatism, CR advocates for mixed-methods research. However, unlike pragmatism, which views ontological and epistemological assumptions as separable from research methods, CR explicitly associates these methods with its philosophical positions. Specifically, CR maintains a layered ontology consisting of three distinct domains: the *real*, the *actual*, and the *empirical* (Holmén, 2020). The *real* domain refers to the deep structures, properties, and mechanisms of entities that exist independently of observation. The *actual* domain encompasses the actual events generated by these underlying mechanisms, irrespective of whether they are observed. Finally, the *empirical* domain represents the observable experiences or events perceived by individuals. CR thus rejects simplistic reductions of reality to only what can be directly observed. Instead, it asserts that many critical aspects of social reality, such as social structures and class relations, are real but not directly observable, requiring inference based on their observable effects.

Methodologically, CR advocates for a pluralistic approach to research, employing any methods that effectively identify and explain causal mechanisms. Given that CR ontology is complex, with multiple layers of reality, both quantitative and qualitative methods can be combined to capture different aspects of a phenomenon (Sobh & Perry, 2006). This methodological flexibility allows researchers to investigate not only observable patterns, but also the underlying structures and causal relationships that shape social and natural phenomena.

2.5 Positioning the Present Study

This study is firmly grounded in a positivist philosophical framework, which emphasises objectivity, empirical evidence, and hypothesis-driven inquiry. The ontological assumption underlying this research is that an objective reality exists, wherein morphological patterns (the structure of words and how morphemes combine) in isiZulu can be systematically measured and analysed. This perspective aligns with the positivist view that reality is independent of human perception and can be understood through structured observation and experimentation.

Epistemologically, the study adopts a stance that valid and reliable knowledge or insights about the relationship between segmentation granularity and downstream task perfor-

mance can be acquired through empirical investigation. By leveraging comprehensive and repeatable experimentation, the research seeks to uncover insights around the supervised morphological segmentation of isiZulu texts using a rule-based approach as a pipeline to generate the syntactic data of different granularity. This approach ensures that findings are grounded in measurable evidence rather than subjective interpretation.

Axiologically, the study reflects the positivist emphasis on objectivity and detachment. Personal values and biases are consciously minimised to maintain the integrity of the research process and outcomes. Ethical considerations are central, with transparency in data selection, methodological soundness, and the reproducibility of results being consistently prioritised in order to uphold the study’s credibility.

2.6 Hypothetico-Deductive Methodology

Since this study assumes the positivist paradigm approach, the most common method for such in scientific research is the hypothetico-deductive methodology. This is an approach that is regarded to be the heart of scientific inquiry (Fosl & Baggini, 2020). This methodology comprises two primary components: the *hypothetico* part, where an explanatory hypothesis is formulated to address the research problem, and the *deductive* part, where testable claims are derived from the hypothesis and empirically examined. The results after the examination can be a substantiated or falsified hypothesis (Park *et al.*, 2020). According to Fosl and Baggini (2020:47) the principle governing this procedure is to, “start with a hypothesis and a set of given conditions, deduce what facts follow from them and then conduct experiments to see if those facts hold and hence whether the hypothesis is false”. Sekaran and Bougie (2016) and Tariq (2015) present this methodology as a seven-step process, as detailed below.

1. *Identify a broad problem area*

The first step in the research process is to identify a general area of interest or concern that serves as the foundation for the research project. This step involves scanning existing literature, identifying gaps, and noting practical challenges in the field.

2. *Define the problem statement*

To conduct a scientific research study, one needs to have a definite aim or purpose (Sekaran & Bougie, 2016). In this step, a clear and concise problem statement should be articulated, which highlights the issue that the study seeks to address and provides the foundation for the research objectives and questions. Preliminary information is gathered to understand the factors contributing to the problem, narrowing the broad area into a specific problem statement.

3. *Develop hypotheses*

Based on observations and existing theories, hypotheses are developed to guide the study. These serve as predictions to be examined through empirical investigation (Popper, 2002). In a scientific study, the conducted hypotheses must meet two key criteria:

- (a) *Testability*: The hypothesis must be capable of being empirically tested.
- (b) *Falsifiable*: The hypothesis should be disprovable, because hypotheses cannot be confirmed, but can only be corroborated until contradicted by new findings.

4. *Determine measures*

This step involves identifying and selecting appropriate tools, instruments, or methods to *operationalise* (quantify or observe) the variables defined in the research hypothesis. Trochim *et al.* (2016) note that in this process, *validity* and *reliability* are critical considerations to ensure that the measures capture the constructs being studied effectively.

5. *Data collection*

Once the measures have been established, researchers proceed with gathering data using suitable methodologies such as surveys, experiments, interviews, or observations. Data collection should follow ethical guidelines and adhere to a well-designed protocol to ensure consistency and minimise bias (Fowler, 2014).

6. *Data analysis*

After collecting the data, researchers analyse it using numerical analysis such as statistical approaches or qualitative techniques appropriate to the nature of the study. This step involves examining patterns, testing hypotheses, and drawing conclusions based on the data's evidence (Sekaran & Bougie, 2016).

7. *Interpretation of data*

The final step in this research process involves interpreting the results of the analysis in the context of the original hypotheses and the broader problem area. The findings provide insights into whether the hypothesis is supported or not. At this stage the researcher has to discuss the implications, address potential limitations, and provide recommendations for future research. In the event that the hypothesis was not supported, the researcher should critically evaluate the reasons behind the found outcome and refine the theory for retesting (Sekaran & Bougie, 2016).

2.6.1 Application in this Research

The hypothetico-deductive methodology serves as the guiding framework for this research, where multiple hypotheses are presented that align with the main objective of the study, which is: *To investigate supervised surface segmentation of isiZulu text using synthetic data generation*. This is achieved through following five sub-objectives presented in Chapter 1 Section 1.4.2.

The following section outlines how each step of the hypothetico-deductive approach is applied in this study:

1. Identifying a broad problem area

As discussed in Chapter 1, isiZulu’s complex agglutinative morphology allows for a single stem to generate numerous word variations, significantly increasing vocabulary size. This vocabulary explosion poses challenges for language modelling, as many word variations may be absent from training data, leading to out-of-vocabulary (OOV) issues where models struggle to recognise or generate unseen tokens effectively.

One potential solution to mitigate data sparsity is morphological segmentation, which reduces vocabulary size by breaking words into their constituent morphemes. Morphological segmentation can be either through canonical or surface segmentation, where the former breaks the words into underlying morphemes, and the latter, which divides them into morpheme-based substrings (Cotterell *et al.*, 2016b). When canonical segmentation is used, it is simpler for one to determine the “correct” segmentation, since the canonical forms of morphemes of the language are well understood from a linguistic point of view. Less can be said about this in a surface segmentation, because the high degree of morphophonological alternation makes it less clear what a suitable segmentation should be and where to place the morpheme boundary, when looking to ensure optimal results of a downstream task.

Both canonical or surface segmentation can be achieved either through rule-based or machine learning-based approaches. Rule-based systems rely on expert-crafted linguistic rules, while machine learning-based methods infer patterns from data, which can be either labelled (supervised) or unlabelled (unsupervised). Studies have shown that supervised approaches generally outperform unsupervised methods in morphological segmentation (Belth, 2024; Ruokolainen *et al.*, 2016; Wang *et al.*, 2016b).

Despite the advantages of machine learning-based segmentation, obtaining high-quality annotated data remains a significant challenge, especially for low-resource languages like isiZulu. The lack of pre-existing labelled datasets makes training supervised models particularly difficult, necessitating alternative data generation.

2. *Defining the problem statement*

The study narrows its focus to the specific problem of data sparsity caused by isiZulu’s morphological complexity. This sparsity negatively impacts the efficiency of language models, particularly in NLP systems, where poor handling of morphological variation results in suboptimal results. Morphological segmentation has the potential to alleviate this issue; however, the problem of annotated data in low-resource languages makes it even harder to gain many insights about morphological segmentation; the problem of data remains a more significant issue in investigating and gaining insight into its effects. In this sense, the present study explores using a hybrid approach to morphological surface segmentation, using a rule-based system, ZRG system, to generate synthetic morphological surface segmented data of different segmentation granularity, to use this data to train supervised machine learning morphological segmenters, and to investigate their performance intrinsically and extrinsically.

3. *Developing hypotheses*

This study is based on two hypotheses:

- (a) In developing a hybrid approach to supervised surface segmentation of isiZulu, using a machine learning approach as a foundation will ensure robustness and efficiency, while a combination of synthetically generated and automatically annotated data would address the requirement of machine learning approaches for large amounts of training data.
 - i. The hybrid approach will exceed the rule-based approach in robustness;
 - ii. The hybrid approach will exceed the rule-based approach in efficiency;
 - iii. The hybrid approach will reach a sufficient level of accuracy in comparison to the rule-based approach.
- (b) The hybrid approach will result in one or more segmenters that improve performance in a downstream task, such as machine translation.

4. *Determining measures*

In determining the appropriate measures for evaluation, this study employs a two-fold approach to assess both the efficiency and effectiveness of the proposed morphological segmentation system. Efficiency is evaluated by examining the ease and speed with which the supervised machine learning model performs segmentation, particularly in comparison to the rule-based approach, which is used as the pipeline for generating data. This involves measuring the processing time of both systems and making a comparison as to which system is the fastest “efficient” in segmentation.

Effectiveness, on the other hand, will be measured in three key ways. Firstly, how robust are the supervised surface segmentation systems in generating new tokens that are not in the vocabulary (tokens that the ZRG could not segment)? The

second and third measures of effectiveness involve evaluating the developed system intrinsically and extrinsically. In the intrinsic evaluation, the system output is directly assessed in terms of predefined standards or criteria that relate to the system’s functionalities or objectives (Jones & Galliers, 1995; Resnik & Lin, 2010). In the surface segmentation context, the intrinsic evaluation assesses how close or similar the system-generated morphs (hypothesis) are to pre-generated (constructed) morphs (ground truth). In text sequential problems, this assessment is usually conducted through quantitative n-grams based metrics such as BLEU (Papineni *et al.*, 2002), CHaRacter-level F-score) (Popović, 2015) and classification metrics such as precision, recall and F1-score. Similarly, this work employs these metrics to evaluate how similar the generated morphs are to the pre-generated morphs.

When conducting extrinsic evaluation, the developed system is treated as an enabling technology and is used to assess its impact on another system, a downstream task (Resnik & Lin, 2010). In this case, the impact of segmentation models on a downstream NLP application, Neural Machine Translation (NMT), is examined. The effectiveness of different segmentation granularities is assessed by evaluating the performance of the NMT system when trained on segmented versus unsegmented data. This comparison provides insight into the extent to which morphological segmentation improves translation quality, as measured by BLEU and chrF scores. Through this comprehensive evaluation framework, the study ensures that both segmentation models and their practical application in the NLP downstream task are well assessed with a view to gain useful insights.

5. Data collection

Two datasets are utilised in this research, which may also fall under the data collection step:

- (a) A dataset for training the morphological segmenters across different levels of granularity.
- (b) A parallel isiZulu-English dataset that is filtered and preprocessed. The isiZulu sentences are segmented using the developed segmenters, with the unsegmented sentences serving as a baseline. This allows for a direct comparison of the effectiveness of the segmentation strategies.

In a broader sense, data collection in this study extends beyond acquiring raw data sets. It involves conducting multiple experiments across different datasets to analyse the impact of morphological segmentation on model performance. This process includes training, validation, and testing, where the data is systematically split to ensure reliable evaluation. Insights are gathered through performance metrics, model robustness, loss trends, and visualisation graphs, which provide a deeper understanding of the model’s learning behaviour, token or morpheme predictions, and overall

system effectiveness. These experimental observations serve as a critical component of data collection, enabling a comprehensive assessment of the developed system.

6. *Data analysis*

Statistical and computational methods are employed to analyse the data that has been collected. This includes intrinsic evaluations being conducted to measure the segmentation models' quality, while extrinsic evaluations assess the impact of segmentation on NMT performance. Comparative analyses are carried out across different granularity levels to identify trends and draw meaningful conclusions.

7. *Interpretation of data*

The results are examined to determine whether the hypothesis is supported. Key findings, such as the relationship between segmentation granularity and their impact on translation quality, are analysed. The implications for future research and applications in low-resource language NLP are discussed, addressing the proposed approach's strengths and limitations.

By applying the hypothetico-deductive methodology, this study ensures a systematic and structured approach to exploring how morphological segmentation can alleviate data sparsity challenges in isiZulu. Each step is aligned with the study's objectives, ensuring that the research generates valid, reliable, and actionable insights. Subsequent chapters will provide in-depth discussions of these steps and their execution within the study's context.

Figure 2.1 summarises the research methodology that allows for the systematic investigation of the research problem and the structured pursuit of the study's objectives.

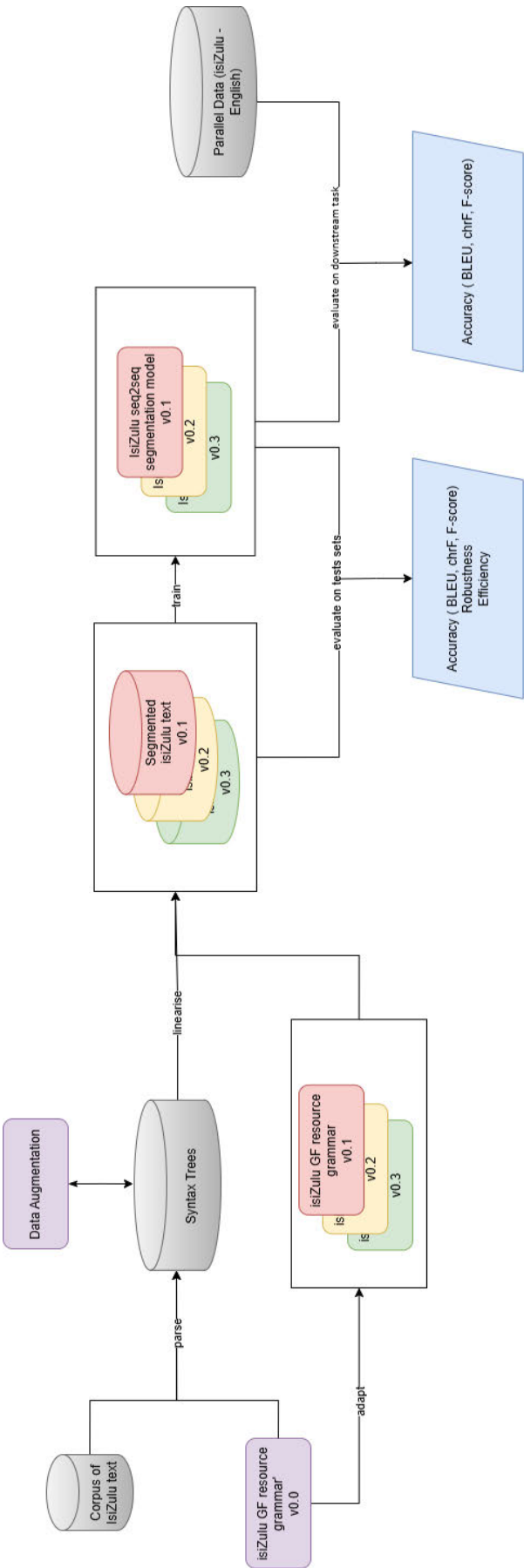


Figure 2.1: Summary of the research methodology.

2.7 Chapter Summary

This chapter outlined the research methodology employed to investigate the possibility of using a hybrid rule-based ZRG and supervised machine learning approaches to conduct morphological surface segmentation of isiZulu text with different granularity levels and gain salient insights through intrinsic and extrinsic evaluation. Grounded in the positivist paradigm, the study adopts the hypothetico-deductive methodology, ensuring a structured and rigorous approach to hypothesis-driven scientific inquiry.

The chapter introduced foundational research philosophies and paradigms, emphasising how ontological, epistemological, axiological, and methodological assumptions inform the research process. These philosophical underpinnings set the stage for the study's methodological choices, particularly its reliance on objectivity, empirical evidence, and systematic observation. The positivist paradigm was identified as the guiding framework, aligning with the study's focus on measurable phenomena and hypothesis testing.

The hypothetico-deductive methodology was presented as the central approach for this research, detailing its steps: identifying a broad problem area, defining a specific problem statement, developing hypotheses, determining measures, collecting and analysing data, and interpreting results. Each step was explicitly contextualised within the study objectives, demonstrating how the methodology supports a comprehensive test of the two formulated hypotheses.

This chapter highlighted key components of methodology, including the selection of intrinsic and extrinsic evaluation metrics, the collection and preparation of datasets, and the use of statistical and computational methods for data analysis. By ensuring alignment between the research questions, methodology, and objectives, the study provides a pipeline to investigate the hypothesis in a systematic and scientifically governed approach and, hence, achieving the primary objective.

In summary, this chapter has provided a comprehensive framework for conducting the research, bridging philosophical considerations with practical methodological choices. Embedding the study within a structured and hypothesis-driven framework helps to ensure the reliability and validity of the findings. Subsequent chapters will build on this foundation, delving into the implementation, experimentation, and evaluation processes, ultimately addressing the broader implications of the research for low-resource languages like isiZulu.

Chapter 3

Morphological Segmentation

3.1 Introduction

This chapter sets out to provide a literature review on morphological segmentation with a number of salient concepts that are associated with it. The review begins with a linguistic background of Nguni languages with a particular focus on isiZulu, covering aspects of phonology, phonetics, and morphology. The chapter then examines different strategies and techniques that are currently employed in the literature to conduct morphological segmentation. This ranges from rule-based approaches to data-hungry approaches such as machine learning and deep learning techniques. The chapter concludes by looking into different metrics that are commonly used to measure the performance of morphological segmentation, and concludes with the trends and open questions that are presented within this domain.

3.2 Background and Linguistic Characteristics of isiZulu

3.2.1 Background and History of isiZulu

This section provides a brief background and describes the linguistic features of isiZulu (and its related languages) that require the kind of segmentation described in this study.

Zulu or isiZulu (as an endonym) is one of the 12 official languages in South Africa, and it is considered to be the most widely spoken indigenous language in the country. IsiZulu speakers comprise approximately a quarter of the population, with around 15.1 million home language speakers out of the population of 62 million people in South Africa (StatsSA, 2022a). The majority of isiZulu native speakers are concentrated in KwaZulu-Natal, where

the language is predominantly spoken in 80% of households, followed by Mpumalanga, where it is the primary language in 27.8% of households, as shown in Figure 3.1. Apart from South Africa, isiZulu is also spoken in other Southern African countries, including Eswatini, Mozambique, and Namibia, although with smaller populations of speakers (Asante & Mazama, 2009).

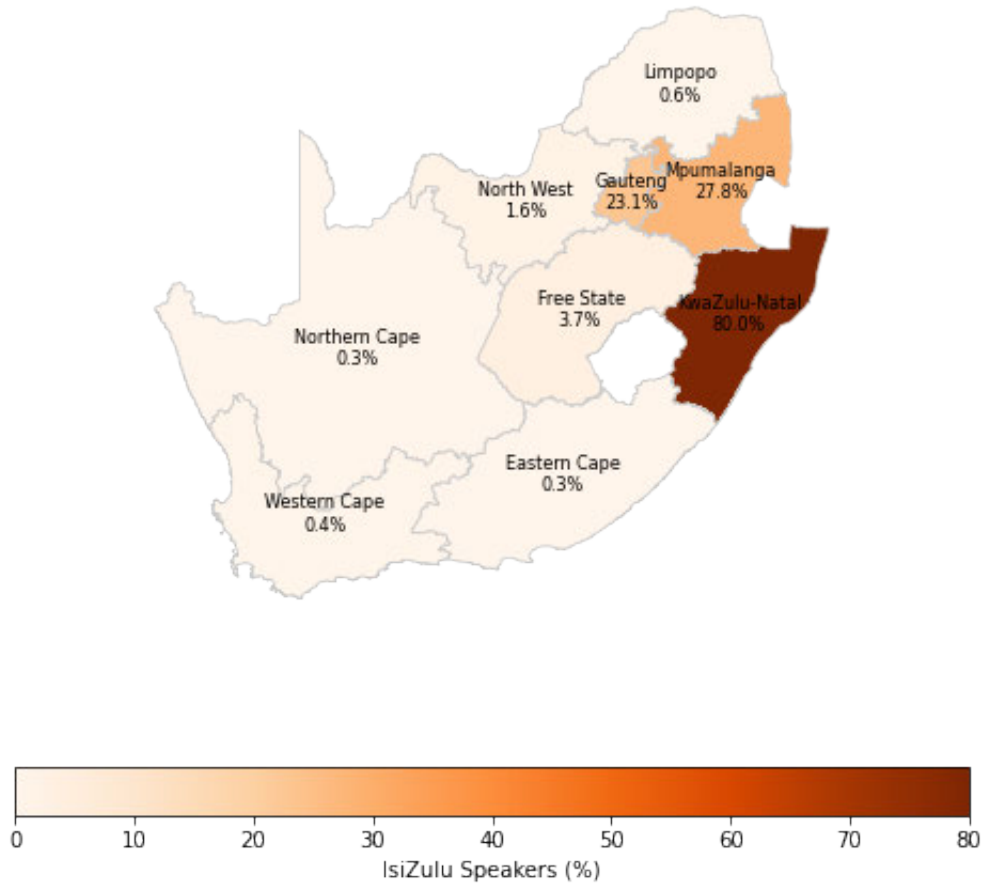


Figure 3.1: Distribution of isiZulu Speakers Across South African Provinces, Based on Statistics 2022 Data

IsiZulu is part of the Nguni language family, which is made up of a group of Bantu languages belonging to a larger Niger-Congo language family. These languages are widely spoken across Southern Africa (Mesham *et al.*, 2021). The Nguni languages, including Siswati, isiNdebele, isiXhosa, and isiZulu, along with others, are classified as the Nguni group (S40) in the Guthrie classification of Bantu languages (Guthrie, 1967). Guthrie’s classification system organises Bantu languages into 15 geographical zones, with the Nguni languages grouped as one of these zones. This classification system provides a framework for understanding both the linguistic characteristics and the geographical distribution of these languages (Flight, 1980). An updated version of this classification was later formulated by Maho (2003), further refining the understanding of the relationships among Bantu languages. The classification structure is summarised in Figure 3.2.

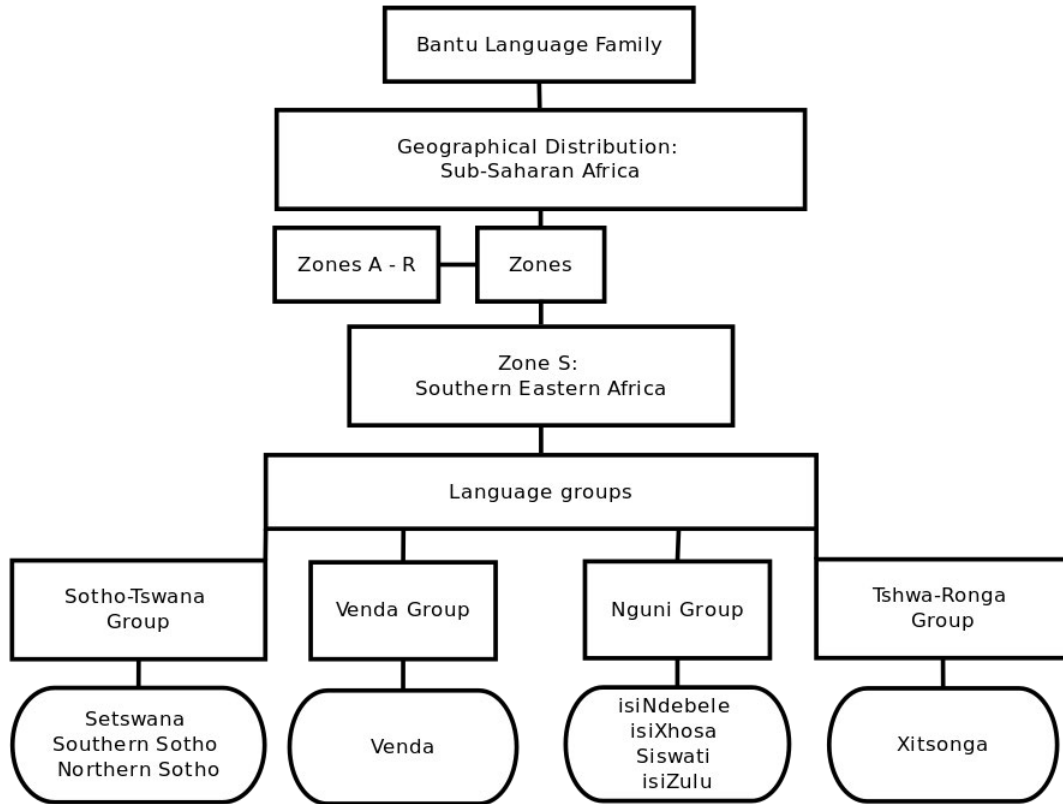


Figure 3.2: Official South African Southern Bantu languages Guthrie group.

In 1995, South Africa’s Constitution recognised four Nguni languages, namely: Siswati, isiNdebele, isiXhosa, and isiZulu, as official languages. However, it is acknowledged that there are other Nguni languages spoken within South Africa’s borders, e.g., Phuthi, Xesibe and Bhaca (Maho, 2003; Zungu, 1999). Following isiZulu, isiXhosa has the second-largest number of home language speakers in South Africa, namely 16.3% of the population. Siswati and isiNdebele have fewer home language speakers in the populations, accounting for 2.8% and 1.7%, respectively, as can be seen in Table 3.1. These languages exhibit significant similarities in their lexical, phonological, and grammatical structures, which render them mutually intelligible to varying degrees (Jacobson, 1986).

Within the Bantu language family, which includes the Nguni languages, Swahili is the most widely spoken language, with an estimated 200 million first and second-language speakers (UNESCO, 2021). Estimations of the total number of Bantu languages vary significantly in the literature. For example, Guthrie (1967) estimates them to be 440, while Bastin estimates 542 (Bastin *et al.*, 1999), while Maho’s update (Maho, 2003) of Guthrie’s classification of the Bantu languages suggests that there are 680 Bantu languages. The Bantu-speaking population (BSP) is estimated at 350 million (Van de Velde *et al.*, 2019).

Table 3.1: Distribution of languages spoken in South Africa.

Language	Speakers (%)
isiZulu	24.40
isiXhosa	16.30
Afrikaans	10.60
English	10.00
Sepedi	8.70
Setswana	8.30
Sesotho	7.80
Xitsonga	4.70
SiSwati	2.80
Tshivenda	2.50
isiNdebele	2.10
Sign Language	0.02
Other	2.10

IsiZulu has been documented for over 180 years, with written records initiated by missionaries in the Natal province, now known as KwaZulu-Natal, in 1830. This pioneering effort led to the publication of the first isiZulu booklet titled *Incwadi Yokuqala* (1837-8), which included the spelling of isiZulu words (UNESCO, 2000). Several works followed this publication, looking into the orthography of isiZulu, such as grammar books (Doke, 1927b; Grout, 1859) and others. These works documented isiZulu words using the Latin alphabet based on phonetic recording, writing the words as they are heard. During this period of growth, significant contributions were made by the Zulu Language Board towards the standardisation of the language; this role was later assumed by the Pan South African Language Board (PanSALB).

3.2.2 Linguistic Characteristics of isiZulu

This section sets out to cover the core linguistic areas of isiZulu, focusing on phonology, phonetics, syntax, semantics, and pragmatics. Since this study primarily focuses on investigating morphological phenomena, several other aspects of isiZulu linguistics - such as discourse analysis, sociolinguistic variation, language acquisition, and historical linguistics - are not covered in this section. For a more comprehensive understanding of phonology and phonetics in isiZulu and related languages, readers can refer to the works of Doke (1927b), Cope (1966), Poulos and Msimang (1998) and Nurse and Philippson (2006), which provide extensive analyses in these areas.

3.2.2.1 Phonetics and Phonology

Phonetics and phonology are two closely related sub-fields within linguistics which focus on the study of human speech sounds. *Phonetics* involves the study of how human speech sounds are produced, classified, and perceived, including their physical and acoustic properties (Davenport & Hannahs, 2020).

On the other hand, *phonology* is concerned with the abstract, systematic organisation of speech sounds within a particular language. It examines the rules and patterns that govern how sounds function in linguistic contexts, such as how phonemes are organised and interact in a given language (Catford, 2001). Central to phonology are *phonemes*, which are the smallest units of sound that are responsible for changing meanings or forms. For example, one can consider the word *ishume*, which is phonemically transcribed as “/iʃu:mi/”. Each phoneme within this word contributes to its unique pronunciation and meaning: /i/ for ‘i’, /ʃ/ for ‘sh’, /u/ for ‘u’, /m/ for ‘m’. In isiZulu, the phonological system is broad, and includes several distinctive features, including two main classes: vowels and consonants.

1. *Vowels*: isiZulu has a straightforward system of vowel sounds, with a total of seven, which are presented by these symbols: /a/, /i/, /u/, /ε/, /ɔ/, /o/ and /e/ (Van de Velde *et al.*, 2019). The first five symbols represent the five basic vowels or five phonemes vowels for isiZulu. The last two symbols are the variant vowels of ɔ and ε, respectively, due to vowel influence. IsiZulu, or, in particular, the Nguni languages, exhibits phonotactic constraints with respect to vowel juxtaposition, meaning that two vowels are not allowed to follow each other. They can either be deleted, merged (coalescence), or become a semi-vowel (consonantalisation) (Van de Velde *et al.*, 2019).
2. *Consonants*: This is the second class that divides sound for isiZulu, i.e., consonants, and this class is further subdivided into plain consonants and clicks (Taljaard & Snyman, 1989). Nguni languages have breathy or murmured consonants (Nurse & Philippson, 2006). Previous studies have considered the number of isiZulu consonants to be around 42, i.e. 33 elementary plain consonants, and nine click consonant sounds (Doke, 1923). More recent studies, however, such as those by Naidoo *et al.* (2005), have increased this number to 59.

A second category of consonants in isiZulu comprises clicks, which are believed to have been inherited from interactions with the San Bushmen (Doke, 1923; Kubeka, 1979). Clicks are distinctive to Nguni languages and are also found in some Sotho languages, likely due to influence from Nguni languages (Doke, 1923). In isiZulu, clicks are present in approximately one-seventh of the vocabulary (Marks, 1967). There are three primary types of click sounds: dental (e.g., /c/), alveolar (e.g., /q/),

and lateral (e.g., /x/). Within these, they can be further distinguished as voiceless, voiced, nalised, prenasalised and aspirated (Naidoo *et al.*, 2005)

In addition to the phonemic inventory, isiZulu consonants exhibit morphophonological alternation. The changes between lexical and surface forms are due to interactions between morphology and phonology, such as assimilation, prenasalisation, palatalisation, and lenition (Eddington, 2012; Herbert, 1977). This variation is particularly relevant for morphological segmentation because it complicates the identification of morpheme boundaries in computational analyses.

3. *Tone*: Tone is a central aspect of isiZulu phonology, with high and low tones playing a significant role in distinguishing meaning (Nurse & Philippson, 2006). This is different from sentence intonation, which alters voice pitch to convey different types of meaning within a sentence or phrase. Intonation in isiZulu plays almost a similar role, but at the word level and, when omitted, may result in misunderstanding. Almost every word in isiZulu and other Nguni languages carries a distinctive tone, which can be inferred from spoken language and affects vowels or syllables, playing a crucial role in distinguishing meaning (Doke, 1923). However, the way in which tones interact in isiZulu is highly complex, and is often influenced by breathy consonants, making it difficult for second-language learners to determine the correct underlying tone of a morpheme. This challenge is further compounded by the fact that tonal information is not reflected in standard isiZulu orthography (Buell, 2005). While tone is a significant linguistic feature of isiZulu, it is not a primary concern in this study, because morphological segmentation relies on written text, where tone is not explicitly marked in standard orthography.

4. *Syllable Structure*

IsiZulu follows a CV (consonant-vowel) dominant syllable structure, meaning that most syllables consist of a consonant (C) followed by a vowel (V). Unlike English, isiZulu does not allow complex consonant clusters or codas (consonants at the end of a syllable), making it a predominantly open-syllable language (Khan & Kadenge, 2016; Naidoo *et al.*, 2005).

IsiZulu phonotactics do not permit consonant clusters at the beginning or end of a syllable. When foreign words with clusters are borrowed, they undergo adaptation to fit the CV structure. For example, the word *school* with a closed syllable [CVC] is resyllabified to “*isikole*” [CV.CV] by vowel insertion. These morphophonological adjustments are necessary for maintaining isiZulu’s syllable constraints (Khan & Kadenge, 2016).

Kotzé and Wolff (2014) explored the use of syllable-based segmentation as a tokenisation strategy for training a statistical machine translation system between English and isiZulu. They syllabified isiZulu texts prior to training to facilitate better alignment between the two languages. This approach leverages isiZulu’s CV syllable structure to improve translation quality by simulating morpheme boundaries with syllable boundaries, thereby enhancing the system’s ability to handle the language’s agglutinative nature.

3.2.2.2 Syntax

In general terms, the syntax of a language includes the rules that govern the construction of grammatical sentences, defining the relationships and order of words within a sentence. IsiZulu, like other Nguni languages, demonstrates a complex interplay between syntax and morphology, significantly influencing its morphosyntactic properties. The predominant sentence structure in isiZulu follows the Subject Verb Object (SVO) order. However, this structure is complexly shaped by isiZulu’s rich agreement system, which is closely tied to its detailed noun class system. This agreement system is essential to the language’s syntax and ensures that subjects, verbs, and often objects are consistent with noun classifications (Pretorius & Bosch, 2003). These classifications affect various grammatical aspects such as verb forms, possessive constructions, and adjective agreement, reflecting a deep syntactic integration of morphological features (Bickel & Nichols, 2007; Keet & Khumalo, 2014).

It should be noted that syntax extends beyond organising words or phrases into sentences; it involves understanding how these elements relate to broader interactive contexts. This notion emphasises the integral connection between syntax and other core linguistic areas such as morphology, semantics, and pragmatics.

However, it is important to relate these syntactic and morphological characteristics also to natural language processing (NLP) systems, which the present study aims to enhance.

3.2.2.3 Semantics

While syntax is concerned with the rules that govern the construction of grammatical sentences, semantics deals with the assignment of meaning to words, phrases, and sentences; simply put, grammatical correctness does not necessarily translate to meaningfulness. For example, the sentence “*isikole siya emfaneni*” (the school is going to a boy) might be recognised as isiZulu by someone with a basic understanding of the language, but it lacks coherent meaning. This illustrates that while sentences can be grammatically correct, they may still fail to convey a clear or meaningful message.

Another notable distinction arises when certain words are present in the syntax of a sentence but do not contribute to the overall meaning. In English, these are known as *stop words*, such as ‘and’, ‘as’, ‘is’, and others. However, in isiZulu, such functional words are less common due to the language’s complex morphology, which relies more on inflectional and derivational processes rather than on independently functioning words.

3.2.2.4 Pragmatics

In the section above semantics is described as the study of meaning in language at the word, phrase, and sentence level, dealing primarily with literal and inherent meanings. Pragmatics extends this by examining how context influences and often transforms these meanings in actual usage. For instance, the interpretation of pronouns like ‘*mina*’ (me) and ‘*nawe*’ (you) can significantly vary depending on the speaker, listener, and situational context. This illustrates how pragmatic factors like the social relationship between the speaker and listener, as well as the circumstances under which the conversation occurs, impact how messages are understood (Hazen, 2014).

3.2.3 Morphological Characteristics: Overview

This section highlights a number of the morphological characteristics of isiZulu. In linguistics, morphology is defined as a study of the internal structure of words and the rules which govern their formation (Haspelmath & Sims, 2013). To ensure a clear understanding of this chapter and the upcoming ones, it is essential to introduce and properly define the underlying concepts and terminologies, as they may vary in meaning depending on the context. For instance, the linguistic term *word* can have different interpretations depending on how it is used.

3.2.3.1 Definitions and Concepts of Words

There has been no universal consensus on the exact definition of a word (Blevins, 2014; Krámský, 1969). For example, an early attempt to define a word, presented by Bloomfield (1995), described it as a “minimum free form”, while Bolinger and Sears (1981) proposed this definition of a word to be “the smallest unit of language that can be used by itself”. Here, “used by itself” refers to the capability of words to be used as stand-alone or free forms. A simple example is the possessive determiner ‘yami’ (my), which depends on the noun class.

Various definitions of a word have been presented in the literature (Booij, 2012; Haspelmath & Sims, 2013; Lyons, 1995). This study adopts the recent definition presented by Haspelmath (2023), which defines a word as “a free morph, a clitic, a root or a compound possibly augmented by non-required affixes, and augmented by required affixes if there are any”. This definition refers to the word form as compared to the lexeme. A lexeme refers to the underlying unit of meaning in a language, which gives rise to various word forms through inflection (Johnston & Schembri, 1999). Clearly, this definition of a word alone emphasises the importance of dedicating a portion of this section to introducing some of the terminologies and concepts that serve as the underlying structure of this chapter and those to follow.

3.2.3.2 Morphemes and Word Formation

With the above definition of *word* in mind, morphology can be rephrased as a study that examines how words are built from smaller units, such as morphemes or morphs, which consist of roots, affixes, and clitics. A *morpheme* is the smallest meaning-bearing unit of a word (Haspelmath & Sims, 2013). It is a minimal segment of language that cannot be further divided without losing its function or meaning, where this meaning can be either lexical (conveying a concrete meaning) or grammatical (providing structural information). On the other hand, according to Zhang (cited by Cao (2022:324)), “Morphs are the phonological (spoken) or orthographic (written) forms to realise morphemes, and they are minimal carriers of meaning”.

In respect of morphemes, they can be classified into two categories: free morphemes (also known as lexical morphemes) and bound morphemes. The former can stand alone as independent words, whereas the latter must attach to other morphemes to convey meaning. Following this classification and the presented word definition, it can be further deduced that every free morpheme is a word. However, free morphemes are relatively scarce in Bantu languages due to their complex morphology, where words are made up of morpheme roots, and one or more morphemes affixes (Faaß *et al.*, 2012).

This structure results in words being composed of multiple morphemes that are joined together to convey specific meanings. For instance, in languages like isiZulu, certain noun stems require prefixes to form complete words, indicating that the stems themselves are bound morphemes. This morphological characteristic highlights the intricate word formation processes inherent in Bantu languages.

Furthermore, in Bantu languages, most words in various major word categories depend on affixes. Typically, words are made up of two or more morphemes: a *root* morpheme, which is considered “the core element of a word, the part that carries its meaning” (Poulos & Msimang, 1998:170), and *affixes* morphemes, which actualise the root as a complete word.

The affixed morphemes are all bound morphemes with distinct and well-defined functions within the word, and are further divided into *derivational* and *inflectional* morphemes.

Derivational morphemes form new words by adding prefixes or suffixes, often changing the word's original meaning and sometimes its grammatical category. Inflectional morphemes, on the other hand, mark grammatical distinctions like tense, possession, and plural forms by adding suffixes. Prefixes are the syllables or elements that precede the root, while suffixes follow the root (Kambarami *et al.*, 2021). Prefixes can include class prefixes, concords, tense formatives, and more, while suffixes encompass, but are not limited to, terminatives, tense markers, and other morphemes that modify the root to indicate grammatical functions such as aspect, mood, and number (Poulos & Msimang, 1998).

3.2.3.3 Morphological Typologies

In terms of morphology, languages are often classified into four main morphological typologies: isolating, agglutinative, fusional, and polysynthetic (Comrie, 1989). Isolating languages have a simplistic morphological structure where there exists a one-to-one correspondence between words and morphemes, with words primarily consisting of free boundary morphemes. In contrast, agglutinative language words are typically formed by combining one or more bound morphemes. The boundaries between these morphemes are clear, and each morpheme in this morphological typology tends to carry a distinct or grammatical function or invariant meanings that are easily identifiable, making segmentation theoretically straightforward (Pirkola, 2001). However, while theoretical segmentation in agglutinative languages may seem easier, automatic morphological segmentation is still challenging because morphemes undergo *phonological changes* (e.g., nasalisation in isiZulu), or *morphophonological alternations* (e.g., assimilation and deletion). Furthermore, some words may be made up of multiple morphemes, leading to long words that may introduce ambiguity in terms of segmentation.

Fusional languages, while also commonly composed of bound morphemes, such as in agglutinative languages. However, they differ from the former in the sense that the morpheme boundaries within words are less distinct. Moreover, a single morpheme may encode multiple morphosyntactic functions within a word, adding complexity to segmentation (Kambarami *et al.*, 2021).

The fourth and the last morphological typology is the polysynthetic or incorporating language. However, Comrie (1989) introduces these terms as different with *polysynthetic* derived from polysynthesis which refers to how large numbers of morphemes (lexical or grammatical) combine into a single word, where the latter refers to the possibility of combining multiple lexical morphemes into a word, with emphasis that a language can be polysynthetic but not incorporated.

Finally, *polysynthetic* languages, also known as *incorporating* languages, combine large numbers of morphemes, both lexical and grammatical, into a single word. The term “polysynthetic” refers to the synthesis of many morphemes (lexical and grammatical), while “incorporating” emphasises the combination of multiple lexical morphemes within one word. Importantly, a language can be polysynthetic without necessarily being incorporating (Comrie, 1989).

3.2.3.4 Morphological Characteristics of isiZulu

IsiZulu, like many other Bantu languages, falls under the category of agglutinative languages. This implies that words in isiZulu are typically formed by joining multiple morphemes, which have a one-to-one mapping between form and meaning (Spiegler, 2011). There are two basic morphological systems that characterise the morphological structure of isiZulu: the *noun classification system* and *system of concordial agreement* (Pretorius & Bosch, 2003).

The noun classification system in isiZulu categorises nouns into various noun classes based on their prefixal morphemes, also known as noun prefixes. Scholars who have worked extensively on the Bantu language family have divided these prefixes into numbered classes to simplify analysis (Bleek, 1862; Meinhof *et al.*, 1932; Van der Spuy, 2001). The earliest known study of noun classes was introduced by Brusciotto (1659) in his work on the grammar of Kongo (H16), which was later expanded by Bleek (1862) in his influential work, *A Comparative Grammar of South African Languages*. In this classification system, noun prefixes in odd-numbered classes typically represent singular nouns, while even-numbered classes represent plural nouns. However, there are exceptions, as some classes are considered neutral and do not strictly adhere to this singular-plural distinction.

Table 3.2 presents 15 noun classes in isiZulu, categorised as either singular or plural, with odd-numbered classes typically denoting singular forms and even-numbered classes denoting plural forms, except for classes 14 and 15, which are neutral. This classification is based on the Bleek-Meinhof system, a widely accepted framework for Bantu noun classes (Meinhof *et al.*, 1932). The table also reflects minor adaptations introduced by Doke’s work in 1927, which expanded Bleek’s system to include additional sub-classes, notably classes 1a and 2a (Doke, 1927a). Notably, isiZulu retains class 15, even though Bleek originally proposed 22 noun classes. In contrast, classes such as 12 (*tu-*), 13 (*ka-*), 16 (*pha-*), 17 (*ku-*), and 18 (*mu-*), are absent in isiZulu, demonstrating variability across Bantu languages.

The role of noun prefixes extends beyond merely indicating noun classes; they also play a crucial role in linking nouns to other words and governing the grammatical structure of various parts of speech (Pretorius & Bosch, 2003). This is achieved through a system of

Table 3.2: Noun classes in isiZulu showing singular and plural examples.

Noun Class	Singular Example	Plural Example
1	<i>um-fana</i> (boy)	-
1a	<i>u-baba</i> (father)	-
2	-	<i>aba-fana</i> (boys)
2a	-	<i>o-baba</i> (fathers)
3	<i>umu-thombo</i> (source)	-
4	-	<i>imi-thombo</i> (sources)
5	<i>i(li)-lizwi</i> (word)	-
6	-	<i>ama-zwi</i> (words)
7	<i>isi-nkwa</i> (bread)	-
8	-	<i>izi-nkwa</i> (breads)
9	<i>in-gane</i> (child)	-
9a	<i>i-bhuku</i> (book)	-
10	-	<i>izin-gane</i> (children)
11	<i>u(lu)-swazi</i> (tree branch)	<i>ama-swazi</i> (branches)
14	<i>ubu-hle</i> (beauty)	
15	<i>uku-bona</i> (seeing)	

concordial agreement, where words in a sentence must align with the noun class to which they belong (Kambarami *et al.*, 2021). This system ensures grammatical consistency across different word categories, including pronouns, verbs, adjectives, and conjunctions and so forth, as illustrated in the example below. This grammatical agreement, commonly referred to as concords, is typically categorised into subject concords, object concords, and possessive concords (Spiegler, 2011).

Example: Consider the isiZulu sentence - *Izingane zikababa zidlala ibhola* ('My father's children are playing football').

In this example, the prefix **zi-** in both *zikababa* and *zidlala* shows concordial agreement with the noun *izingane* (Class 10). This ensures that both the possessive construction *zikababa* and the verb *zidlala* are grammatically aligned with the subject of the sentence *izingane*. This demonstrates a grammatical correlation between different word types. Consequently, to manipulate the subject of the sentence and still retain a semantically valid sentence, one would also need to modify the prefixes and concords of various words in the sentence due to this concordial agreement system, making the morphological complexity distinct.

3.2.3.5 IsiZulu Orthography

In discussing the morphology of agglutinating languages, it is crucial to consider the orthography of Southern African Bantu languages. These languages can be categorised into two writing systems: disjunctive and conjunctive. A language is considered to be disjunctively written when a single linguistic word is split into multiple orthographic words (Taljard & Bosch, 2006). In contrast, conjunctively written languages like isiZulu maintain a one-to-one correspondence between linguistic and orthographic words (Bosch & Pretorius, 2002).

For instance, one can consider the phrase “I will see them”, which in Southern Sotho, a disjunctively written language, it is written as, *‘ke tla ba bona’* with four distinct orthographic words written separately. However, in a conjunctively written language like isiZulu, this phrase is written as one orthographic word that corresponds to its linguistic word, *‘ngizobabona’*, with its constituent morphemes *‘ngi + zo + ba + bona’*. Despite both languages belonging to the Bantu family, their writing systems differ significantly.

It is not common for languages of the same family to use different writing systems. However, as noted by Wilkes (1985), this distinction is necessary for languages like isiZulu due to phonological processes such as vowel coalescence, vowel elision, and consonantalisation. These processes often result in morphemes that are not syllabic, making disjunctive orthography impractical, and could result in sentences that are difficult to read and pronounce.

3.3 Morphological Segmentation: Overview

The above section provided an overview of the characteristics of Nguni languages, with a specific focus on isiZulu, examining its complex morphological structure and writing system. These linguistic features present significant challenges for researchers and developers who are creating essential language processing technologies. Consequently, this complexity poses a threat to the long-term digital adaptability of these languages. Understanding isiZulu and similar languages through morphological segmentation is crucial for enabling the development of such technologies. Morphological segmentation plays a salient role in NLP by enabling the division of words into their constituent morphemes. This section provides an overview of the concept of morphological segmentation, discussing various strategies used to perform this task, and outlining how the effectiveness of these strategies is typically evaluated.

3.3.1 Introduction to Morphological Segmentation

Morphological segmentation, or decomposition, has been an active area of research for many years. One of the earliest and foundational works in this domain, which laid the groundwork for using automatic procedures to identify morphemes and their internal boundaries, is that of Harris (1955). Over time, a wide range of morphological segmentation techniques have emerged in computational linguistics. Researchers generally categorise the task of morphological segmentation into two types: canonical segmentation and surface segmentation.

3.3.2 Canonical Segmentation

Canonical segmentation involves breaking down a word into segments while restoring the orthographic changes of its morphemes (Cotterell *et al.*, 2016a). This technique reverses the orthographic changes that occur during word formation, treating the task as a sequence-to-sequence problem where a word is mapped to its canonical morphemes, separated by segmentation boundaries (Rouhe *et al.*, 2022). To illustrate, one can consider the canonical segmentation of the word *ngebhubesi* (meaning “with a lion”):

$$1. \text{ ngebhubesi} \rightarrow \text{nga} + \text{i} + \text{bhubesi}$$

In this example, the word *ngebhubesi* is segmented into morphemes that represent the underlying canonical forms or lemmas, with the phonological changes reversed. The resulting morphemes, *nga+i+bhubesi*, are considered canonical morphemes. As a result, determining the correct morpheme boundaries in this form is relatively straightforward.

3.3.3 Surface Segmentation

Surface segmentation involves breaking a word into morphemes as they appear in their surface form, reflecting their actual usage in language. The surface form of a word refers to the way it is pronounced or written after phonological or morphological processes have been applied. In other words, it represents the final output of linguistic transformations rather than the underlying morphemes in their canonical form. The concatenation of these segments based on morphemes (morphs) results in the original orthographic word (Mager *et al.*, 2020). This task is also treated as a sequence-to-sequence problem, where a word is split into its morphs by inserting morpheme boundaries. The surface segmentation of the word “*ngebhubesi*” can be seen in the following examples:

$$1. \text{ ngebhubesi} \rightarrow \text{nge-bhubesi}$$

2. $ng\text{e}bh\text{u}besi \rightarrow ng\text{-}e\text{-}bh\text{u}besi$

3. $ng\text{e}bh\text{u}besi \rightarrow ng\text{-}eb\text{h}u\text{b}esi$

Surface segmentation only captures the morphemes in their surface form without reversing the orthographic changes that occurred during word formation. As a result, various morpheme splits can be produced for a single word. For example, during the formation of the word “*ngebhubesi*”, the underlying canonical morphemes *nga* (with) and *i* (the) combine. Since isiZulu prohibits the occurrence of two vowels in succession, the vowels *a* from *nga* and *i* from *i* fuse to form a new morpheme, *e*. From a linguistic perspective, different segmentations are possible. In segmentation (1), the *e* combines with *ng*, while in segmentation (2), it stands between *ng* and *bhubesi*. In segmentation (3), the *e* attaches to *bhubesi*.

This segmentation approach presents an ambiguity in terms of determining where the morpheme boundary should be placed, as there is no single “correct” answer from a linguistic perspective. However, certain segmentations may be more useful from a computational standpoint, depending on the specific application or processing requirements.

3.4 Approaches to Morphological Segmentation

Two main approaches are commonly used to perform morphological segmentation: rule-based approaches and machine-learning approaches (Anand Kumar *et al.*, 2010). Rule-based systems rely on handcrafted linguistic rules, while machine learning models are trained on annotated data to automatically infer patterns. This section provides a discussion of these techniques, highlighting key studies that have adopted them, as well as the strengths and weaknesses associated with each method. Understanding these approaches is essential for developing effective NLP systems for agglutinative languages, where accurate morphological segmentation is crucial for tasks such as machine translation and language modelling.

3.4.1 Rule-Based Approaches

Rule-based approaches or knowledge-based systems in morphological segmentation rely on predefined linguistic rules and handcrafted lexicons to segment words into constituent morphemes. These sets of rules are usually generated based on experts’ knowledge and the morphological structure of the given language (Eiselen & Puttkammer, 2014b). To create a morphological segmenter using this technique, one does not need large amounts of data, thus making them easier to implement and execute for low-resource languages.

3.4.1.1 Finite State Machines

Finite state machines (FSM), also known as *finite-state automata* (FSA), are among the most widely used rule-based approaches in morphological segmentation. FSMs are formal mathematical models that consist of a finite number of states, through which the machine can transition based on given inputs in a process called *transitioning* (Wang, 2019). At any given time, the FSM is in exactly one state, and it moves to the next state according to predefined rules (*transition rules*), typically triggered by specific tokens (e.g., characters or words). The FSM can either accept or reject a sequence, thereby processing the input to determine if it conforms to a particular language. As Hapke *et al.* (2019:7) explain, FSMs can be thought of as “trees of if ... then ... else statements” that respond to or generate specific tokens or actions. A variant of FSMs is the *finite-state transducer* (FST), which performs state transitions like FSMs, but also has the additional capability of transforming input sequences into output sequences, effectively mapping one language or form into another.

In the literature, several foundational works have applied finite-state transducer (FST) methods for morphological analysis. For example, Brun (1998) created a morphological analyser for French verb morphology using a two-level FST system developed by Chanod (1994). This system links the surface form of a string to its morphological analysis. Another notable contribution is that of Beesley (1996), who developed the Arabic Finite-State Lexical Transducer, a two-level morphological system based on lexicons and rules. This system is adaptable and has been applied to other languages like English, German, and French, demonstrating its versatility even for languages with complex morphology. Similarly, Oflazer (2003) presented a two-level morphological analyser for Turkish, which features agglutinative morphology, using the *XRCE* Finite State Toolkit by Karttunen and Beesley (1992). This framework generates surface forms from a lexical form, further highlighting the adaptability of FSTs.

When considering Bantu languages, Swahili is notable as the first Bantu language to have a fully developed morphological analyser. This system, developed by Hurskainen (1992), was termed SWATWOL, where “SWA” stands for Swahili and “TWOL” refers to the two-level morphology approach. The system utilises a central TWOL module, which processes the input as a set of transducers compiled from morphophonological alternation rules based on the work of Karttunen *et al.* (1987). These rules allow for the handling of the morphophonological changes that occur during word formation. The system also incorporates a set of lexical entries, which serve as the foundation for analysing and processing the Swahili language, making it one of the earliest comprehensive models for Bantu languages.

A widely recognised FST system is the *Xerox Finite-state Automata* introduced by Beesley and Karttunen (2003), which models both the lexicons and morphosyntax of a language. It utilises the *Finite-State Lexicon Compiler (lexc)*, created by Karttunen (1993), to handle lexicons. To model morphophonological alternations (i.e., changes between lexical and surface forms), the system employs the *twolc* tool, developed by Karttunen and Beesley (1992). The interface *xfst* allows for regular expressions to be compiled into finite-state networks, and all components are eventually compiled into FSTs to form a lexical transducer, which performs the morphological analysis.

Using the *Xerox Finite-state Automata*, Pretorius and Bosch (2003) developed a finite-state morphological analyser for isiZulu, which later evolved into a finite-state morphology demo known as *ZulMorph*¹. The work was subsequently extended by Bosch *et al.* (2006) to create computational morphological analysers for six South African Bantu languages: isiZulu, isiXhosa, Siswati, Ndebele, Northern Sotho, and Setswana (Pretorius *et al.*, 2015). Similarly, Eiselen and Puttkammer (2014b) built upon this work by developing text resources for ten South African official languages, including morphological decomposers that follow a rule-based approach derived from previous work, such as that of Bosch *et al.* (2006). These decomposers segment words into their constituent morphemes and categorise each morpheme based on its grammatical properties, aiding in the analysis of the linguistic structure.

Beyond South Africa, Xerox Finite-State Tools have been employed in other agglutinative languages. For instance, Makedonski (2005) developed and implemented a finite-state automaton (FSA) for the nominal paradigm of the Turkish language. Additional implementations include work for Ekegusii by Elwell (2008), Kazakh by Kairakbay and Zaurbekov (2013), and Laz by Önal *et al.* (2021), each of which leveraged FSTs to handle the complex morphology inherent in these languages.

3.4.1.2 Logical Programming or Reasoning-Based Systems

This category encompasses multiple methods that encode expert knowledge using propositional logic, enabling systems to perform logical reasoning through various types of inference rules. The approaches provide a structured framework for reaching sound conclusions based on premises that are encoded as symbols, and operators representing expert knowledge in a logically inferable manner.

Logical reasoning-based approaches include deductive, inductive, and abductive reasoning. Deductive reasoning uses established inference rules to form conclusions that are logically certain, meaning that if the premises are true, the conclusion must also be true (Manicas & Kruger, 1976). In contrast, inductive reasoning allows for probable conclusions, but is

¹<https://portal.sadilar.org/FiniteState/demo/zulmorph/>

not guaranteed, meaning that the premises provide some evidence but do not necessarily confirm the conclusion’s validity (Hawthorne, 2004). Finally, abductive reasoning seeks the most likely explanation or hypothesis for a given set of observations, even though it may not be definitively proven, making it useful for handling uncertainty and incomplete information (Paul, 1993).

Further work following this method can be found in Spiegler, Golénia, and Flach (2010), known as Deductive-Abductive Parsing (DEAP), which is employed for morphological analysis looking at languages with complex morphologies, namely Arabic and isiZulu. DEAP is a hybrid system, combining rule-based logical reasoning with some statistical techniques. In their implementation, Context-Free Grammars (CFGs) (Stolcke, 1995) were used. These CFGs consist of rules that describe possible sequences of categories or labels, along with a dictionary that links morphemes to their respective labels. CFGs entail a formal type of grammar that generates all possible combinations of a string by defining finite, predefined production rules (Stolcke, 1995). This structure helps to organise and segment morphemes within a sentence according to logical rules.

Similarly, Spiegler, Van Der Spuy, and Flach (2010) presented the Ukwabelana project that utilises a semi-automatic approach that implements a partial morphological grammar for isiZulu, based on Definite Clause Grammar (Pereira & Warren, 1980). This grammar was designed to create surface-segmented isiZulu data by defining structured grammatical rules for segmenting words. This corpus resulted in 10 000 morphological surface segmented word types. This labelled surface segmented corpus has served as the basis for several works that employ surface segmentation for isiZulu (Cotterell *et al.*, 2015; Liu & Prud’Hommeaux, 2022).

Most of the aforementioned rule-based tools performed well in low-resource settings such as that of Nguni languages. Yet, challenges have been reported to be associated with them in the literature. The most common challenge is their dependence on experts’ knowledge for the development, maintenance, and expansion of their rules (Du Toit & Puttkammer, 2021). This subsequently results in their initial development and, in some cases, in their maintenance being an expensive and time-consuming endeavour.

Another challenge is that rule-based systems rely on predefined rules to produce correct segmentation for a given word token. This feature limits their effectiveness to known or expected instances, meaning that they often perform poorly on words with complex morphological structures not fully covered by the rules (Du Toit & Puttkammer, 2021; Villegas-Ch *et al.*, 2024). As a result, they are less robust in handling linguistic variability and unseen vocabulary.

Lastly, Mzamo (2021) argues that since the majority of these rule-based approaches produce various valid segmentation, among these possibilities, only one is correct, presenting a need for disambiguation. In their work, they refer back to the work of Bosch (2020),

who presented a comparative study on isiZulu morphological systems that demonstrate this feature. Despite such setbacks, these methods have been considered useful when dealing with small datasets, and these methods can accurately segment these texts since they are guided by linguistic experts, making them useful for various downstream tasks once they are developed (Grönroos *et al.*, 2016; Zbib *et al.*, 2012).

3.4.1.3 Review of the isiZulu Resource Grammar (ZRG)

3.4.1.3.1 Grammar Formalism and GF: Overview

The Grammatical Framework (GF) is a highly modular grammar formalism that separates abstract syntax from concrete syntax. Abstract syntax defines language-independent structures using a Logical Framework (LF), while concrete syntaxes map these structures to language-specific forms, handling details such as inflection and word order (Ranta, 2004). GF supports multilingual grammars by handling linguistic variation across languages (morphology, syntax) and shared structures. It is modular and scalable, allowing for the reuse of grammar components across different languages (Ranta, 2007). This capability is encapsulated in the GF Resource Grammar Library (RGL) (Ranta, 2009), which provides a reusable set of grammatical rules and structures for a wide variety of languages, including rules for word inflection, agreement, and syntax, thus streamlining the development of multilingual systems.

GF is built on a type-theoretical foundation with strong static type checking, ensuring that grammars remain consistent and error-free. Its Portable Grammar Format (PGF) (Angelov *et al.*, 2010) allows for the deployment of standalone GF applications across multiple programming environments, such as Haskell, JavaScript, Java, C#, and C. GF has been successfully applied in various domains including natural language generation (Nguyen, 2018), translation systems (Ranta *et al.*, 2015), and dialogue systems (Bel-Enguix *et al.*, 2006).

GF is built upon two fundamental processes: *parsing* and *linearisation* (Ranta, 2004). Parsing is the process by which a natural language string is analysed to produce an abstract syntax tree (AST), representing its underlying grammatical structure independent of any particular language. Linearisation, on the other hand, is the reverse process, where an abstract syntax tree is converted back into a natural language string according to a specific concrete syntax. These complementary processes allow GF to function both as an interpreter and as a generator, thereby facilitating robust multilingual language processing.

3.4.1.3.2 Purpose and Functionality of ZRG

The Zulu Resource Grammar (ZRG) by Marais and Pretorius (2023) is an innovative implementation within the GF that specifically models the morphology and syntax of isiZulu, which is characterised by agglutinative morphology, conjunctive orthography, and extensive morphophonological alternation. Building on GF’s core principles of abstract syntax, concrete syntax, and its Logical Framework (LF), the ZRG extends the capabilities of GF by addressing the linguistic challenges inherent to isiZulu. A key innovation of ZRG is its ability to effectively model the morphophonological alternations, treating certain aspects of morphology essentially as syntactic phenomena. This is achieved through the use of GF’s orthography engineering features, most notably by employing a specialised binding token (BIND), which is used to allow for the dynamic construction of tokens, or surface forms, at runtime and without introducing spaces. This design decision not only bypasses the potential intractability associated with pre-compiling all possible morpheme combinations but, also maintains linguistic transparency and robustness in token segmentation.

3.4.1.3.3 Modeling IsiZulu Morphosyntax

IsiZulu’s linguistic features, which consists of agglutinative morphology, rich system of noun classes, and extensive agreement mechanisms, pose a significant modelling challenge. ZRG addresses these by:

1. Modularising morphological components: Each morpheme (e.g., noun prefixes, concords, tense markers) is treated as an independent grammatical entity, yet is later bound together into a single orthographic token at runtime.
2. Encoding morphophonological alternations as syntactic agreements: For instance, the grammar uses parameters to “remember” abstract morpheme forms, ensuring that agreement phenomena (such as the alternation between -na- and -ne-) are correctly reflected in the linearisation process.
3. Extending GF’s standard framework: While GF’s RGL provides a foundation for many languages, ZRG requires significant extensions and deviations to handle isiZulu’s morphological complexity without incurring intractable grammar sizes or overgeneration.

3.4.1.3.4 Chunk Parsing Extension for Robustness

GF-based parsers can be brittle when faced with input that does not perfectly conform to the grammar. To improve robustness, ZRG was extended with a chunk parsing mechanism. This extension introduces new chunk categories that mirror standard syntactic categories (e.g., *ChunkNP* for noun phrases, *ChunkVP* for verb phrases), allowing the parser to decompose sentences into well-formed “chunks” rather than failing entirely on unrecognised segments. Such a mechanism is particularly beneficial in a pro-drop language like isiZulu, where agreement information must be preserved even when subjects are omitted. The chunking extension thus enhances the parser’s ability to recover from errors and to produce partial yet informative parses, which is another feature.

3.4.1.3.5 Evaluation and Limitations

The initial application of ZRG, as reported by Marais and Pretorius (2023), demonstrates its pioneering role as the first openly available parser for isiZulu using GF. It was evaluated on a corpus of carefully curated sentences; where the ZRG demonstrated high token-level coverage by successfully parsing over 97% of tokens in the evaluation set. However, its reliance on a handcrafted lexicon (with nearly 17 000 entries) and explicit morphological rules means that certain linguistic phenomena (e.g., particular auxiliary forms, subjunctive agreements, or adverbial prefixes) still lead to unparsed tokens that are returned as generic symbol chunks *Symb*. Moreover, the parser’s performance is sensitive to out-of-vocabulary items and deviations from expected grammatical structures. While these limitations highlight areas for further refinement, the current version of ZRG constitutes a robust baseline and a valuable tool for both syntactic analysis and the generation of segmented data.

3.4.1.3.6 Implications for Data Generation and Future Work

The flexibility of ZRG extends beyond parsing: by adapting its core morphology operations, researchers can generate annotated, surface-segmented datasets that reflect specific segmentation strategies. This capability is particularly advantageous in resource-scarce environments, where manual annotation is expensive and time-consuming. Recently, Mkhwanazi and Marais (2024) demonstrated that ZRG-based segmentation can match or even surpass existing annotated datasets, highlighting its potential for training more effective segmentation models. This work builds on that foundation, aiming to further explore and enhance segmentation techniques while advancing language modelling for isiZulu.

3.4.2 Statistical Machine Learning Approaches

In contrast to rule-based approaches, machine learning techniques learn patterns from large annotated data sets or raw data to perform various tasks (Goldwater *et al.*, 2009; Murphy, 2012b). These techniques learn from examples and data, rather than relying on explicitly defined rules like in rule-based systems. There are four commonly established machine learning methods: supervised, unsupervised, semi-supervised, and reinforcement learning (Murphy, 2012b). The first three are the most commonly used in morphological segmentation. Within the machine learning paradigm, this study distinguishes between statistical machine learning and deep learning approaches, each having been applied to morphological segmentation with varying success. Statistical approaches typically rely on probabilistic models and feature extraction, while deep learning methods, particularly those involving neural networks, automatically learn features and have been shown to excel in tasks with complex morphologies (Halidanmu *et al.*, 2017).

This study classifies traditional machine learning approaches as statistical methods, as their learning algorithms rely primarily on mathematical and statistical principles to solve problems. Machine learning, a subfield of Artificial Intelligence (AI), has made a significant impact across various industries, including robotics (Vuong, 2021), optimisation (Liakos *et al.*, 2018), and healthcare (Ghassemi *et al.*, 2020), among others. It has also been applied extensively in NLP, with use cases in retail (Krishna *et al.*, 2018) and recommender systems (Singh, 2018).

NLP, itself a subfield of AI, aims to enable machines to understand and generate human language, both written and spoken. A notable characteristic of statistical machine learning (SML) is that they are usually trained on datasets with some explicitly defined (world representation) features that have been manually engineered for that specific task (Taye, 2023). There are different learning techniques that can be used to learn patterns from data ranging from supervised, unsupervised and semi-supervised. Below is an overview of these different learning techniques used to conduct morphological segmentation.

3.4.2.1 Unsupervised Morphological Segmentation

In unsupervised segmentation, algorithms analyse raw text to identify morpheme boundaries without relying on labelled data. These approaches do not require target values, making them particularly valuable for languages with limited or no annotated datasets, provided there is sufficient digital text available. Due to the scarcity of high-quality annotated corpora, researchers have increasingly focused on unsupervised machine learning techniques. This section reviews several prominent techniques developed for morphological segmentation across languages with morphological structures similar to isiZulu.

Minimum Description Length (MDL)

The Minimum Description Length (MDL) principle is one of the commonly used strategies in morphological segmentation, and it was originally introduced by Rissanen (1978). It views learning or prediction as a process of data compression and involves two mathematical components: $L(M)$, the description length of a model M and $L(D|M)$ the description length of dataset D encoded by the model M which represents a set of hypotheses in the model. The optimal model MDL is the one that minimises the function $L(M) + L(D|M)$ (Grünwald, 2007).

In the context of unsupervised morphological segmentation, the MDL principle plays two primary roles: (i) it helps to select the most optimal parameter values without requiring labelled data, making it advantageous for low-resource languages where segmented data is scarce; and (ii) it guides the segmentation process by searching for the most optimal solution among the possible segmentations, helping to identify the most efficient way to break down words into morphemes (Magistry & Sagot, 2013).

An early method presented by Brent (1993) employed the Minimum Description Length (MDL) principle combined with probabilistic techniques to discover linguistic consistencies in word structures, specifically targeting English spelling patterns. This approach worked by identifying relevant sets of suffixes within a corpus using a heuristic search algorithm. De Marcken (1995) expanded on this foundation to create an unsupervised learning algorithm based on the MDL principle, which segmented continuous streams of text into words. Starting with individual characters as the baseline lexicon, the algorithm iteratively added lexical items that maximised corpus compression while balancing this with the cost of introducing new items. This balance was key to optimise segmentation without overfitting the model to the data.

A notable advancement in this area was the work of Goldsmith (2001), who developed *Linguistica*, a system for word segmentation specifically tailored to stem+suffix languages like those found in the Indo-European family. *Linguistica* applied an MDL-based optimisation to detect common stems and suffixes, known as *signs*, without relying on labelled data. By refining morpheme segmentation iteratively, the system was able to closely match human morphological analysis. Importantly, this approach allowed for the unsupervised segmentation of words into meaningful morphemes, marking a significant step in developing data-driven approaches for morphological analysis without morphosyntactic labelling.

However, despite its success, this method has not been widely adopted for languages with complex morphology, such as isiZulu. MDL-based methods face challenges in these contexts due to the need to account for numerous possible morpheme combinations. One notable work that employs this method is that of Nkosi (2017), which investigates various segmentation decision metrics for isiZulu morphological analysis, utilising the MDL principle as implemented in *Morfessor*, which is to be discussed below.

Morfessor

These constitute a family of generative models that perform morphological segmentation of words using raw, unsegmented data. Various iterations of these models have been developed over time, a few of which are mentioned here. The oldest model in this family is known as the *Morfessor Baseline* (MBL) approach, introduced by Creutz and Lagus (2002a), and is a popular unsupervised learning system that utilises the Minimum Description Length (MDL) principle as a cost function. This model aims to minimise the overall description length by balancing two factors: the size of the morphological lexicon and the representation of words in the corpus. The goal is to find the most efficient segmentation, where the morph *lexicon* consists of unique morphs (minimal units of meaning or form). Unlike *Linguistica*, which was primarily tested on morphologically simpler languages such as English, Spanish, French, Italian, and Latin, Morfessor was applied to both English and Finnish, with Finnish presenting a significantly more complex morphological system.

The *Morfessor Baseline-Freq-Length* model, introduced by Creutz (2003), builds on the MBL by integrating prior distributions of morph length and frequency. Specifically, it uses a gamma distribution to model the lengths of the morphs, improving the segmentation of complex languages. This unsupervised model captures patterns in both the frequency and length of morphs, resulting in a more accurate morphological analysis than its predecessor, which did not explicitly account for morph length. However, despite these improvements, certain limitations persist. When the word morphs generated by the model are compared with those segmented by linguists, challenges arise, such as the under-segmentation of common strings, over-segmentation of rare strings, and violations of language-specific morphotactic rules.

To mitigate the issues of morphotactic violations and over- or under-segmentation, Creutz and Lagus (2004) introduced a refined variation of Morfessor called *Morfessor Categories-ML*. This model builds upon the segmentation produced by the baseline models, but applies additional analysis using maximum likelihood (ML) estimation along with heuristic processes to control lexicon complexity. Each segmented morph is assigned one of four categories: *prefix*, *stem*, *suffix*, and an additional *non-morpheme* or “noise” category, which is used to label segments that do not fit neatly into the other three categories.

These morphological categories help to address some of the shortcomings of the MBL. For example, over-segmentation is reduced by heuristically joining morphs that were incorrectly classified as non-morphemes with their neighbouring morphs, while under-segmentation is corrected by enforcing the segmentation of redundant morphs (those containing another identifiable morph), but without allowing them to be incorrectly labelled as non-morphemes. Additionally, this model incorporates a first-order Hidden Markov Model (HMM) to encode positional information about the morphs, making the model sensitive to context. This context-awareness aids in maintaining the morphological rules of a language, and reduces the chances of morphotactic violations, thus improving the overall segmentation accuracy.

A subsequent model, *Morfessor Categories-MAP* (MCM), introduced by Creutz and Lagus (2005), further refined the approach taken by *Morfessor Categories-ML*. This model replaced the maximum likelihood (ML) approach with a more sophisticated probabilistic maximum a posteriori (MAP) method. In this framework, both the lexicon and the corpus representation are explicitly considered probabilistic entities, and the goal is to maximise the posterior probability of these entities. This change eliminates the need for heuristic functions to control the lexicon size, which was a key characteristic of the earlier models.

One of the most notable updates in MCM is the use of a *hierarchical lexicon structure*. This hierarchical approach helps address limitations such as under-segmentation by expanding larger morphs into their constituent parts while simultaneously avoiding over-segmentation. The model ensures that non-morpheme segments are not generated by rigorously maintaining the structure of morphemes. This hierarchical mechanism allows for more accurate segmentation by considering the relationships between morphs, and thus provides a more linguistically informed approach to morphological segmentation.

The above-mentioned models were the first four from the Morfessor family (Creutz & Lagus, 2007) and have been widely used as baselines for comparing unsupervised morphological segmentation and for various extensions. The official update, known as *Morfessor 2.0*, was detailed in Virpioja *et al.* (2013), with new features. Over time, newer versions emerged to address some of the limitations in these models. One notable extension is *Allomorfessor*, introduced by Virpioja *et al.* (2009) and Kohonen *et al.* (2009), which addresses the linguistic phenomenon of *allomorphy*, where a single morpheme can have multiple surface forms. This model extends the Morfessor framework to accommodate the presence of allomorphs, refining the segmentation process by allowing for morphs to be grouped under a common abstract form.

There have been several works that use these different baseline models to benchmark the performance of unsupervised morphological segmentation systems. For example, in relation to the Nguni languages, Mzamo *et al.* (2019b) and Mzamo *et al.* (2020) benchmarked their isiXhosa unsupervised segmenter models using the oldest of these models, the MBL

(Creutz & Lagus, 2002b). Similarly, other works on Nguni languages, such as Meyer and Buys (2022) and Moeng *et al.* (2022), used MBL as a comparison model for their subword-level unsupervised segmentation systems. The work of Spiegler *et al.* (2008), which explores varying levels of supervision, uses the MCM to develop an unsupervised morphological segmenter for building a morphological analyser for isiZulu.

Entropy-based Models

The next set of models to be discussed in this study are those using the Branching Entropy technique. The Branching Entropy model was introduced by Jin and Tanaka-Ishii (2006), and builds on the concept of accessory variety (AV) (Feng *et al.*, 2004). AV refers to the number of distinct “accessor varieties” - the characters appearing before or after a particular string in an unlabelled corpus. Several works have leveraged this technique in unsupervised morphological segmentation (Feng *et al.*, 2005; Guo *et al.*, 2012; Sun & Xu, 2011). This method performs more effectively in large corpora, because the broader word coverage allows for more accurate calculation of preceding and following characters.

However, a notable challenge arises due to AV’s proportionality to corpus size, leading to inflated AV values in larger datasets. These higher values may not accurately reflect the true uncertainty in the data. To address this, previous studies attempted to normalise AV values using *frequency-based uncertainty classes*. However, this normalisation process is not straightforward, as determining the appropriate settings for these classes and threshold values is complex and context-dependent.

To address the issue of normalising AV, Jin and Tanaka-Ishii (2006) proposed a more robust technique known as *Branching Entropy* (BE). Instead of relying on simple counts of preceding or succeeding characters, BE calculates the entropy of accessor characters for a given string, representing uncertainty in a probabilistic manner. This allows BE to remain effective regardless of corpus size, solving the scaling issues associated with AV.

Branching Entropy operates on the observation that words often have distinct beginnings and endings, characterised by a sharp increase in uncertainty about the characters that precede or follow them. By analysing these fluctuations in uncertainty, the method identifies word boundaries where the branching entropy spikes, signalling a shift in the predictability of the string (Sudoh, 2015). BE for an n-gram $x_{i...j}$ is calculated as:

$$h(x_{i...j}) = - \sum_{x \in V} p(x|x_{i...j}) \log p(x|x_{i...j}),$$

where V is the set of surrounding characters, and $p(x|x_{i...j})$ represents the probability of character x in the context of $x_{i...j}$. BE is divided into two types: Left Branching Entropy (h_L), based on preceding characters, and Right Branching Entropy (h_R), based on following characters. For new word detection, the smaller value of h_L and h_R is compared to a threshold; if it exceeds the threshold, the n -gram likely forms a word boundary.

Mzamo *et al.* (2019a) presented two unsupervised morphological segmentations for isiXhosa: (1) The IsiXhosa Branching Entropy Segmenter (XBES), this is an entropy-based segmenter operates as a character-level language model to predict segmentation boundaries based on the generated conditional entropy probability, (2) the IsiXhosa Heuristic Maximum Likelihood Segmenter (XHMLS), which is a model that uses heuristics to produce possible segmentations and finally uses probabilistic graphical models to rate the most probable word segmentation. This work was followed by that of Moeng *et al.* (2022), where an entropy-based model based Mzamo *et al.* (2019a) XHMLS algorithm was used to conduct unsupervised surface segmentation for Nguni languages. Another work using this approach is that of Dlamini *et al.* (2020), which uses a maximum entropy morphological analyser to segment the isiZulu words to train morpheme-based embeddings.

Morpho Challenge Algorithms

Morpho Challenge is a peer-operated competition that utilises statistical machine learning algorithms to perform word morphological analysis. This series of competitions began in 2005, and aimed at fostering the development of effective morphological analysis algorithms from various practitioners. The models of the participants are evaluated on the basis of specific downstream tasks, datasets, and metrics established by the competition organisers.

The objective of the first Morpho Challenge was to create algorithms capable of segmenting raw text into morphemes that could be useful for a downstream task, with performance being automatically compared against a linguistic reference “gold standard” analysis or state-of-the-art evaluation tasks. In the follow-up Morpho Challenge of 2007, the competition remained unsupervised, requiring participants to not only segment raw text, but also to perform morphological analysis on the identified morphemes (Kurimo *et al.*, 2007). By 2010, the Morpho Challenge introduced a semi-supervised component, allowing participants to utilise a small sample of morphologically segmented data in addition to the unsupervised approaches (Virpioja *et al.*, 2011).

As noted by Kurimo, Virpioja, Turunen, and Lagus (2010:87), “more than 50 algorithms have already been tested in the Morpho Challenge evaluations”. For an in-depth exploration of other techniques submitted to the Morpho Challenge, readers can refer to these foundational works (Kurimo, Virpioja, & Turunen, 2010; Kurimo, Virpioja, Turunen, Blackwood, & Byrne, 2010; Kurimo *et al.*, 2006, 2007, 2009), which offer comprehensive

insights into the algorithms and methods utilised in the competition. Although none of these approaches were specifically implemented for Nguni languages, the Morpho Challenge represents a significant effort in advancing the field of morphological segmentation, and it serves as a valuable foundation for continued research in this area.

3.4.2.1.1 General Limitations Across Methods

While unsupervised approaches have advanced morphological segmentation, several limitations persist, particularly for languages with complex morphology such as isiZulu. Many of the discussed methods face challenges with agglutinative languages, since morphemes are concatenated in ways that create long-distance dependencies. These methods often struggle to accurately capture such structures, resulting in reduced segmentation quality (Ansari *et al.*, 2019). Additionally, the lack of linguistic constraints in purely statistical approaches can lead to inconsistent segmentations, requiring extensive post-processing or manual adjustments to improve accuracy. Moreover, methods that rely solely on statistical patterns without integrating linguistic insights may fall short in identifying rare or irregular morphemes, which are common in morphologically rich languages. As a result, segmentation performance in these cases may not meet the requirements for downstream tasks, presenting the need for more robust models that can handle the intricacies of agglutinative and low-resource languages effectively.

Since this study primarily adopts a supervised approach, the focus will now shift towards examining supervised models and evaluating their effectiveness for morphological segmentation.

3.4.2.2 Supervised Morphological Segmentation

Unlike unsupervised morphological segmentation, which learns to segment from raw text data without labels, supervised methods require a labelled or annotated dataset to guide the model in identifying morpheme boundaries. In supervised learning, an algorithm constructs a mathematical function using the labelled data, mapping the input (word forms) to the desired output (morpheme boundaries). These labelled datasets are typically hand-encoded by linguistic experts, making the process highly dependent on the quality and accuracy of the annotations.

Several supervised techniques have been developed over the past few years, each employing various approaches to morphological segmentation. This section discusses some of the most commonly used techniques, with a particular focus on those developed for Nguni languages or languages with similar morphological structures.

Hidden Markov Models

Hidden Markov Models (HMMs) (Rabiner & Juang, 1986) are directed graphical generative models commonly used to model sequence distributions $p(x, y)$ over two sequences of variables, $\mathbf{X} = (x_1, x_2, \dots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \dots, y_n)$, representing input and output sequences, respectively. HMMs are particularly useful for time series data and are widely applied in estimating the joint distribution $p(x_i, y_i)$ as well as the conditional distribution $p(y_i | y_{i-n}, \dots, y_{i-1})$, where the current output depends on past observations.

(1) *Hidden states* represent unobservable discrete states that influence observable data, and (2) *Observable symbols* are the visible outputs probabilistically linked to the hidden states. The model transitions between hidden states based on probabilities, emitting observable symbols in the process. This enables HMMs to effectively capture sequential dependencies, making them valuable for applications like speech recognition (Chavan & Sable, 2013), part-of-speech tagging (Banko & Moore, 2004; Kupiec, 1992), and morphological segmentation (Cing & Soe, 2020; Tukur *et al.*, 2020).

The primary goal in such undertakings is to infer hidden structures, such as morphemes, from observable data like words, often employing a dynamic programming algorithm known as the Viterbi algorithm (Viterbi, 1967) to find the most likely sequence of hidden states. However, their application to Nguni languages appears to be limited. With these languages, a significant drawback of HMM models is their sensitivity to rich, interdependent features. This results from how the HMM takes an inflexible assumption about the input $\mathbf{x}$ about its format and also its dependence structure, making it challenging to capture complex dependencies between multiple features that represent $\mathbf{x}$ effectively.

3.4.2.2.1 Maximum Entropy Markov Models

Another undirected graphical model is the Maximum Entropy Markov Model (MEMM), also known as *MaxEnt*. These techniques aim to attain the simplest optimal solution and are based on the *maximum entropy principle*, which states that the least biased model is the one that uses a probability distribution that seeks to optimise the entropy with all known facts (Ratnaparkhi, 1997). Mathematically this is an exponential model that uses a training set $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$, and it seeks to train a model θ that maximises $P(\mathbf{Y}|\mathbf{X}, \theta)$. Thus, this model is unlike the HMM and has the ability to include additional interdependent features in its models of $P(\mathbf{Y}|\mathbf{X})$ without having to assume the form of $P(X)$ (Sutton & McCallum, 2012), hence solving the HMM dependency problem.

Sarikaya *et al.* (2007) effectively utilised Maximum Entropy Markov Models (MEMMs) to integrate morphological and lexical features, enhancing language modelling tasks. This model segments Arabic words into morphological components, prefixes, stems, and suffixes,

using predefined affix lists and rule-based criteria. Furthermore, Cing and Soe (2020) employed this technique for joint word segmentation and part-of-speech (POS) tagging, as well as for conducting morphological analysis of the Myanmar language.

These studies highlight the versatility and effectiveness of MEMMs in addressing morphological segmentation challenges across various languages. Although, to the best of the author’s knowledge, no prior work has directly applied MEMMs to Nguni languages, these techniques provide valuable insights. The approaches discussed in the following sections build upon these foundational models, adapting them to the morphological characteristics of isiZulu and related languages.

However, a common limitation reported in these works is the issue of *label bias*, where the model tends to favour states with fewer outgoing transitions, leading to less accurate predictions (Lafferty *et al.*, 2001). This constraint reduces the model’s flexibility to select among equally probable transitions, especially in complex segmentation tasks.

Conditional Random Field Models

The final graphical model to be discussed is the Conditional Random Fields (CRF) model. CRFs were introduced by Lafferty *et al.* (2001) as a technique for word segmentation and tagging. This probabilistic model retains all the advantages of Maximum Entropy Markov Models (MEMMs) while addressing their main disadvantage: the label bias problem. Rather than modelling the joint distribution $p(x, y)$ of local transitions, as in MEMMs, CRFs model the joint probability of the entire label sequence $\mathbf{Y} = (y_1, y_2, \dots, y_n)$ given the entire input sequence $\mathbf{X} = (x_1, x_2, \dots, x_n)$.

The graphical structure constitutes significant differences between CRFs and previously discussed models, including the HMMs, and MEMMs. While HMMs and MEMMs are defined using directed graphs, CRFs rely on an undirected graph, known as a Markov random field. The simplest variant of the first-order CRFs is known as linear-chains, which is equivalent to the first-order HMMs. The first-order CRF can be obtained by defining a first-order Markov random field over the output $\mathbf{y}$ where it is conditioned on the output globally on the input sequence $\mathbf{y}$. To estimate the model’s parameters, several iterative algorithms are employed:

1. *Structured Perceptron Algorithm* (Collins, 2002): This algorithm iteratively trains on tagged instances using the Viterbi search algorithm to correct prediction errors by updating model parameters through simple addition and subtraction.
2. *Conditional Maximum Likelihood Estimation* (Cox & Reid, 1987): This technique formulates parameter estimation as an optimisation problem, seeking to solve the maximum likelihood by minimising a function with a unique solution. The iterative

scaling method used in the original work of Lafferty *et al.* (2001) remains one option, while more efficient methods like gradient descent have been explored in the literature (Vishwanathan *et al.*, 2006).

While these methods aim to estimate the model parameters with precision, exact estimation is often computationally expensive or infeasible. As a result, alternative parameter estimation approaches have been proposed, such as the Beam search (Vishwanathan *et al.*, 2006) and Pseudo-Likelihood Learning (Besag, 1975), which provide faster yet reasonably accurate solutions.

The work of Ruokolainen *et al.* (2013) implemented a supervised word segmentation approach using a small morphologically segmented dataset for languages including English, Arabic, Finnish, and Turkish, achieving superior results compared to state-of-the-art unsupervised methods (Poon & Domingos, 2008) and semi-supervised methods (Kohonen *et al.*, 2010).

One notable work that has received considerable attention for Nguni languages is that of Cotterell *et al.* (2016b), which developed a deterministic morphological segmentation system called CHIPMUNK. This system is based on a semi-Markov conditional random field (semi-CRF) model that integrates MMaxEnt principles, enabling the use of feature-rich representations. CHIPMUNK was tested on six languages, including isiZulu, using the *Ukwabelana* dataset. The CRF approach is particularly well-suited for tasks involving surface segmentation, as demonstrated by Moeng *et al.* (2022), who employed a feature-based CRF model to achieve effective surface segmentation across four Nguni languages.

However, despite its popularity for word segmentation in low-resource languages, the CRF-based method has some drawbacks. One key challenge is its high computational complexity, particularly during the training phase. As the feature space expands with textual data, the model’s complexity increases, resulting in slower training times and higher memory consumption (Lafferty *et al.*, 2001). Another drawback is the need for manual feature engineering, which often requires linguistic expertise to extract useful features for the model. This need for specialised knowledge can limit the scalability and the adaptability of CRFs to other languages without a similarly rich linguistic background (Chen *et al.*, 2015).

Alternative Statistical and Supervised-Based Methods

Apart from the graphical models discussed above, such as Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs), and Conditional Random Fields (CRFs), several other machine-learning techniques have been successfully applied to the task of supervised morphological segmentation. These techniques include but are not limited to, Memory-Based Learning, Support Vector Machines (SVMs), and Tree-Based Models.

In Puttkammer and Du Toit (2021), the model developed for canonical segmentation and syntactic morpheme tagging of Nguni languages utilises the TiMBL (Tilburg Memory-Based Learner) to approach the segmentation task. TiMBL, as described in Daelemans *et al.* (2004), transforms segmentation into a classification problem using the k-nearest neighbours (k-NN) algorithm along with a weighting algorithm. This enables the model to classify morphemes by leveraging stored data instances and making predictions based on similarities between input data and previously seen instances.

TiMBL operates as a memory-based learner, meaning that it stores instances of labelled data and compares new data points to the stored examples to predict their segmentation. This method is particularly useful for handling low-resource languages, where large annotated datasets may not be available. Prior to this, Quasthoff *et al.* (2016) also employed TiMBL for segmenting both isiZulu and German. Despite the stark contrast in resource availability, 8 000 words were labelled for isiZulu versus 70 000 for German, where the TiMBL achieved over 96% accuracy in both languages. This demonstrates the model’s ability to perform well in both high-resource and low-resource settings.

General Limitations Across Supervised Methods

Using statistical machine learning for supervised morphological segmentation has several limitations. A primary issue is overfitting, often due to limited labelled data or poor model design, leading to models capturing noise rather than generalisable patterns, which reduces their performance on new data. Feature extraction also poses challenges, as it typically requires linguistic expertise to engineer effective features, and thus limiting scalability across languages with diverse morphological structures. Additionally, dataset quality is critical; inconsistencies or errors in labelled data can hinder model accuracy, especially in low-resource settings (Wang *et al.*, 2016b). Addressing these limitations requires careful data preparation and robust feature engineering - including deep learning for automatic feature extraction, as well as thoughtful model design in order to improve generalisation.

3.4.2.2.2 Semi-Supervised Morphological Segmentation

This section concludes the discussion on statistical machine learning approaches by presenting a semi-supervised method, which combines both supervised and unsupervised methods into one technique (Spiegler *et al.*, 2008). This method uses labelled and unlabelled data to train a machine learning strategy. This strategy is useful when dealing with low-resourced languages and where there is typically little to no label dataset available to fully take advantage of the supervised methods.

Kohonen *et al.* (2010) explored semi-supervised learning of concatenative morphology using the MBL model (Creutz & Lagus, 2002a). In their approach, a model is trained to identify morphemes through maximum a posteriori (MAP) estimation. The key idea in this method is that the model leverages labelled data to directly inform its analysis of specific morphological patterns. For words present in the labelled data, their segmentations are fixed, which guides the model’s learning process. The unsupervised portion of the data plays a salient role, as the model seeks to generalise morphological patterns from the large amount of unlabelled data it processes.

The work of Shalnova and Golénia (2010) presents a novel semi-supervised method for morphological segmentation that effectively combines supervised learning model Stem Extraction algorithm (SSE) for stem extraction with unsupervised learning GBUMS (Graph-Based Unsupervised Morpheme Segmentation) for affix segmentation. This approach addresses the challenges posed by under-resourced fusional and agglutinating languages, isiZulu and Finish, by maximising the utility of limited annotated data and exploiting patterns in unannotated corpora.

In the rule-based approach section, a “DEAP: Deductive-Abductive Parsing for morphological analysis” (Spiegler, Golénia, & Flach, 2010) was mentioned, which also looked at languages with complex morphology, namely Arabic and isiZulu. Since their system is hybrid, one can also categorise this system as being semi-supervised despite using a rule-based deductive strategy to extract rules that inform the unsupervised model, blending the two methodologies into a hybrid system for morphological analysis.

3.4.3 Deep Learning Approaches

While conventional statistical machine learning techniques have demonstrated promising results, they often require careful feature engineering and domain expertise to extract the relevant features needed to represent data as feature vectors for model training. Similar to rule-based systems, this process can be expensive, time-consuming, and sometimes challenging to mathematically capture in a way that can be effectively used by models. In some cases, essential features are difficult to quantify or express in a form conducive to machine learning.

In response to these limitations, a subclass of machine learning known as *deep learning* has gained significant popularity and preference over traditional statistical-based methods. Deep learning models leverage a powerful technique called *representation learning*, which enables machine learning algorithms to discover both the representation of data and the mapping from input to output automatically (Goodfellow *et al.*, 2016). This approach eliminates the need for extensive manual feature engineering, as deep learning models can automatically learn useful features from raw data in a relatively short period. This

not only improves efficiency, but also reduces the time and cost associated with model development.

Deep learning’s ability to learn hierarchical representations of data has revolutionised many fields, particularly NLP (Mikolov *et al.*, 2011; Schwenk, 2012) and computer vision (Krizhevsky *et al.*, 2012; Shalounova & Golénia, 2010), where extracting useful features manually can be incredibly complex. As such, considering some languages may have complex morphology, deep learning has also gained popularity in solving the task of word morphological segmentation or decomposition.

In this section, a brief survey of studies that have employed deep learning techniques for the task of word segmentation is presented, focusing on supervised, unsupervised, and semi-supervised methods. These approaches, leveraging the representation learning capabilities of deep neural networks, have demonstrated significant improvements in tasks involving morphologically rich languages, reducing the need for manual feature extraction and simplifying the modelling process.

3.4.3.1 Sequence-to-Sequence Modelling

The problem of word segmentation is commonly viewed as a sequential problem, where the input and output are presented as a sequence of words or characters. One popular method used to solve such problems is the use of sequence-to-sequence (seq2seq) models, as proposed by Sutskever *et al.* (2014) and Cho *et al.* (2014). These models rely on a framework known as a *deep neural network* (DNN), which consists of two key components, namely the encoder and the decoder. Figure 3.3 provides a simple illustration of the encoder-decoder architecture, showcasing how the input is transformed into an output sequence through these two stages.

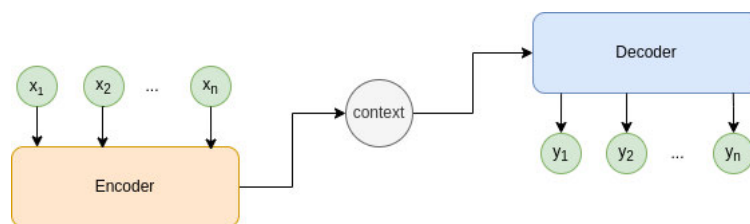


Figure 3.3: A simple diagram of the encoder-decoder.

Encoder-Decoder

1. *Encoder*: This part of the deep learning framework reads and processes the input sequence. The encoder accepts a source sequence of varying length, denoted as:

$$\mathbf{x} = (x_1, x_2, \dots, x_{T_x}),$$

and produces a fixed-length vector known as the context vector ($\mathbf{c}$). This vector serves as a summary representation of the input sequence, computed based on the hidden states ($\mathbf{h}_t$) at each time step t (where $t = 1, 2, \dots, T_x$). The hidden state is updated recursively using a function f , which takes as input the current token (x_t) and the previous hidden state ($\mathbf{h}_{t-1}$) (Niu *et al.*, 2021):

$$\mathbf{h}_t = f(x_t, \mathbf{h}_{t-1})$$

After processing all T_x tokens (or time steps) in the input, the encoder aggregates the hidden states into a single fixed-length context vector $\mathbf{c}$ via a function q :

$$\mathbf{c} = q(\{\mathbf{h}_1, \dots, \mathbf{h}_{T_x}\})$$

where f and q represent non-linear activation functions commonly used in deep learning, such as sigmoid (Cybenko, 1989), tanh (Chen & Chang, 1996), and rectified linear unit (ReLU) (Nair & Hinton, 2010).

The non-linearity introduced by these activation functions is crucial in multi-layered networks, as it ensures that the model can learn complex hierarchical representations rather than behaving as a simple linear transformation of inputs. Without non-linearity, a deep network would collapse into an equivalent single-layer model, limiting its ability to capture intricate patterns in data (Jurafsky & Martin, 2025).

2. *Decoder*: The decoder is responsible for generating the output sequence of varying length in an autoregressive manner, and it is mathematically represented as:

$$\mathbf{y} = (y_1, y_2, \dots, y_{T_y}),$$

It leverages the context vector $\mathbf{c}$ (which encapsulates the input sequence information), along with the previously generated output token y_{t-1} and an initial hidden state (often derived from the encoder's final hidden state $\mathbf{h}_{T_x}$) (Niu *et al.*, 2021). At each time step t (with $t = 1, 2, \dots, T_y$), the decoder updates its hidden state using a function f :

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, y_{t-1}, \mathbf{c})$$

This function allows the decoder to map the encoded context information to the target sequence and progressively generate tokens one by one up to the output length T_y . Since each computation of the hidden state $\mathbf{h}_t$ depends on the fixed context vector $\mathbf{c}$, the decoder effectively maintains alignment with the input sequence while constructing the output (Jurafsky & Martin, 2025).

Feedforward Neural Network (FNN)

Figure 3.3 reflects the simplest form of the encoder-decoder model, often called a *feed-forward neural network* (FFN). In this configuration, the input parameters are treated as independent of each other, which forces each position in the sequence to learn its features independently using a fixed-sized context vector $\mathbf{c}$ (Jurafsky & Martin, 2025).

As an instance of supervised machine learning, the FFN aims to predict an output $\hat{\mathbf{y}}$ given an input $\mathbf{x}$, while being trained on a ground-truth target y . The FFN learns to approximate the true y by optimising parameters such as weights ($\mathbf{W}_i$) and biases ($\mathbf{b}_i$) for each layer i . The network approximates y by optimising parameters such as *weights* ($\mathbf{W}_i$) and *biases* ($\mathbf{b}_i$) at each layer i .

This optimisation is achieved using a *loss function*, which quantifies the discrepancy between the model's predictions ($\hat{\mathbf{y}}$) and the true values ($\mathbf{y}$) (Jurafsky & Martin, 2025). During training, the model iteratively updates its parameters to minimise the loss function. This is accomplished through the *backpropagation algorithm* (Rumelhart *et al.*, 1985), which computes the gradient of the loss function with respect to the weights and propagates the error backwards through the network, enabling efficient parameter updates and improved learning.

While effective for certain tasks, the FFN architecture lacks the ability to capture temporal or contextual dependencies between different time steps. This limitation becomes particularly evident in tasks such as machine translation, where meaning is often distributed across multiple words in a sequence (Goodfellow *et al.*, 2016). However, since FFN processes each input independently, it struggles to model long-range relationships within sequences.

3.4.3.1.1 Recurrent Neural Networks (RNNs)

To address this limitation, a more advanced framework known as the *Recurrent Neural Network* (RNN) (Rumelhart *et al.*, 1986) was introduced. Unlike FFN, which processes each input independently, RNNs share the same weights or parameters across different time steps, introducing a feedback mechanism that allows the model to retain information from

previous steps. This recursive structure enables the model to retain contextual information and learn dependencies across time, making it particularly well-suited for sequential data.

Despite the benefits of sharing weights in a recurrent manner, *RNNs* still has limitations in handling certain types of sequential problems. Specifically, they present difficulty with tasks where the current input does not only depend on past inputs $0, \dots, t-1$, but also on future inputs $t+1, \dots, t_n$. For example, in a machine translation problem, a word at time step t might have multiple possible meanings, and its correct interpretation may depend more on subsequent words rather than on the preceding ones. In this case, relying solely on past inputs will not help the model to disambiguate the meaning of the word at t .

To address this challenge, *Bidirectional Recurrent Neural Networks* (BiRNNs) have been introduced by Schuster and Paliwal (1997). BiRNNs process the input sequence in both directions, combining two separate RNN layers, one that moves *forward* in time, from the start of the sequence to the end, and another that moves *backwards* in time, from the end of the sequence to the beginning. This way, the output y_t at each time step t is informed by both past and future context, providing a more comprehensive understanding of the sequence. The BiRNN consists of two hidden states for each time step t (Niu *et al.*, 2021):

1. $\vec{\mathbf{h}}_t$: The hidden state generated by the forward RNN, which processes the input sequence from the start to the end.
2. $\overleftarrow{\mathbf{h}}_t$: The hidden state generated by the backward RNN, which processes the input sequence from the end to the start.

These two hidden states are concatenated at each time step to form a combined hidden state:

$$\mathbf{h}_t = [\vec{\mathbf{h}}_t^T, \overleftarrow{\mathbf{h}}_t^T]^T$$

RNNs operate using a technique called *backpropagation through time* (BPTT) (Werbos, 1990), an extension of standard backpropagation used in FNN. Since RNNs process sequential data by maintaining hidden states over time, BPTT unrolls the network over multiple time steps and computes weight updates across all previous states in the sequence. Unlike standard backpropagation, which only propagates gradients backwards through layers, BPTT propagates gradients both through the network layers and across time, ensuring that past inputs influence parameter updates for future predictions (Kag & Saligrama, 2021).

In the training of RNN systems, two key processes are involved: *forward propagation* and *backward propagation*. During the forward pass, the input sequence is processed one step at a time, with each time step updating the hidden states based on previously computed

states. The loss function is computed at the final output layer (or at each time step, depending on the setup). In the backward pass, gradients are computed with respect to the model's parameters, weights (W) and biases (b), and then propagated backwards through all previous time steps (Sherstinsky, 2020). Since RNNs share the same weights across all time steps, BPTT updates these parameters globally, allowing the network to adjust based on long-term dependencies within the sequence. Initially, weight parameters are typically initialised with small random values, while biases are often set to zero (Montavon *et al.*, 2012). Through multiple iterations, the model gradually minimises the loss function, improving performance on sequential tasks.

However, during this training process, two significant issues can arise due to the recursive nature of weight multiplication across different time steps:

1. *Vanishing Gradient Problem*: This occurs when the gradients (i.e., the partial derivatives of the loss with respect to the weights) become very small as they are propagated back through the network (Narkhede *et al.*, 2022). As a result, the weight updates become negligible, and the model struggles to learn long-range dependencies because the information from earlier time steps vanishes. Over time, the gradients approach zero, effectively preventing the model from learning from distant past inputs.
2. *Exploding Gradient Problem*: This is the opposite issue, where the gradients become excessively large as they are propagated back. When the weights are too large, the gradients grow exponentially during backpropagation, causing the model to become unstable (Goyal *et al.*, 2018). The values can explode towards infinity, leading to divergent behaviour during training, making it impossible to converge to a meaningful solution.

Long Short-Term Memory and Gated Recurrent Unit

These two problems significantly hamper the performance of RNNs, particularly when trying to learn long-term dependencies in sequential data. To address the limitations of Recurrent Neural Networks (RNNs), particularly the vanishing and exploding gradient problems, two advanced adaptations of RNNs were introduced: Long Short-Term Memory (LSTM) (Hochreiter, 1997) and Gated Recurrent Units (GRU) (Cho *et al.*, 2014). These architectures provide mechanisms for better handling long-term dependencies in sequences while avoiding the common issues associated with basic RNNs.

In contrast to the classical RNN, the LSTM introduces a more complex internal architecture. It incorporates self-loops (internal recurrences) that allow gradients to flow through the network over long time steps, reducing the likelihood of the vanishing gradient problem. The LSTM architecture uses gates to regulate the flow of information. There are three such gates, namely: input gate, forget gate and output gate. The input gate determines

which information is added to the current state, the forget gate decides what to discard from the previous state, and the output gate controls which part of the cell state is used as output for the next cell (Goodfellow *et al.*, 2016). These mechanisms allow LSTMs to selectively retain relevant information while discarding unnecessary details, making them effective for processing long sequences.

On the other hand, the GRU as proposed by Cho *et al.* (2014), is a simplified variant of LSTM. The GRU also employs a gating mechanism, but with fewer components, combining the forget and input gates into a single gate. This single gate controls the flow of both new information and past information. Unlike LSTM, GRU does not have a separate memory cell, thus simplifying its structure. It has the *update gate*, which determines how much of the previous hidden should be retained and the *reset gate*, which decides how much of the previous state should be forgotten or reset for the next step.

The key difference in this architecture is that LSTM has a dedicated memory cell and separate gates for controlling input, output, and forgetting, making it larger and more powerful for handling complex dependencies in large datasets (Goyal *et al.*, 2018). GRU, with its fewer parameters (no memory cell and fewer gates), is computationally more efficient, and can outperform LSTMs on smaller datasets where model size and training time are crucial.

The LSTM and GRU methods significantly improve the ability to maintain long-term dependencies compared to classical RNN models. These methods can also be combined to form bi-models such as *Bidirectional LSTM* or *Bidirectional GRU*, which allow the models to process information both forward and backwards through the input sequence. While these RNN variants successfully capture contextual information through a context vector $\mathbf{c}$, which encodes the semantic details as a fixed-length vector, they do have limitations when handling long input sequences.

Bahdanau (2014) highlighted this issue, noting that encoding input sequences $\mathbf{x}$ of varying lengths into a fixed-length context vector $\mathbf{c}$ compresses the input information, which degrades performance as the sentence length increases. This fixed-length encoding forces the model to “squeeze” important details into a limited space, making it difficult to preserve all the necessary information for long sequences, such as in translation tasks.

To address this problem, Bahdanau (2014) introduced the attention mechanism, which enables the context vector $\mathbf{c}$ to have a variable length, dynamically adjusting to the input sequence. Instead of relying solely on a single fixed-length context vector, the *attention mechanism* allows the model to focus on different parts of the input sequence at each decoding step, improving the model’s ability to handle longer and more complex sentences.

In their work, Bahdanau (2014) proposed a BiRNN that integrates the attention mechanism. In this architecture, the *attention block* is added to the decoder, which now consists

of both an RNN and the attention mechanism. This attention block calculates the current context vector $\mathbf{c}_t$ at each decoding step by considering the concatenated hidden state $\mathbf{h}_t$ and the relevant parts of the input sequence, rather than relying solely on the final hidden state.

3.4.3.1.2 Transformer

In 2017, Vaswani *et al.* (2017) introduced a revolutionary architecture called the *Transformer*, which also leverages the attention mechanism, further enhancing the model’s ability to weigh different parts of an input sequence and thus capture contextual relationships more effectively (Vaswani *et al.*, 2017). The Transformer architecture quickly gained popularity due to its ability to model long-term dependencies in data without the recurrence constraints of RNNs, primarily through its *Self-Attention* mechanism. Among its various components, two have been particularly critical: the Self-Attention Mechanism and Multi-Head Attention, which are discussed in the next section (Sharma *et al.*, 2023).

Key Components of the Transformer Architecture:

1. *Self-Attention Mechanism:* This mechanism allows the model to assess the relevance of different tokens in the input sequence, regardless of their position. Each token attends to every other token, enabling the model to capture dependencies between words that are far apart in the sequence, without the limitations of sequential processing seen in RNNs.
2. *Parallel Processing with Multi-Head Attention:* Another significant advantage of the Transformer architecture is its ability to process sequences in parallel. This is achieved through the Multi-Head Attention mechanism, which splits the attention process into multiple “heads”. Each head attends to different parts of the sequence independently, allowing the model to capture various aspects of the input sequence simultaneously. This dramatically improves the model’s computational efficiency, allowing it to handle large datasets and longer sequences more effectively than with RNNs.

Unlike traditional RNN-based frameworks, the Transformer moved away from sequential processing, despite retaining the common sequence-to-sequence structure of encoder-decoder models. In the Transformer, both the encoder and decoder components remain essential: the encoder processes the input, while the decoder generates the output; however, its operational framework is distinct and more efficient.

The primary role of the encoder in the Transformer architecture is to extract features from input sentences using multiple layers automatically. A key component in this process is the self-attention mechanism, which assigns different weights to different parts of the

input sequence without regard to their position. Another crucial feature that makes the Transformer preferable is its ability to perform parallel processing through the *Multi-Head Attention Mechanism*, which splits the attention process into multiple independent attention heads, allowing the model to capture diverse contextual information efficiently. Further discussions of the Transformer architecture are provided in Chapter 5.

3.4.3.2 The Deep Learning Morphological Segmenters

Wang *et al.* (2016a) presented a series of supervised morphological segmentation models based on LSTM neural networks. Their approach is based on learning to predict morpheme boundaries directly from raw input characters, without relying on hand-crafted linguistic features. The authors introduced three novel architectures: Window LSTMs, Multi-Window LSTMs, and Bidirectional Multi-Window LSTMs. These models use fixed-width character windows to capture local context and to learn sequence structures automatically. The authors evaluated their models on Hebrew and Arabic datasets, demonstrating competitive performance compared to feature-rich supervised methods, especially with limited training data.

Le *et al.* (2022) employed a deep learning-based model for the morphological segmentation of Innu-Aimun, an indigenous language of Canada. Their method uses a Transformer-based encoder-decoder architecture with a multihead self-attention mechanism, incorporating pre-trained embeddings at character and affix levels. The authors manually collected a small corpus and created a golden test set using native speakers. Their model, T-DeepLo, was then evaluated against several baseline approaches, including Finite-State Transducer (Richardson & Tyers, 2021), Morfessor (Virpioja *et al.*, 2013), and Adaptor Grammar-based methods (Eskander *et al.*, 2020). The proposed deep learning model outperformed all baselines, demonstrating promising results for morphological segmentation in this low-resource, polysynthetic language scenario.

Similarly, Premjith *et al.* (2018) presented a deep learning-based strategy for the segmentation of Malayalam texts, which is an agglutinative language. The authors created a large annotated dataset by inflecting nouns and verbs. They then preprocessed the data by romanising Malayalam words (converting the Malayalam writing system to Roman script), segmenting them into syllables, and marking morpheme boundaries. This was then presented as a sequence labelling problem, at which they developed three systems based on RNNs, LSTM networks, and GRUs. The models were designed to automatically learn the rules for identifying morpheme boundaries and separating them from the original words. This process allowed for a better analysis of the grammatical structure of the words, i.e., morphological analysis. The RNN, LSTM, and GRU models achieved high accuracy rates, with the best performance reaching 98.16% using GRU.

In the work of Jamtsho and Muneesawang (2020), the segmentation of Dzongkha, the national language of Bhutan which lacks explicit word delimiters was formulated as a supervised syllable tagging problem. The authors implemented a Deep Neural Network (DNN) and a Bi-LSTM model, where the DNN uses a window approach to capture contextual information, and the Bi-LSTM further enhances performance by capturing dependencies across syllables. Pretrained syllable embeddings were used, and the best-performing model was the Bi-LSTM, which achieved an accuracy of 95.25%.

Closing with recent models that look at the segmentation of isiZulu with deep learning techniques, attention is paid to the work of Moeng *et al.* (2022), which presented a comprehensive study on morphological segmentation for Nguni languages (isiZulu, isiXhosa, isiNdebele and Siswati), investigating both supervised and unsupervised approaches. For supervised canonical segmentation, the authors applied two seq2seq models i.e. Bi-LSTM with attention mechanism, and Transformer-based architectures, with a Transformer architecture achieving the best performance (average F1 score of 72.5% across four languages). For supervised surface segmentation, the authors employed two models based on CRF (Lafferty *et al.*, 2001), which are the *Feature-based CRF* and *Bi-LSTM CRF*. Here the feature-based CRF outperformed neural approaches, reaching an average F1 score of 97.1%.

In the unsupervised setting, Moeng *et al.* (2022) presented bi-LSTM that uses entropy for character-level language modelling. This dual approach allowed the model to capture contextual information from both sides of the word, which is a hallmark of bi-directional architectures. However, unsupervised methods performed poorly as compared to Morfessor (Creutz *et al.*, 2007b), which obtained an average F1 score of 27.69% in all languages while their method obtained an average of 25.73%.

All these models were evaluated using precision, recall, and F1 scores, treating segmentation as morpheme identification rather than mere boundary detection. The scores that were found for seq2seq models demonstrated their capabilities to handle the complex morphological structure of Nguni languages effectively.

Liu and Prud’Hommeaux (2022) presented a study on model generalisability for morphological segmentation in cross-linguistic low-resource scenarios. They emphasised the challenges posed by typical model evaluation designs that assume a monolingual focus, which can lead to misleading conclusions about model performance, especially when data sets are small or unrepresentative. In this work, the authors compared three classes of supervised models, i.e., Morfessor (Kohonen *et al.*, 2010) an original semi-supervised model, a seq2seq bi-GRU with attention mechanism and Order- k CRF, a CRF model that takes into consideration the previous k characters.

The authors used 11 languages from six different language families, assessing how well these models generalize when exposed to new data sets. IsiZulu was part of these languages, and

they used Ukwabelana data for its segmentation. The finding of Liu and Prud’Hommeaux (2022) demonstrated that model performance and rankings based on a single small dataset often fail to generalise, with considerable variability in results across different samples. In this work, they also investigated factors influencing model generalizability, where they found that morpheme overlap and the average number of morphemes per word between training and test sets were more influential than dataset size alone.

Meyer and Buys (2022) explored the technique of using a *subword segmental language model* (SSLM) that jointly learns subword segmentation and language modelling in an unsupervised manner. Their model uses a neural architecture combining LSTMs and a lexicon to learn to segment words into subwords while being trained as an autoregressive language model. They evaluated their approach in four Nguni languages of South Africa. The authors compared the SSLM to baseline models using character, BPE, and *unigram language model* (ULM) tokenisation on both language modelling and unsupervised morphological segmentation tasks. The SSLM outperformed baselines on language modelling across the four languages on average. For morphological segmentation, a word-level variant of their model outperformed existing unsupervised methods by a large margin.

Recently, Pranjić *et al.* (2024) introduced LLMSegm, a novel supervised approach for surface-level morphological segmentation using large language models (LLMs). They formulated the task as a binary classification problem, where the model predicts whether a morpheme boundary exists between each pair of adjacent characters in a word. The authors leveraged the *Glott500-m* LLM, which is pre-trained on more than 500 languages, to fine-tune it on labelled morphological segmentation data. Their method encodes each input word twice, once as the original word and once with a potential morpheme boundary inserted in order to utilise both the LLM’s subword tokenisation and explicit boundary information.

LLMSegm was investigated across seven morphologically diverse languages, of which four form part of Nguni languages and are low-resourced and three are English, Finnish and Turkish, where, except for English, all have agglutinative morphology. Which outperformed several baseline methods in both semi-supervised settings (using only 900 labelled examples for the three languages) from Morpho Challenge 2010 (Kurimo, Virpioja, & Turunen, 2010) and for low-resource languages from NCHLT (Eiselen & Puttkammer, 2014b). The approach achieved improved F1 scores and accuracy compared to competing methods on 6 out of 7 datasets.

3.5 Morphological Segmentation Metrics

In this section, various evaluation methods that are commonly used in evaluating morphological segmentation systems are perused. Where the evaluation of basic NLP processes or tasks such as morphological segmentation or analysis, parsing, lemmatisation, word embedding and the like, which are commonly used to enhance performances of other NLP tasks such as translation systems, information retrieval, and so forth, it is common for their performance to be divided into two evaluation criteria: *intrinsic* (direct) evaluation or *extrinsic* (indirect) evaluation.

Within the intrinsic evaluation method, the system’s output (predicted morphemes) is directly assessed in terms of predefined standards or criteria that relate to the system functionalities or objectives (Jones & Galliers, 1995; Resnik & Lin, 2010). In this evaluation method, the output generated by the system can either be manually evaluated or can be evaluated using external data, which is referred to as a reference set or ground truth. This form of evaluation is usually not sufficient to yield a good overview of the holistic performance of the developed process or system.

Since these morphological segmenters are created to enhance the performance of downstream NLP tasks, the measure of the quality of the morphological segmenters in relation to the impact it has on the effectiveness of the downstream task is referred to as extrinsic evaluation (Resnik & Lin, 2010). In this evaluation strategy, the morphological segmentation task is treated as an *enabling technology*, whose sole responsibility is its contribution to a downstream application.

Both evaluation methods are essential to comprehensively assess a segmentation model’s effectiveness. Several studies have shown that strong performance in intrinsic evaluation does not always translate to better results in extrinsic tasks, and vice versa (Chiu *et al.*, 2016; Jones & Galliers, 1995). Therefore, conducting both types of evaluations ensures a holistic assessment of a segmentation model’s capabilities and its practical usability in real-world applications.

3.5.1 Intrinsic evaluation

Within intrinsic evaluation, the evaluation process can vary depending on whether the system is supervised, unsupervised, or semi-supervised. In the supervised model, the output is simply a list of morphemes or substrings divided by morpheme boundaries, which can be evaluated by directly comparing them to a pre-generated (constructed) reference list (ground truth). However, with unsupervised models, where the data is unlabelled, there is no straightforward correspondence between the generated morphemes and the reference labels. This challenge is also present in semi-supervised systems, which only

have partial labelling. As a result, many systems have sought to establish an organised method to map the generated morphemes to those in the reference set.

Spiegler and Monson (2010) outlined several computational requirements for an effective automatic metric in morphological analysis. These include correlating with real-world NLP tasks, being readily computable to avoid excessive time costs, being robust against manipulation while reflecting the true morpheme distribution, and being readily interpretable, with scores that directly relate to the quality of the morphological analysis. Although these criteria were initially proposed for morphological analysis, they are equally relevant for evaluating the effectiveness of morphological segmentation, given that segmentation is a fundamental component of morphological analysis.

Since the present study takes a supervised route, it explores the evaluation metrics that are commonly used for this task in the literature, which are classification metrics (F1 score, recall and precision), Bilingual Evaluation Understudy (BLEU) (Papineni *et al.*, 2002) and CHaRacter-level F-score (Popović, 2015).

3.5.1.1 Classification Metrics

Classification metrics entail the commonly used metrics “but poorly motivated” ways to assess the performance of classification models, especially in the context of information retrieval and binary classification tasks (Powers, 2011). The commonly used measures within this category are recall, precision, and F-score (F1 Score), which help to assess the accuracy of classification models.

In the case of binary classification, it allocates input objects, either negative (-) or positive (+), based on the relevant feature(s) being investigated. However, classification systems are prone to errors, particularly in tasks such as medical diagnosis, spam filtering, and morphological segmentation. Two types of errors can occur: *Type I error* (False positive - FP); in this error, a system incorrectly predicts a positive outcome when the actual result should be negative. Similarly, a *type II* (False Negative - FN) error occurs when a system wrongfully predicts a negative outcome when the actual outcome is positive. In the classification task, the true outcomes (ground truth) and prediction can be viewed in four different ways, as illustrated in Table 3.3. Here is a brief explanation of these outcomes in relation to segmentation tasks:

Table 3.3: Contingency table: confusion matrix.

Predicted	Actual	
	Positive	Negative
Positive	True Positives (TP)	False Positives (FP) (Type I Error)
Negative	False Negatives (FN) (Type II Error)	True Negatives (TN)

1. *True Positives (TP)*: Correctly predicted positive outcomes (correctly segmented morphemes in segmentation tasks).
2. *True Negatives (TN)*: Correctly predicted negative outcomes (correctly identified non-morpheme boundaries).
3. *False Positives (FP)* (Type I Error): This is an instance where the Model incorrectly identifies a boundary or segmentation where there is none. model predict: *u-m-fana* instead of *u-mfana*
4. *False Negatives (FN)* (Type II Error): Instances where the Model fails to identify an existing morpheme boundary or segmentation. model predict: *umfana* instead of *u-mfana*

From these different counts, three metrics have been introduced, namely the precision, recall, and F1 score.

1. **Precision**, this measures a ratio of true positive results out of all the predicted positives. This is a *FP* (Type I Error).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

2. **Recall**, also referred to as *sensitivity*, measures the ratio of true positives out of all the predicted positives. This simply reflects the *FN* (Type II Error)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3. **F-score**, this measures the harmonic average between recall and precision and gives a balanced measure of these two metrics.

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

With *precision*, one seeks to understand in all the instances that the system has classified as positive how many are correctly classified. In the *recall*, the question is, in all those instances which are true positives, how many were correctly classified? It is worth mentioning that these metrics do not evaluate the order of the segments produced by means of a model, which is a crucial aspect of subword modelling, especially for generative models. Instead, they treat morpheme segmentation as a *set-based* task, where the presence or absence of elements (morphemes) is evaluated without accounting for their positional arrangement. In a nutshell, they focus on comparing predicted morphemes (or boundaries) with the true ones, ignoring the sequence or position of morphemes. Below is an example, where a model can give full points:

Consider the isiZulu word “ngizokubona,” which means “I will see you.” Correct segmentation (ground truth): *ngi-zo-ku-bon-a*. If, for example, the model predicts “ku-zo-ngi-bon-a” which gives “*kuzongibona*” (it will see me) a valid isiZulu word. These will be given a perfect score even though they are wrong.

3.5.1.2 N-gram Based Evaluation Metrics

3.5.1.2.1 Bilingual Evaluation Understudy (BLEU)

BLEU is a precision-based evaluation metric that implements a modified version of *n-gram* (a sequence of “n” amount of elements) that measures the overlapping of words between the candidate set and reference(s) set (Papineni *et al.*, 2002). This metric was originally developed for translation systems to calculate the precision of matched n-grams between the predicted and reference sentences. With the n-gram, it allowed the metric award systems that adhere to the order of words. In the calculations, BLEU also includes a *brevity penalty* to adjust for length differences between predicted and reference sentences.

In the task of segmentation, morphemes are produced as a sequence of strings that are divided by morpheme boundaries, as compared to white spaces in sentences. Such similarities make it possible for one to use BLEU as a way to intrinsically assess the performance of word morphological segmentation more in particular when you are experimenting with agglutinative language like isiZulu, where a word can have several morphemes that can be split like a sentence and whiles approach was first used for unsupervised word segmentation, while maintaining the morpheme order which is necessary. Where there are also unlike binary boundary-based metrics, it can give partial credit for partially correct segmentation. The BLEU mathematical formula is defined in terms of the Brevity Penalty (BP). which is as follows:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

where:

1. c = length of the candidate segmentation.
2. r = effective reference length (closest reference length to the candidate)

Where the BLEU is finally formulated as:

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log(\text{precision}(n)) \right)$$

where:

1. - N = maximum n-gram length (commonly 4)
2. - w_n = weights for each precision (often set to $\frac{1}{N}$)

The BLEU score ranges from 0 to 1, where a score closer to 1 signifies better segmentation, while a score of 0 indicates no matches with the reference segment.

3.5.1.2.2 CHaRacter-Level F-score (chrF)

The chrF follows the same logic as BLEU, except that the n-grams are considered at the character level (Popović, 2015).

1. Character N-gram precision:

$$\text{Precision} = \frac{\text{Number of matched character n-grams}}{\text{Total number of character n-grams in generated text}}$$

2. Character N-gram recall:

$$\text{Recall} = \frac{\text{Number of matched character n-grams}}{\text{Total number of character n-grams in reference text}}$$

F1 score Calculation: The chrF (similar to the F1 score above) score is derived from the harmonic average of precision and recall, hence using the F1 score formula as above, it results in:

$$\text{chrF} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

3. Character N-gram Length (n): The value of n can be adjusted to compute different character n-gram lengths (e.g., 1-gram, 2-gram, etc.), and the final chrF score can be calculated as a weighted average of the scores for different values of n . Similarly to BLEU above, the score also ranges from 0 to 1, where 0 means poor score and 1 means perfect score.

The above metrics are commonly used for supervised morphological segmentation, and are not very relevant in unsupervised and semi-supervised settings since the data set has no labels to use as ground truth. One way in which such learning can be evaluated is through the usage of methods that do not assess the models based on the reference morphemes and those generated, but based on the co-occurrence of morphemes of the words (Creutz *et al.*, 2007a), which was used in Evaluation of Morpho Challenge models, commonly referred as MC evaluation metric. The other method is that introduced by Spiegler and Monson (2010), known as *EMMA* (Evaluation Metric for Morphological Analysis), which works by

matching the predicted morphemes or morphemes-label pairs to those in the ground truth, using a one-to-one mapping. A variant of this metric is called *EMMA-2* (Virpioja *et al.*, 2011), which is optimised to be quicker by replacing the one-to-one mapping of morphs with two many-to-one assignment problems of morphemes. Another method that is of relevance in the unsupervised setting is *BPR* (Boundary Precision and Recall) (Grönroos *et al.*, 2014), which assesses the ability of a model to identify segmentation boundaries, through a direct comparison of the reference set and the predicted set.

3.5.2 Extrinsic Evaluation

The metrics outlined above evaluate the performance of the developed system based on predefined objectives. However, assessing the system in isolation is not sufficient, because morphological segmentation is typically implemented as an enabling technology. Therefore, to comprehensively evaluate and compare the performance of the developed models, this study extends the analysis by examining their impact on a translation system. This type of assessment is referred to as extrinsic evaluation, which provides a holistic overview of how effective the developed system is in terms of enhancing another NLP system.

3.5.2.1 Neural Machine Translation

With approximately 7 000 languages spoken worldwide (Leben, 2018), Machine Translation (MT) has become a vital tool for enabling communication across linguistic boundaries. MT systems facilitate the automatic translation of natural language sentences from one language to another. Over the years, MT has evolved significantly, transitioning from rule-based systems that relied heavily on linguistic knowledge, to data-driven approaches such as Statistical Machine Translation (SMT) (Brown *et al.*, 1990), and more recently, Neural Machine Translation (NMT) (Bahdanau, 2014; Cho *et al.*, 2014; Kalchbrenner & Blunsom, 2013).

The availability of large bilingual datasets has propelled the increased adoption of SMT and NMT. SMT systems leverage parallel corpora to learn phrase alignments and translation rules statistically. However, these systems often struggle to model long-distance dependencies effectively (Dwivedi *et al.*, 2025). The limitations of SMT gave rise to the era of NMT, which has since become the dominant paradigm in MT. Unlike SMT, NMT utilises continuous representations rather than discrete symbolic alignments (Srivastava *et al.*, 2018). NMT systems leverage a single, large neural network to model the translation process, taking advantage of deep learning’s capability to automatically learn features, unlike SMT systems that rely heavily on extensive feature engineering (Tan *et al.*, 2020).

Despite its advantages, NMT is highly data-dependent, requiring vast amounts of parallel

data to learn effectively. In resource-scarce settings, SMT has been observed to outperform NMT due to its lower data requirements (Koehn & Knowles, 2017). This limitation poses a significant challenge to the development of NMT systems for low-resource languages. Another notable drawback of NMT is its reliance on a fixed-sized vocabulary, which stores only the most frequent words. Words outside this predefined vocabulary are typically mapped to a generic *[UNK]* (unknown) symbol (Luong *et al.*, 2014). Consequently, any OOV word is represented as *[UNK]* and often decoded as an empty string, rendering the system ineffective at translating rare or unseen words.

In languages with complex morphology, such as isiZulu, Finnish, and Turkish, which exhibit agglutinative characteristics, a single stem can combine with multiple morphemes, leading to a combinatorial explosion of word forms. This poses a significant challenge for NMT systems because a small fixed vocabulary cannot represent all possible forms of roots and stems adequately. Consequently, the issue of data sparsity becomes even more pronounced, further limiting the system’s ability to effectively handle such languages.

To address the issue of data sparsity, particularly in NMT systems, researchers often employ subword tokenisation strategies that divide long words into smaller pieces. These strategies create an open vocabulary, allowing NMT systems to generalise better to unknown words (Salesky *et al.*, 2020). One approach in this regards is character-level tokenisation, which splits words into individual characters, offering a highly granular representation (Lee *et al.*, 2017). Another method is syllabification, where words are divided into syllables, capturing intermediate structures between characters and complete words (Kotzé & Wolff, 2015; Liu *et al.*, 2021). Additionally, Byte Pair Encoding (BPE), initially developed as a data compression method (Gage, 1994), was adapted for NMT by Sennrich (2015). BPE iteratively merges frequent character sequences to form subword units, effectively encoding rare words into more frequent subword patterns (Ding *et al.*, 2019; Salesky *et al.*, 2020; Wang *et al.*, 2020). Another approach is morphological segmentation, where words are segmented into morphs or morphemes based on linguistic knowledge, creating units with clear semantic or grammatical meaning that are then used in NMT systems (Chen & Fazio, 2021; Pan *et al.*, 2020; Tukeyev *et al.*, 2020).

Among the subword tokenisation methods, BPE has been the most widely adopted due to its unsupervised nature and few tunable hyper-parameters (Tan *et al.*, 2020). However, several studies have highlighted its limitations. For instance, Bostrom and Durrett (2020) noted that BPE often oversplits the roots of rare words, resulting in poor translations. Additionally, Tukeyev *et al.* (2020) observed that BPE struggles with morphologically rich languages, leaving many words unsplit and leading to larger vocabularies compared to morphological segmentation. Furthermore, BPE does not consider the morphology of a language during segmentation, which can degrade performance in language modelling tasks (Ataman *et al.*, 2017; Park *et al.*, 2021).

In contrast, morphological segmentation has been found to be a better alternative to BPE when morphologically annotated data is available. Studies by Mager *et al.* (2022) and Ataman *et al.* (2017) reported that morphological segmentation outperforms BPE in NMT tasks, especially for languages with complex morphology. Moreover, morphological segmentation has been shown to deliver better results than syllabification (Kotzé & Wolff, 2015). Character-level tokenisation, although simple, is generally not recommended, because it produces overly long sequences, which can lead to spurious word generation and suboptimal results in NMT systems (Lee *et al.*, 2017).

These findings highlight the importance of tailoring segmentation strategies to the linguistic characteristics of the target language for optimal NMT performance. To further explore this, the present study investigates the effect of morphological segmentation on an NMT system, specifically evaluating a translation system that maps morphologically segmented isiZulu sentences of varying granularities to word-based English sentences.

3.6 Chapter Summary

The objective of this chapter was to provide a comprehensive review of morphological segmentation. The chapter began with an exploration of the background and history of isiZulu in South Africa, examining the language in relation to its linguistic features, such as phonetics and phonology. Several key definitions of linguistic terms were introduced, laying the groundwork for the concepts discussed in later sections. Special attention was given to terms like “word”, as there is no universal consensus on its definition.

The next section delved into morphology and the various typologies of languages. IsiZulu, with its agglutinative structure and conjunctive orthography, was highlighted as having a particularly complex morphology. This morphological structure presents a significant challenge, leading to data sparsity, as a single stem can combine with various suffixes to form multiple words. Traditional corpora often fail to capture these variations, which negatively impacts the performance of language processing systems trained on such data.

To address this issue, morphological segmentation was discussed as a crucial solution, distinguishing between two main types: canonical and surface segmentation. The chapter provided a thorough review of different approaches to morphological segmentation, including rule-based methods, machine learning, and deep learning techniques. Finally, the chapter explored the metrics commonly used to evaluate segmentation performance, setting the stage for the next discussions in subsequent chapters.

Chapter 4

Data Preparation

This chapter provides an overview of the data preparation steps required to train supervised morphological segmenters. It has been established in the previous sections that annotated data is a requirement for one to create a supervised system, either for statistical machine learning or a deep learning system. When it comes to creating morphologically segmented datasets, two common approaches have been consistently used: manual annotation by language experts, and rule-based generation using predefined linguistic rules. In some cases, these methods are combined in a semi-automatic process to improve efficiency. However, manual and rule-based processes can be time-consuming and costly (Kituku, 2022; Swart, 2000), thus posing challenges for generating the large datasets typically required for high-performing models.

Regarding isiZulu, the development of morphologically segmented datasets has encountered similar limitations. As far as could be determined through an Internet search, only two public datasets for morphological segmentation in isiZulu are currently available: one is accessible online, while the other is not publicly available, but can be accessed by certain individuals or institutions.

This chapter aligns with the two objectives presented in Chapter 1, particularly *Sub-objective 1* and **Sub-objective 2**, which address the use of the ZRG to parse a large corpus of isiZulu text with a view of obtaining a treebank and to augment the treebank to obtain additional synthetic data. These feed directly to the primary objective of the study, namely, investigating a hybrid approach to morphological surface segmentation of the isiZulu text using synthetic data generation.

This chapter is organised as follows: Section 4.1 and 4.2 introduce the two datasets that are often used for morphological segmentation of isiZulu texts. Section 4.3 provides a reflection in terms of ZRG as a pipeline to produce the morphological surface segmented data, including a reflection on its limitations and strengths. Section 4.4 gives details

on the corpora used for creating the annotated data, while Sections 4.5, 4.6, and 4.7 outline the processes involved in transforming raw isiZulu texts into sentence data, covering parsing, linearisation, and data augmentation, respectively. Finally, Section 4.9 presents the chapter summary.

Additionally, all scripts, configurations, and data set details can be accessed in the GitHub repository¹ for further reference and reproducibility.

4.1 Ukwabelana Dataset

Spiegler, Van Der Spuy, and Flach, 2010 used fictional works and the isiZulu Bible to compile word lists and sentences which were used for the work known as ‘*Ukwabelana*’ (sharing), corpus a morphological corpus for isiZulu. The Ukwabelana corpus was developed as the first open-source morphologically analysed corpus for isiZulu with the aim to address the scarcity of computational linguistic tools for isiZulu. The morphological annotated words aimed to develop a morphological analyser.

4.1.1 Dataset Composition

In the Ukwabelana project, several corpora were produced, which have been used for several works in the literature (Cotterell *et al.*, 2015; Eskander *et al.*, 2016; Liu & Prud’Hommeaux, 2022; Mkhwanazi & Marais, 2024; Moukangwe *et al.*, 2014) and more. The annotated dataset included 10 000 lines of words, which were morphological surface segmented, and 3 000 POS-tagged sentences. The raw data comprised 100 000 unlabelled common isiZulu word types and 30 000 raw sentences.

Beyond the corpus itself, the project provided a morphological grammar for isiZulu, designed to generate hypotheses about possible word roots with the aim of producing parses that align with isiZulu grammar rules. Additionally, a semi-automatic morphological analyser and a POS tagger for morphologically analysed words were developed to further support the computational linguistic applications.

However, these resources are not publicly available online, because the previously provided URL is no longer functional. Despite this, recent works have continued to utilise some of these resources, indicating that they remain accessible within certain research communities or institutions.

¹https://github.com/Sthsha/supervised_surface_segmentation

4.1.2 Development and Annotation Process

The development process employed a semi-automatic annotation framework with the aim of improving the efficiency of annotation compared to manual annotation by experts. The system used a partial grammar written in (Definite Clause Grammars) DCG format, with approximately 240 morphological rules and 300 morpheme dictionary entries. This grammar focused primarily on closed morpheme categories, while leaving open morpheme categories such as noun and verb roots for hypothesis generation during the analysis process.

The parsing algorithm uses an abductive reasoning approach, working top-down through the rules and left-to-right through the text when matching dictionary entries. The algorithm could hypothesise unknown morphemes, particularly roots, when encountering previously unseen forms. This parsing algorithm generated multiple possible analyses for each word, which were in line with isiZulu grammatical structure, and which were then presented through a web interface for expert validation.

Through the Web interface, experts could review the generated hypotheses, select single or multiple correct parses, and provide manual analyses when the system failed to generate correct options. Once validated, the segmented text was compiled into the corpus, making it a foundational resource for training and evaluating computational models for isiZulu.

4.1.3 Limitations and Challenges

This project presented a sound start for developing morphologically analysed datasets for isiZulu, but the project also faced several challenges. The limited range of source materials restricted the variety of registers and vocabulary in the corpus. The grammar's coverage was incomplete, requiring ongoing updates and refinements. The parsing system often generated large numbers of hypotheses that needed manual filtering, and some words also required complete manual analysis. The use of multiple annotators also created challenges for maintaining consistency.

4.2 National Centre for Human Language Technology Text Corpora

In an effort to advance research and development (R&D) in South Africa, the National Centre for Human Language Technology (NCHLT) project was launched to create foundational language resources for South Africa's official languages, including isiZulu (Eiselen & Puttkammer, 2014b). This initiative was instrumental in supporting language technology

development for under-resourced languages. Additionally, the NCHLT project developed comprehensive text resources for ten South African languages, including isiZulu, encompassing collections of genre-classified texts, lexicons, frequency lists, and named-entity lists. These text corpora are important for various natural language processing tasks such as morphological segmentation, part-of-speech tagging, and machine translation, laying essential groundwork for researchers and developers working to expand language technology for South Africa’s under-resourced languages.

4.2.1 Dataset Composition

For isiZulu, the NCHLT corpus development involved several key steps. Initially, a significant portion of the data was sourced from government websites (gov.za) and documents, and a smaller portion was sourced from outlets such as news platforms, magazines, and scientific articles, ensuring a diverse linguistic representation. The project successfully created a substantial monolingual isiZulu corpus comprising 1.64 million tokens, along with a comprehensive type lexicon of 54 374 unique words and a named entity lexicon containing 11 634 entries. This foundation provided essential resources for computational linguistics work in isiZulu.

4.2.2 Development and Annotation Process

The project adopted a comprehensive four-layer annotation framework addressing token, orthographic, morphological, and morphosyntactic aspects of the language. At the token level, texts were segmented into paragraphs, sentences, multiword expressions, words, and punctuation. The orthographic layer focused on correcting errors, such as non-standard spellings, confusables, run-ons, and splits, while preserving the integrity of the original text. The morphological layer handled both lemmatisation and detailed morpheme-level analysis, and the morphosyntactic layer provided part-of-speech tagging along with grammatical function annotations (e.g., subject and object concords).

For isiZulu, characterised by its conjunctive writing style and rich morphological complexity, the segmentation process was implemented using a rule-based approach inspired by Bosch *et al.* (2006). This system recursively identified and stripped known affixes (covering grammatical classes such as relatives, negatives, concord classes, and derivational affixes) from each token. The remaining constituent was then verified against lexicons of valid roots and stems, with only those combinations accepted that were linguistically plausible.

The development process was systematically structured, with isiZulu as the prototype for all Nguni languages. Eiselen and Puttkammer (2014a) created detailed annotation protocols based on *Expert Advisory Group for Language Engineering Standards* (EAGLES)

standards and implemented a two-phase annotation process with expert verification. They utilised a specialised annotation environment called LARA II to ensure consistency and provide pre-generated morphological analyses for annotator selection. The final annotated corpus for isiZulu contained 44 324 tokens, all fully annotated across the four layers and aligned with parallel English texts.

4.2.3 Limitations and Challenges

The isiZulu morphological decomposer, while fully automatic, is based on a rule-based, recursive approach that faces some limitations. One immediate challenge is the inherent complexity of isiZulu’s agglutinative morphology and conjunctive orthography, which can result in multiple plausible segmentations for a single word. This ambiguity means that, in some cases, the system may produce alternative segmentations, and its overall accuracy is highly dependent on the completeness of the affix lists and the quality of the lexicon of valid roots and stems.

Additionally, while the decomposer systematically strips affixes and verifies the remaining constituent against linguistic resources, it sometimes fails to capture less frequent or irregular morphological patterns. During evaluation, Eiselen and Puttkammer (2014a) indicated that there remains more research to be conducted to improve morphological decomposition performance, particularly for the conjunctively written languages such as isiZulu.

Resource constraints also affected the overall process. The limited availability of annotated data and the need for expert verification meant that some outputs required manual review and correction, reducing the effective automation in practice and highlighting the necessity for ongoing refinement of both the rules and the underlying lexicons.

4.3 Reflection on ZRG’s Limitations and Advantages

In Section 3.4.1.3, an overview of ZRG was provided, presenting it as an implementation of isiZulu morphosyntax using the GF programming language. The ZRG leverages the GF runtime to parse text into syntax trees and linearise these trees into natural language. Additionally, the GF runtime’s orthography engineering capabilities allow for custom linearisation, systematically binding subwords together to form grammatically correct surface tokens. This makes it possible to parse an isiZulu token, obtain a syntax tree, and apply custom linearisation to produce its surface-segmented version.

However, the ZRG was not originally designed as a surface segmentation tool, but as a general-purpose parser and natural language generation system. While it can be adapted

for segmentation through GF’s orthographic engineering features and the use of specific binding tokens, it was not inherently optimised for this task. As a result, although parsing an isiZulu token to obtain a syntax tree and applying custom linearisation to generate a surface-segmented version can produce the desired result, this process presents certain challenges.

This section critically evaluates the limitations of the ZRG, including slow parsing performance, brittleness to input variations, and ambiguity in segmentation. It also draws parallels with challenges encountered in previous isiZulu segmentation projects, such as Ukwabelana and NCHLT, to contextualise ZRG’s performance. Additionally, it reflects on the advantages of the ZRG, including its broad linguistic coverage, and flexibility in output linearisation. This balanced assessment supports the rationale for using ZRG as an investigative tool for segmentation rather than assuming it to be a definitive solution.

4.3.1 Limitations of the ZRG

4.3.1.1 Slowness in Parsing

Despite its strong linguistic foundation, the ZRG suffers from efficiency issues. In practice, using the GF runtime with the isiZulu grammar can be slow when parsing text. The grammar is large and complex, and GF’s parsing algorithm may explore many possibilities, which leads to high computational cost for each token. This slowness becomes apparent when scaling up to large corpora: parsing thousands of tokens with ZRG can be time-consuming, making it less ideal for real-time or high-throughput applications.

The underlying cause of this is that GF performs exhaustive parsing according to the grammar rules, and isiZulu’s agglutinative structure means a single word can contain many morphemes, each with multiple possible interpretations. As a result, parsing a single complex word may involve constructing and comparing many candidate parse trees, slowing down the overall process.

4.3.1.2 Brittleness to Input Variations

Another limitation of ZRG is its brittleness – a sensitivity to even minor deviations (e.g., spelling errors) or unexpected forms in the input. ZRG, which is a form of rule-based system, tends to be strict: an input word must exactly match one of the patterns encoded in the rules (and be in the lexicon) to be successfully parsed. If the input has a spelling error, an unrecognised loanword or name, or a slightly different morphological construction not covered by the rules, ZRG may fail to parse it at all.

The ZRG’s known lack of robustness to unexpected input means that it does not gracefully handle tokens outside its predefined lexicon or constructions. For example, if ZRG encounters a perfectly valid isiZulu word that was omitted from its lexicon, it will likely reject it (fail to parse), whereas a more forgiving system might attempt a partial analysis. If a word or morpheme is missing from the ZRG lexicon, the parser fails to analyse the word correctly, often defaulting to a *MkSymb* (Symbolic Marker) placeholder, indicating an unknown lexical item. This is comparable to Ukwabelana’s parsing errors, where the DCG grammar sometimes failed to generate correct hypotheses for novel words. In summary, ZRG’s strictness ensures high precision for known inputs but makes it fragile in the face of unexpected or out-of-vocabulary items, much like other rule-based approaches.

4.3.1.3 Ambiguity and Multiple Segmentations

A further challenge with using ZRG (and indeed any rich morphological analyser) is ambiguity – the possibility of multiple valid segmentations for a single word. IsiZulu words often allow multiple morphological parses depending on context. The ZRG, when parsing in isolation, will typically return all analyses that fit the grammar. This leads to multiple possible segmentations for some tokens. In this study’s use of ZRG for segmentation, it was observed that certain words had more than one plausible breakdown according to the grammar rules. This is not a *flaw* per se – it reflects genuine linguistic ambiguity – but it requires additional steps to handle. Without broader context, the grammar itself cannot know which interpretation is correct, similar to how the ZulMorph analyser “produces all possible morphological analyses for the token” and does not disambiguate among them.

The ZRG is more constrained than ZulMorph in the sense that it encodes strict grammatical agreements, but ambiguity still arises for certain prefixes or homographs. This study does not attempt to present a full disambiguation of ZRG’s output, but it simply selects the most likely parse in cases of ambiguity, based on naïve probability calculations. This strategy, while practical, highlights the point that ZRG alone is not a disambiguation tool – it generates candidate segmentations, and it may require a contextual filter or external model to choose the correct one. However, while multiple morphosyntactic analyses are often possible, the segmentations resulting from them are often the same, reducing the ambiguity with respect to the segmentation task.

The need to resolve multiple possible segmentations was also encountered in prior corpora: the NCHLT project had to define a canonical segmentation for each word (deciding among alternatives) (Eiselen & Puttkammer, 2014a), and the Ukwabelana corpus creators had to manually validate outputs of their parsing algorithm to ensure a single, correct segmentation for each entry (Spiegler, Van Der Spuy, & Flach, 2010). These experiences reflect the limitations of ZRG in handling ambiguity.

4.3.2 Strengths of the ZRG

4.3.2.1 Comprehensive Coverage

One of the primary advantages of the ZRG is its extensive linguistic coverage. As a GF resource grammar, it encodes the core morphology and syntax of isiZulu providing a wide range of morphological rules and a large lexicon for the language. In practice, the ZRG includes nearly 17 000 lexical entries, which means it can recognise and generate a vast number of isiZulu word forms (Marais & Pretorius, 2021). This broad coverage is crucial for handling isiZulu’s complex agglutinative morphology and conjunctive orthography.

Clearly, many of the challenging phenomena in isiZulu (such as morphophonological alternations and high degrees of affix fusion) can be “handled in a principled way” by the ZRG and its runtime system. In other words, ZRG was built to systematically cover the language’s grammatical constructions, giving it the ability to parse or generate structures across a wide linguistic spectrum. This comprehensive foundation makes ZRG a strong starting point for segmentation tasks, because it should, in theory, be able to analyse most well-formed isiZulu words.

4.3.2.2 Linguistic Robustness and Correctness

The design of ZRG is grounded in linguistic theory, which contributes to its robustness in terms of producing grammatically valid analyses. The grammar has been “built to be correct” with respect to isiZulu morphosyntax. This means that if ZRG produces a parse or a segmentation for a given token, that analysis is consistent with the known linguistic rules of isiZulu. Such linguistic validity is a clear strength compared to purely data-driven approaches, since ZRG is less likely to propose implausible segmentations.

Furthermore, because GF resource grammars serve as high-level linguistic libraries, developers can rely on ZRG’s well-tested morphological rules instead of crafting ad-hoc solutions. Earlier work has shown that ZRG and the GF runtime successfully handle many isiZulu-specific morphological complexities, reinforcing its role as a structured grammar-based system.

This strength is also evident in ZRG’s role as the first openly available isiZulu parser, capable of handling not only isolated word segmentation but also constructing parsed syntactic structures, showcasing its comprehensive linguistic modelling. In summary, ZRG’s foundation in explicit linguistic principles ensures that every segmentation is linguistically justified, distinguishing it from statistical segmenters that rely solely on observed patterns without inherent grammatical validation.

4.3.2.3 Custom Linearisation (Flexible Output)

Another significant advantage of the ZRG is the flexibility of its linearisation, which can be leveraged for custom output formatting. GF grammars are bidirectional: they can parse input strings into abstract syntax trees and linearise (generate) strings from those trees. By customising the linearisation rules of ZRG, it becomes possible to generate morphologically segmented output rather than normal orthographic text. In practical terms, this means that one can adjust the grammar’s output to insert segment boundaries (such as hyphens or spaces) between morphemes in a flexible manner, thereby having the ability to customise the granularity of surface segmentation.

This capability is extremely useful; it allows one to create a segmented dataset automatically from raw text by linearising the trees into a segmented format. The ability to define how the grammar realises words is a type of controllable output generation unique to a rule-based framework like GF. This custom linearisation gives researchers the freedom to experiment with different segmentation granularity using the same underlying grammar.

This flexibility is a clear strength of the ZRG: it is not only a parser, but also a generator that can be tailored to produce the kind of segmented output required for various experiments. In contrast, datasets produced by other approaches often have a static nature; once a segmentation decision is made, the output is fixed and cannot be easily altered to reflect alternative segmentation strategies. Such inflexibility limits the ability to explore and evaluate different segmentation models or strategies. However, with ZRG, the segmentation style can be dynamically adjusted, providing a valuable advantage for investigating the optimal approach to morphological segmentation in isiZulu.

With this pipeline, this work can generate a large-scale surface-segmented dataset, allowing for an investigation into the feasibility of using such data for a supervised machine learning system. In addition, it allows an assessment of how well this approach can complement and address the previously discussed challenges associated with ZRG.

4.3.2.4 Massive Surface-Segmented Data Pipeline

Building on the flexible output provided by the ZRG, this work implements a pipeline capable of generating a massive surface-segmented dataset. By systematically processing raw isiZulu text into a segmented format at varying levels of granularity, this pipeline enables the creation of a massive dataset that is both linguistically rich and computationally adaptable.

This extensive segmented dataset not only serves as a comprehensive linguistic resource, but also allows one to investigate the feasibility of using such synthetic data for training supervised machine learning systems. Given the limitations of the ZRG, such as its slow

parsing performance, brittleness to input variations, and ambiguity in segmentation, the generated data offers a practical means to complement these challenges. In effect, the surface-segmented corpus can be employed to train models that may overcome some of the ZRG’s inherent limitations by learning segmentation patterns directly from data.

Furthermore, the flexibility in segmentation granularity provided by the ZRG means that alternative segmentation strategies can be explored. This adaptability is especially valuable when assessing how well supervised machine learning systems can replicate or even improve the segmentation quality produced by the ZRG. The present study aims to utilise the generated synthetic data for investigating supervised morphological surface segmentation of isiZulu at different granularity levels. The goal is to leverage the strengths of both rule-based systems and machine learning approaches to overcome the weaknesses of each approach.

These various limitations and challenges highlight the need for continued research and development in the morphological decomposition for isiZulu, particularly in developing more robust approaches to handle its complex agglutinative structure and improving the accuracy of rule-based approaches.

4.4 Data Acquisition

This section describes the data collection and pre-processing steps followed in this study. For the task of morphological segmentation, it is crucial that the dataset is of high quality and contains as many grammatically correct isiZulu texts as possible. To achieve this, the datasets were sourced from open-source repositories that provide good-quality data.

Grammatically correct words are essential for morphological segmentation, especially since the parsing depends on a predefined grammar or lexicon. In grammar-based approaches, the parsing system uses specific rules or lexicon entries to identify and parse a word. Thus, if the dataset contains grammatically incorrect words or other deviations not covered by the lexicon, the parser may fail to recognise the word structure or meaning, resulting in incorrect segmentations or an inability to parse the text altogether. Therefore, ensuring grammatical correctness in the dataset helps the parser to apply its rules effectively, leading to accurate and reliable segmentation outcomes.

Christodouloupoulos and Steedman (2015) presented the *Bible Corpus*², which includes texts in 100 languages and is composed of the Old and New Testaments. The dataset, structured in XML format, contains approximately 30 000 lines of isiZulu Bible verses.

²<https://christos-c.com/bible/>

The *Autshumato monolingual isiZulu corpus*³ is taken from Autshumato parallel corpora, which was originally created by Groenewald and Fourie (2009). This data combines various sources such as magazines, newsletters, and documents crawled from the government domain (*.gov.za). The total number of isiZulu sentences in this dataset was 230 000.

*NCHLT isiZulu Text Corpora*⁴ is a list of sentences that were developed by Eiselen and Puttkammer (2014a) during the NHCLT project discussed above. The total number of isiZulu sentences in this dataset was about 75 000 after the cleaning process.

The *Jehovah's Witness Dataset* (JW300) is an extensive collection of parallel texts in more than 300 languages, created by Agic and Vulic (2019) to facilitate multilingual NLP. The isiZulu portion comprises more than 1 million sentences, crawled from the `jw.org` website and stored as XML files. Although this data set is substantial, it is no longer available as an open-source resource.

Finally, *Ukwabelana isiZulu Sentences* was used, which was developed by Spiegler, Van Der Spuy, and Flach (2010), which includes approximately 30 000 raw isiZulu sentences as described above.

4.4.1 Data Pre-Processing

To maintain consistency and high quality across these datasets, several pre-processing steps were employed to clean the data, and several were web-scraped with some unwanted characteristics that were not useful for the task of creating a morphological segmented dataset. For example, the JW300 and Bible corpus were initially stored in XML format, which contains additional metadata, tags, and structures that are not directly useful for the morphological segmentation task. Therefore, the first step was to extract only the relevant textual content from these XML files and convert them into a simpler plain text format, making further processing and cleaning more straightforward.

The main aim of this chapter is to explain the methodology undertaken to create a morphologically segmented dataset using a rule-based system that can parse clean isiZulu sentences or words. The first preprocessing step was to lowercase every sentence in the different datasets; this was necessary since the segmentation task is not dependent on uppercase or lowercase. To further clean the dataset and to ensure its quality for segmentation, punctuation, and nonalphabetic characters such as numbers were removed. In cases where normalisation techniques were necessary (such as handling diacritics or standardising word forms), these were applied to maintain a consistent dataset across diverse sources.

³<https://repo.sadilar.org/handle/20.500.12185/581>

⁴<https://repo.sadilar.org/handle/20.500.12185/321>

4.4.2 Data Organisation

This section explains the organisation of the collected datasets, including the structure and metadata used to ensure efficient data handling and traceability.

1. *JSON Structure Implementation:* To facilitate consistent organisation, a JSON format was used to structure the data. Each entry in the JSON file represents a single sentence, ensuring that each sentence is clearly separated and easily accessible. This structure allows for the straightforward selection and manipulation of individual sentences or tokens for various NLP tasks. The JSON entries follow a clear format:

```
[
  {
    'sentence': '<isiZulu text>',
    'origin': '<source dataset>',
    'date_of_access': '<date>'
  }
]
```

Each entry includes the sentence itself, its origin (e.g., Bible Corpus, Autshumato, or Ukwabelana), and the date when the data was accessed. This standardised format simplifies the loading, filtering, and transformation of data across different processing stages.

2. *Source Tracking and Metadata Inclusion:* Metadata inclusion ensures that each data point retains information about its origin and acquisition details, enhancing traceability and data management. The field “*origin*” documents the source of the data set (e.g., Bible Corpus, NCHLT or JW300), while the “*date_of_access*” field records the date when the data were accessed or downloaded. This approach allows for tracking the provenance of each sentence, ensuring that researchers can selectively analyse sentences based on their source, collection date, or any future metadata additions.

4.5 Data Parsing

In the previous chapters, in line with *Sub-objective 1*, this study proposed a method using ZRG to parse a large corpus of isiZulu text and generate a treebank.

The first process of achieving this is through *parsing* isiZulu text, using the ZRG that is based on the GF runtime. The GF runtime can parse natural language strings into abstract syntax trees. This section outlines the parsing methodology, highlighting data

diversity, parallel processing strategies, and error handling. Below are different steps and processes undertaken to parse isiZulu data, in line with the *Sub-objective 1*, which proposed using a *Zulu Resource Grammar* (ZRG) “to generate to parse a large corpus of isiZulu text to obtain a treebank”.

4.5.1 Dataset Preparation and Batching

Using the dataset described above, a random number of different data sources (“origin”) was selected to be part of the sentences parsed to ensure variety. This selection process aimed at avoiding biases associated with one domain and to represent a broad range of morphosyntactic patterns typical of isiZulu, which has the potential to make the model generalise better across different words. All chosen sentences were then divided into batches of five, a configuration optimised to manage processing load and facilitate parallel execution. By batching in this way, each batch can be processed independently, allowing the parsing system to handle multiple batches concurrently, since parsing is a time-consuming task.

4.5.2 Parallel Processing with Docker

A parallel processing strategy was crucial, given the computational intensity of parsing with ZRG, which is observed across almost all grammar formalisms (Angelov, 2009). A Docker-based environment was used to support this, enabling resource-controlled execution across multiple server cores, which had 80 cores and 100 Gigabytes of memory. The Docker container was configured to allocate a fixed number of CPU cores (20) and memory resources (40GB), ensuring efficient utilisation without overloading the server. This setup also streamlined process management, reducing latency by running multiple parsing tasks simultaneously and allowing the server to handle large datasets more effectively.

The parsing script was adapted to initiate parsing on multiple cores by setting the maximum parallel processes, which are defined based on available server resources. The system minimises memory usage spikes and maintains stable performance across varying data loads by controlling the number of concurrent parsing threads.

4.5.3 Runtime Parsing and Output Management

Each batch of five sentences was parsed using the *Portable Grammar Format* (PGF) (Angelov, 2009), a GF formalism optimised for relatively fast runtime parsing and linearisation. Although PGF provides a compact and efficient format for grammar-based parsing, the complexity of ZRG still results in slow parsing speeds, particularly for large corpora. Despite this, PGF enables structured iterative parsing, allowing for better error tracking

and recovery during batch processing. For each batch parsed, the output includes both a JSON file and a log file. The JSON file contains successfully parsed sentences with their respective syntax trees, while the log file documents any failures along with specific error messages. This dual-output approach facilitates debugging and ensures transparency in tracking parsing errors.

Each parsed sentence can produce multiple abstract syntax trees due to isiZulu’s complex morphology, which allows for multiple valid interpretations. A greedy selection approach was employed to refine the output: from the finite set of possible parse trees, only the two trees with the highest probability were selected. This selection process, informed by a probabilistic ranking system, prioritises the most plausible syntactic structures, enhancing the quality of the segmented dataset.

4.5.4 Error Handling and Output Consolidation

Throughout the parsing process, sentences that did not parse were logged in detail, including the reason for the failure. These logs are essential for error analysis, providing insights into the limitations of the ZRG in handling certain morphological or syntactic constructions. Upon completion of each batch, the *JSON* outputs from successful parses were consolidated into a comprehensive dataset, while the *.log* files were used to compile a record of failures. This consolidated approach allows for the straightforward retrieval of both successful and unsuccessful parses, streamlining dataset evaluation, and further refinement of the parsing pipeline.

4.5.5 Parsing Output Example

The Figure 4.1 represents a parsed sentence from a batch file from the dataset, which aims to demonstrate the handling of isiZulu morphology and syntax using the ZRG. Each sentence in the batch is parsed into multiple abstract syntax trees, which are structured representations of the sentence’s syntactic and morphological components.

4.5.5.1 Batch Structure and Parsing Process

Each batch follows a structured format to ensure efficient organization and retrieval of parsed sentences. The key elements include:

1. *Batch ID*: Each batch is assigned a unique identifier, e.g., `batch_10004`. This identifier helps organise and retrieve parsed sentences efficiently.

2. *Input Sentence*: The original isiZulu sentence, e.g., “*esikwaziyo ukuthi bekulwa abafana ababili kodwa asazi impi isuswe yini*”, (meaning “what we know is that two boys were fighting, but we do not know what caused the battle”) is used as an example here, which in a batch text-file there would be four more other sentences.
3. *Parsed Trees*: Each sentence is parsed to generate the most likely trees using the statistical parsing algorithm of the GF runtime, capturing possible syntactic structures. In this example, two parse trees are presented for each sentence, reflecting the most probable syntactic interpretations based on a greedy selection approach. The output format is as follows:
 - *Structure of Parse Trees*: Each parse tree is presented as a structured combination of syntactic units. The primary units in the parsed trees include:
 - (a) *Chunks* (e.g., **ChunkPhr**): Represent key phrases within the sentence, such as verb phrases or noun phrases.
 - (b) *Syntactic Labels* (e.g., **Symb_Chunker**, **VP_Chunker**): These labels represent specific grammatical roles or structures within isiZulu, such as symbolic markers (**Symb_Chunker**) or verb phrases (**VP_Chunker**).
 - (c) *Morphosyntactic Tags* (e.g., **PPos**, **TPresTemp**): Capture tense, polarity, and morphological properties, reflecting isiZulu’s agglutinative nature. For instance, **TPresTemp** denotes present tense, and **PPos** indicates positive polarity.

Each parsed tree provides a unique syntactic analysis of the sentence, illustrating how the sentence structure is interpreted by the parser. For example, in the first parse tree of `batch_10004`, the phrase “*esikwaziyo*” (meaning “what we know”) is categorised as **Symb_Chunk**, representing a symbolic or nominal chunk. In contrast, the phrase “*impi isuswe*” (meaning “what caused the battle”) is analysed as a predicate, with polarity and tense information attached. This detailed categorisation provides insights into how different sentence components are syntactically organised, which is crucial for understanding the linguistic structure and accurately capturing the morphosyntactic patterns of isiZulu.

- *Handling of Morphological Constructions*: The example shown below in Figure 4.1 illustrates how ZRG examines isiZulu’s conjunctive constructions to create an abstract syntax tree when a sentence such as “*esikwaziyo ukuthi bekulwa abafana ababili kodwa asazi impi isuswe yini*” is parsed.

During the parsing, two trees were generated; however, upon inspection, the trees were almost similar in linearisation. This is mostly the result of parsing at the sentence level rather than at the word level, which resulted in a word being considered in context, where a word’s morphological analysis was influenced by the words around it, which reduces ambiguity. Henceforth, the first tree was used for further processing.

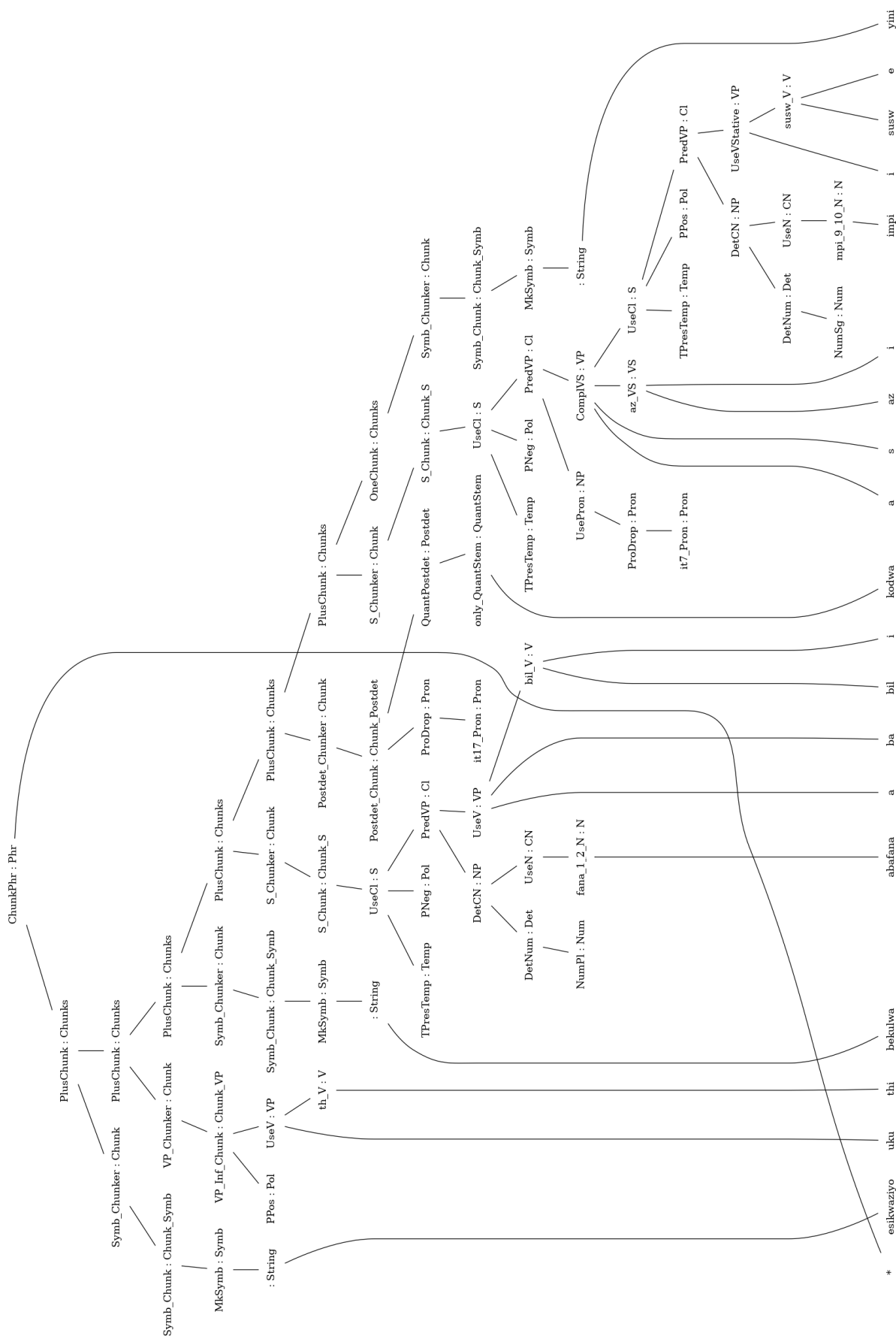


Figure 4.1: Parse abstract syntax tree generated by ZRG with default segmentation.

4.5.6 Processing Trees

A parsing system like the GF runtime, when used with the ZRG, relies on a lexicon that serves as a base of lexical entries, allowing the parser to recognise and segment words into meaningful parts. In a language like isiZulu, with its rich and complex morphological structure, creating a comprehensive lexicon that covers every possible form and morpheme is challenging. Figure 4.1 illustrates the syntax tree resulting from the parsing of the sentence: “esikwaziyo ukuthi bekulwa abafana ababili kodwa asazi impi isuswe yini”. Parsing with the ZRG often results in the lexical item *MkSymb* (or *Symb_Chunk*), indicating a limitation where ZRG fails to parse a word entirely. *MkSymb* is essentially a fallback mechanism that ZRG uses when it encounters a word or morphemes outside its lexicon. Instead of parsing the term, it leaves it as a placeholder (*MkSymb*) within the tree, indicating an unsegmented component.

Several of the trees that were generated had this symbol showing the limitation of the ZRG. At first glance, the simplest idea was to reduce and discard all the sentences that had such symbols and retain only those that were correctly parsed. This approach was not feasible, as most trees had this symbol. Thus, discarding trees containing *MkSymb* would result in the loss of potentially useful information from the successfully parsed portions of those trees, since, in many cases, only a small segment of a tree is unrecognised, while the remaining parts are correctly parsed.

A workaround to this was to use *pgfaux*⁵, which allowed retaining and working with the correctly segmented sub-components using a specific function implemented called *sub-trees_of_cat*. This function can identify and extract only the subtrees (or portions of the parsed tree) that match a specified category, such as “Chunk”. This function allows the extraction of successfully segmented tree parts while skipping the *MkSymb* sections. This function traverses the tree recursively; it filters out incomplete segments without losing valid parsed information. The extracted subtrees, being valid segments, are suitable for training and analysis, even if parts of the original sentence remain unsegmented. This approach maximises the utility of the data, preserving valuable syntactic and morphological information that contributes to building a more robust dataset.

4.5.7 Example of Application

To explain this more visually, consider the example of a tree in Figure 4.1, which contains both segmented and unsegmented components. In this tree, certain components, such as *Symb_Chunk* (*MkSymb* “esikwaziyo”), *Symb_Chunk* (*MkSymb* “bekulwa”), and *Symb_Chunk* (*MkSymb* “yini”), indicate segments that were not successfully parsed due to lexicon constraints.

⁵<https://pypi.org/project/pgfaux/>

However, even within these unparsed sections, there are several parsed subtrees that provide valuable syntactic and morphological information. The function *subtrees_of_cat* from the *pgfaux* module was used to identify and extract these successfully segmented components. Two examples of successfully parsed subtrees within the structure are shown in Figure 4.2 and 4.3. However, these subtrees are just extracted from the original trees and can also be wrong in terms of lexical item mapping. This is caused by the ZRG when it has to rely on a chunking strategy, which then results in a less likely parse for trees and does not influence the grammatical correctness of a generated text when linearised.

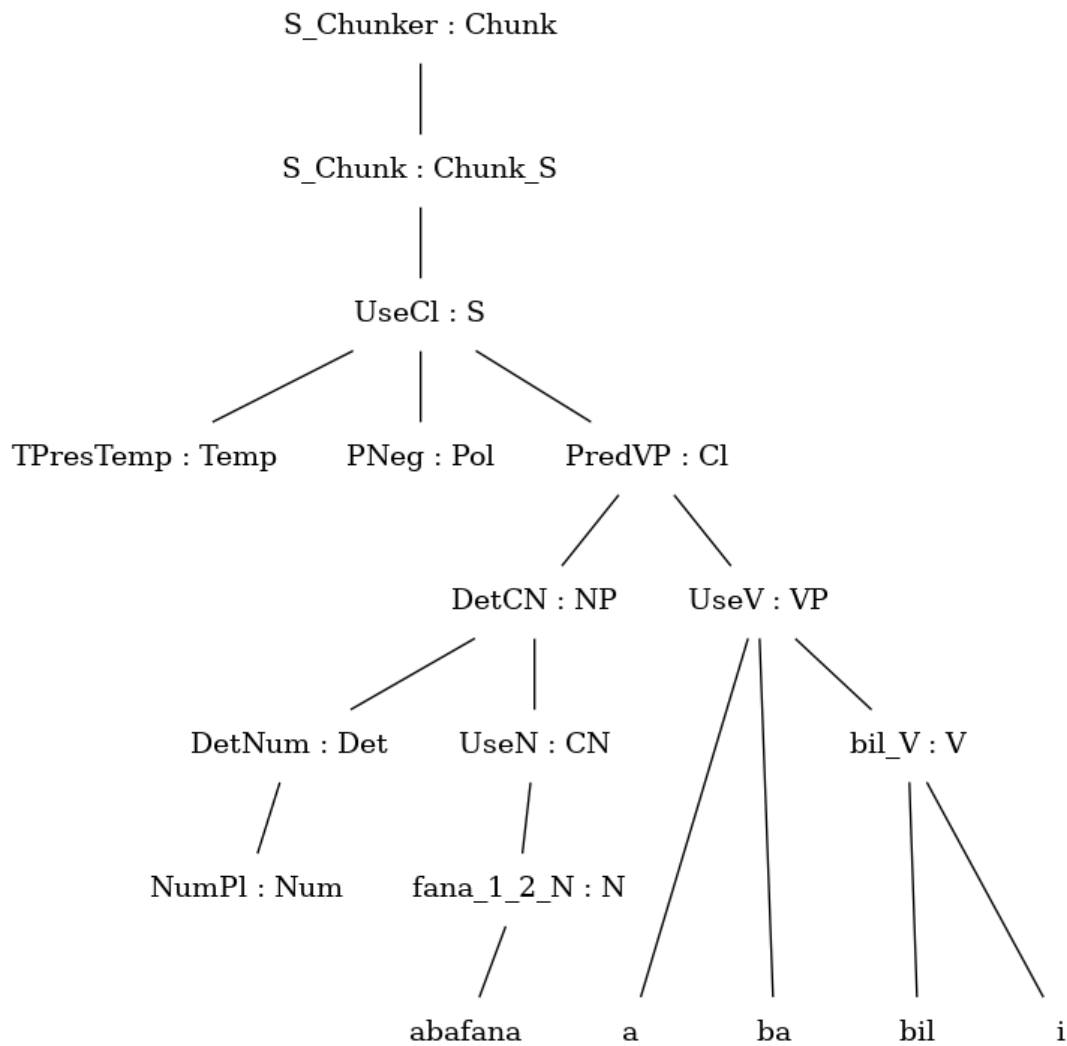


Figure 4.2: The first found subtrees from the bigger tree.

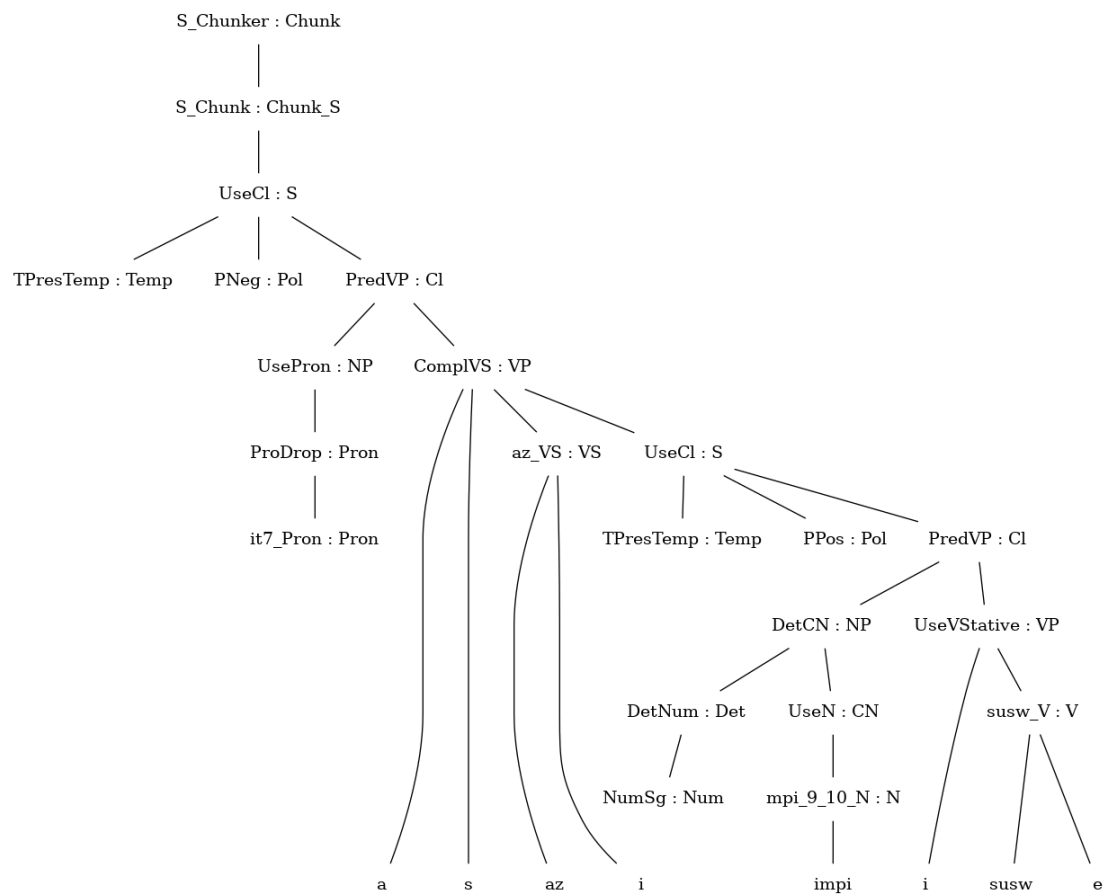


Figure 4.3: The second found subtrees from the bigger tree.

4.6 Data Augmentation

Data augmentation (DA) plays a critical role in enriching the dataset, specifically in low-resourced languages where it is limited. Data augmentation is the process of creating synthetic data by adding slightly modified copies of existing data, which aims to improve the diversity of training data, thus helping the model to better generalise to unseen testing data (Li *et al.*, 2022). Traditionally, this technique was used in computer vision to increase the number of images, through transformations such as flipping, rotating, and scaling, while preserving semantic content (Shorten *et al.*, 2021).

In the NLP context, it is considered challenging to preserve the semantics and meaning of texts during the DA, influenced by the inherent intricacies of a language or the discrete nature of language (Feng *et al.*, 2021). Despite such shortcomings, DA is continuously gaining attention as new techniques emerge, primarily in low-resource settings where available digital data is limited. More specifically, in the advent of Large Language Models (LLMs), which have shown remarkable performance across various NLP tasks, presents an opportunity to address these challenges by offering a scalable, high-level solution for data augmentation (Pranjić *et al.*, 2024; Zha *et al.*, 2023).

However, augmentation strategies that involve manipulating language remain challenging, particularly for low-resource languages, where models lack sufficient training on language-specific morphological data. Consequently, when these models are tasked with handling morphology-dependent processes, they often underperform due to their limited exposure to the morphological complexities unique to these languages. The ZRG allows for an augmentation that is guaranteed to be morphosyntactically correct. Furthermore, the segmentation use case does not require semantics to be preserved, since the task is focused at morphosyntactic analysis.

4.6.1 Workflow for Data Augmentation

The data augmentation workflow developed is a comprehensive, rule-based system that generates new data from existing datasets by altering syntactic features of trees such as polarity and plurality, and lexical items such as nouns and verbs to create diverse syntactic variations. In the above section, a detailed strategy regarding how parsing was done to create a dataset of abstract syntax trees was presented, which can be used for data augmentation to increase the datasets and make them as diverse as possible. This process enables the generation of synthetic yet linguistically accurate data, adding variety and depth to the training set.

This process was deemed necessary since the isiZulu language has an agglutinative structure and complex morphology, which can result in data sparsity and, consequently, can

result in a model that does generalise well in terms of new words with different morphological structures. Therefore, creating diverse training data can potentially improve the performance of segmentation models.

4.6.2 Augmentation Techniques

The data augmentation method techniques work by iteratively replacing various linguistic components within the syntactic tree representations of isiZulu sentences. These components include nouns, verbs, tense markers, polarity, pronouns, and other grammatical markers found on the abstract trees generated by the ZRG. The process performs multiple augmentation “runs,” each time generating new tree structures by shuffling elements across trees, which helps create unique variations of the original data. However, since this is a naive approach, the resulting sentences may not always be semantically correct; this limitation is discussed below. The key augmentation techniques used are as follows:

1. Noun Shuffling

The script identifies and extracts nouns within each syntax tree (using patterns like `(UseN noun_N)`). In each run, nouns are shuffled across trees, replacing each noun with another randomly selected noun from the pool of all available nouns. This technique introduces new agreement contexts for verbs without altering the overall sentence structure, enabling the model to encounter different noun contexts for similar sentence structures.

2. Verb Shuffling

Similarly, verbs are identified (pattern `(UseV verb_V)`) and shuffled across sentences. By substituting verbs with other verbs from the dataset, the script generates syntactic variations that do not always retain the semantic correctness of a sentence, but provide exposure to different actions and contexts. This method enhances the diversity of verb usage across different syntactic structures, providing a richer set of examples within the dataset.

3. Stative Verb Shuffling

For stative verbs, which describe states rather than actions, a separate function shuffles elements tagged as `(UseVStative verb_V)`. These verbs are more context-specific, and shuffling them enriches the dataset by demonstrating different states or conditions within similar syntactic patterns.

4. Pronoun Replacement (ProDrop)

Pronouns, especially those in dropped-pronoun (ProDrop) languages like isiZulu, are another key component. The script shuffles pronouns marked with (ProDrop pronoun) across sentences. This technique enhances the diversity of the dataset by exposing different pronoun contexts.

5. Complement Verb (CompIV2) Shuffling

The script also manipulates complement verbs, tagged as **CompIV2**, which represents verbs requiring an object noun phrase. This process involves taking a transitive verb and complementing it with an object noun phrase. If an object morpheme is present together with the verb, the resulting segmentation produces different morpheme sequences that incorporate this object morpheme. Consequently, this transformation generates a new variation of the verb, as it now appears in a different agreement context.

6. Polarity Shuffling (PPos and PNeg)

Polarity markers, denoted by **PPos** (positive) and **PNeg** (negative), are essential for expressing affirmative or negative statements. The script shuffles these markers to create new sentences that may have the opposite polarity from the original. This augmentation technique enables the model to handle both affirmative and negative constructions more effectively.

7. Number Marking (NumSg and NumPl)

The script identifies singular and plural markers (**NumSg** and **NumPl**) within the trees, allowing these markers to be shuffled across sentences. This approach enables the model to learn number agreement and plurality in different syntactic contexts.

8. Tense Alteration

Finally, the script handles tense markers, including **TRemPastTemp** (remote past), **TPresTemp** (present), **TPastTemp** (past), and **TFutTemp** (future). By substituting these tense markers within sentences, the model encounters sentences in various temporal contexts, enhancing its ability to accurately interpret and generate time-sensitive statements.

4.6.2.1 Combining Augmented Data

After each run, the script combines the newly augmented data with the original dataset, resulting in a comprehensive set of syntactically diverse sentences. By preserving unique tree structures and only replacing certain elements, this augmentation strategy maintains grammatical integrity while significantly expanding the dataset. The augmented data is then saved to a file, providing a richer dataset for training segmentation models.

4.6.2.2 Example of Data Augmentation

To illustrate this process further, one can consider two sentences that have different parsing trees but have some relatedness in the tree structure or share similar lexical items.

Consider the isiZulu sentence: “*amantombazane amahle amathathu ahlala ekhaya nomama wabo*” (The three beautiful girls live at home with their mother). The parsing tree for this first sentence is illustrated in Figure 4.4. Similarly, consider a second isiZulu sentence: “*ubaba kade enezimoto ezimbili ezimbi ezishayelwa abafana bakhe abadlala ibhola*” (The father had two bad cars driven by his sons who play soccer). The graphical representation of this second sentence is represented as an abstract syntax tree structure in Figure 4.5. All these abstract syntax trees are produced using the default ZRG system.

Table 4.1 provides a detailed overview of various data augmentation manipulations applied to isiZulu sentences. These manipulations include adjustments in tense, polarity, adjective, and verb choices, enabling a syntactically and semantically diverse array of sentence variations. With this approach, a single parse tree can generate multiple sentence variations. For instance, changing the tense in Tree 1 (*T1*) can produce different sentences, depending on whether the modification applies to one instance of **TPresTemp** or multiple. Such selective changes result in various sentences with distinct temporal contexts, allowing nuanced control over data generation.

This flexibility extends to other lexical items that appear multiple times within a sentence, such as adjectives or noun classes. The augmentation process can target one or more instances of these items, or even leave them unchanged, creating diverse sentence structures. Each manipulation affects not only the modified word, but also related words, due to *isiZulu’s agreement system*. For example, changing from plural to singular influences noun prefixes, verb agreements, and possessive concords across the sentence, demonstrating the language’s rich morphosyntactic dependencies.

Although this strategy is advantageous for its flexibility, allowing for the seamless manipulation of trees and substitution of lexical units across trees, it does not always yield semantically coherent phrases, although grammatical correctness is maintained. Since ZRG enforces correct morphosyntax, every generated sentence adheres to isiZulu’s gram-

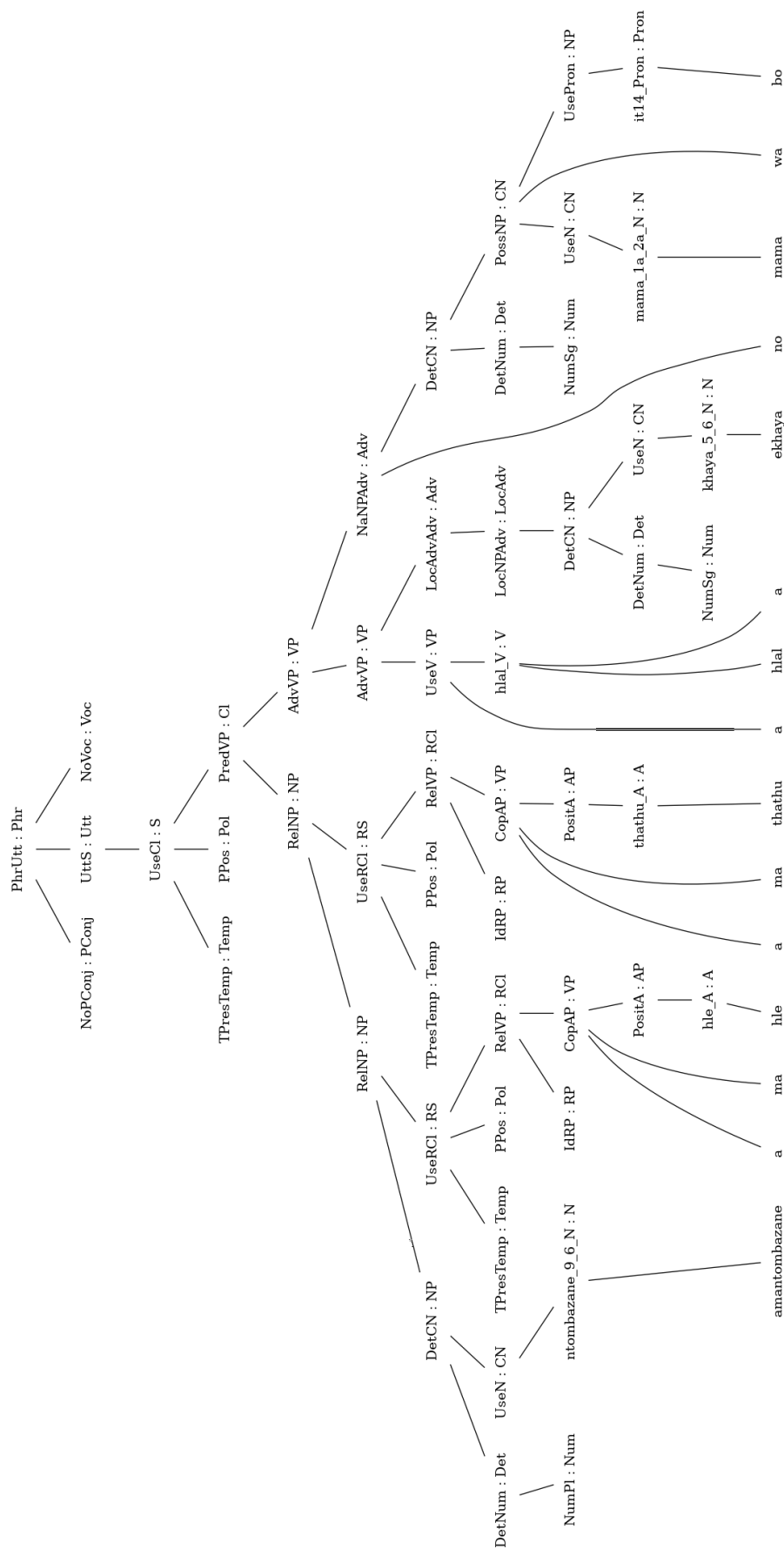


Figure 4.4: The abstract syntax tree for “*amantombazane amahle amathathu ahlala ekhaya nomama wabo*”.

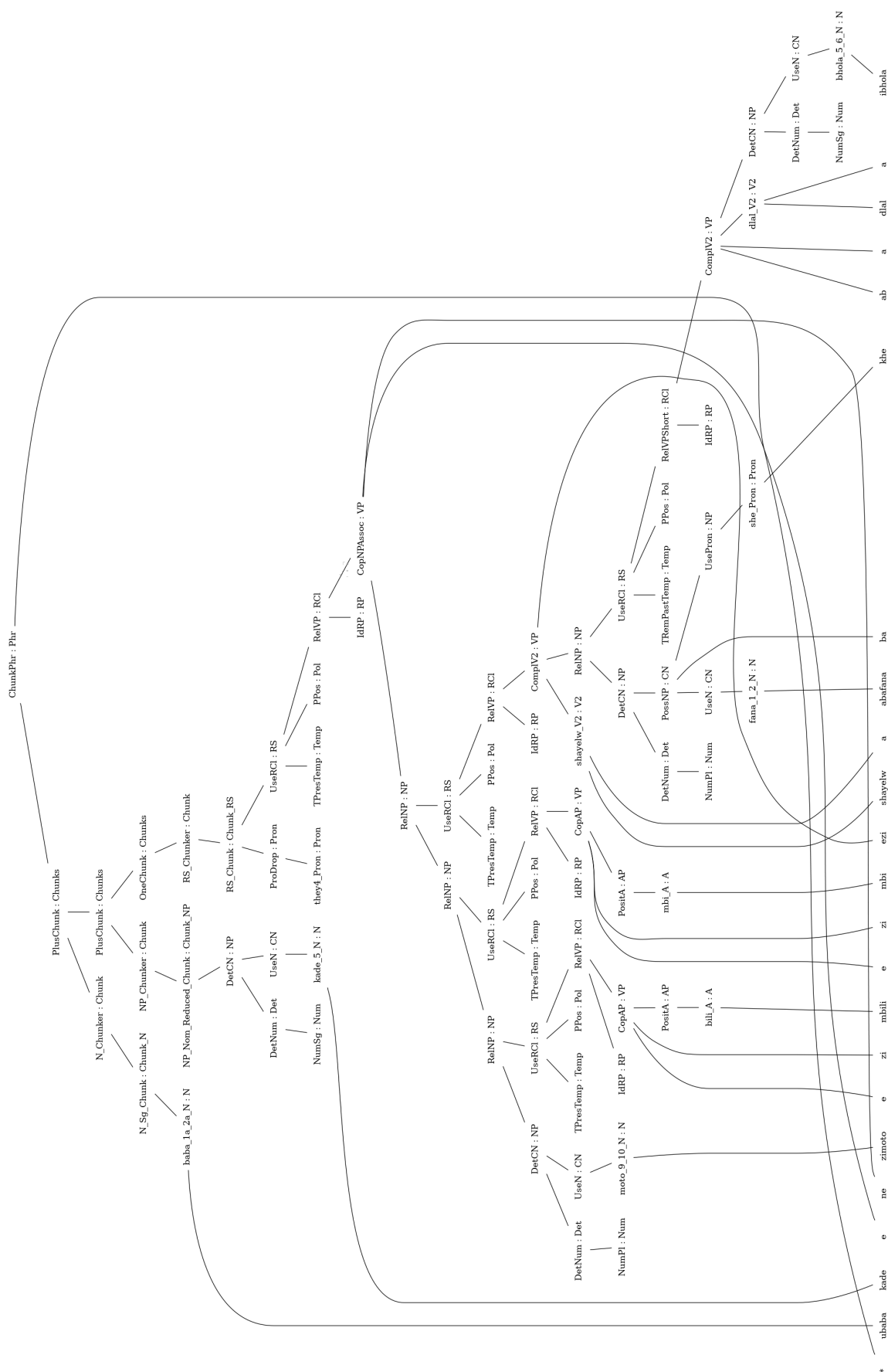


Figure 4.5: The abstract syntax tree for “ubaba kade enezimoto ezimbili ezimbi ezishayelwa abafana abadlala ibhola”.

Table 4.1: Manipulation of terms and resulting sentences.

Manipulation	Term Replaced	New Term	Resulting Sentence
Tense Change T1	TPresTemp (Present)	TPastTemp (Past)	<i>amantombazane abemahle abemathathu ebahlala ekhaya nomama wabo</i> (The three girls who were beautiful who lived at home with their mother)
Tense Change T2	TPresTemp (Present)	TPastTemp (Past)	<i>ubaba kade enezimoto ezimbili ezimbi ezazishayelwa abafana bakhe abadlala ibhola</i> (The father had two bad cars that were driven by his sons who play soccer)
Polarity Change	PPos (Positive)	PNeg (Negative)	<i>amantombazane amahle amathathu awahlali ekhaya nomama wabo</i> (The three beautiful girls do not live at home with their mother)
Polarity Change	PPos (Positive)	PNeg (Negative)	<i>ubaba kade enezimoto ezinhle ezingashayelwa abafana bakhe abadlala ibhola</i> (The father had two beautiful cars that are not driven by his sons who play soccer)
Adjective Change T2&T1	thathu_A (three)	mbili_A (two)	<i>amantombazane amahle amabili ahlala ekhaya nomama wabo</i> (The two beautiful girls live at home with their mother)
Verb Change	hlal_V (live/stay)	dlal_V (play)	<i>amantombazane amahle amathathu adlala ekhaya nomama wabo</i> (The three beautiful girls play at home with their mother)
Combined Change	TPresTemp (Present) hle_A (beautiful) thathu_A (three) hlal_V (live/stay)	TPastTemp (Past) khuL_A (big) mbili_A (two) dlal_V (play)	<i>amantombazane amahle amabili adlala ekhaya nomama wabo</i> (The two big girls played at home with their mother)

mathematical rules. However, ZRG does not evaluate semantic coherence, meaning that certain lexical substitutions may result in grammatically correct but contextually odd sentences.

However, despite this limitation, the augmentation technique remains valuable for training models that require a wide range of syntactically varied input, especially in low-resource settings where data diversity is crucial. This approach is especially relevant for training morphological segmenters at the token level rather than at the sentence level, because token-level segmentation does not depend on the semantic coherence of the sentence, which can influence segmentation.

Semantic correctness is less critical in token-level segmentation, as explored in this work; instead, accurate grammar and precise morpheme boundaries are more important. This data augmentation technique effectively supports these requirements, as it ensures grammatical accuracy and consistency in morphological structure without maintaining sentence-level semantic coherence.

4.7 ZRG Linearisation

The final step in creating surface segmented data is known as *linearisation*; in this step, the ZRG uses the GF runtime to linearise abstract syntax trees into natural language strings. ZRG systematically uses a specific token for binding subwords together at runtime to form grammatically correct surface tokens.

4.7.1 Different Linearisation Strategies

The linearisation strategy in ZRG is designed to reliably generate surface-segmented texts by using a specific binding token. In the data augmentation process, abstract syntax trees are created and then manipulated at the lexical level. These abstract trees are then converted into human-readable sentences through linearisation, where the binding token is used to fuse morphemes seamlessly. This binding token is left empty to ensure that segments join together naturally without added spaces, resulting in grammatically correct isiZulu output.

When surface segmentation is in view, there is an open question as to where the morpheme boundary must be placed within a token, since a token can have different morpheme boundaries that can be justified linguistically. Thanks to the core morphology operations of the ZRG, it can be easily tuned to create different possible segmentation strategies that are more or less fine-grained. This flexibility is one of the main advantages of the ZRG-based approach, allowing for customised segmentation strategies that can be adapted to the specific requirements of a given NLP task.

This work investigates three distinct linearisation strategies: *Lin A*, *Lin B*, and *Lin C*, each of which affects how words are segmented. These strategies are designed to explore various approaches to morphological segmentation and assess their performance intrinsically (within segmentation quality itself) and extrinsically (in downstream NLP tasks).

4.7.1.1 Lin B: Middle-Grained Segmentation

This approach relies on the segmentations that are naturally built into the ZRG internal structure, with an additional segmentation between noun prefixes and stems. Segmentations are introduced only where morphological “gluing” would naturally occur, following the GF style, where fused morphemes are segmented according to the grammatical structure. This method emphasises morphological consistency based on linguistic principles without over-segmenting.

4.7.1.2 Lin A: Fine-Grained Segmentation

Lin A takes a finer approach, retaining all segmentations from Lin B but further breaking down morpheme sequences that may otherwise appear as single strings. This aggressive segmentation approach aims to maximise the visibility of individual morphemes within word forms, even if they are typically written as single segments. The goal is to expose all underlying morphemes, which can be beneficial for tasks requiring high morphological granularity.

4.7.1.3 Lin C: Coarse-Grained Strategy

Lin C reduces segmentation by keeping some natural morpheme boundaries in place while consolidating sequences of verb and copulative prefixes into single segments. This approach focuses on minimising the segmentation of highly fused morpheme sequences, which can help models to interpret certain verbal and copulative structures as unified units. Lin C provides a balance between morphological detail and cohesion, potentially benefiting tasks where over-segmentation could introduce noise.

4.7.2 Application of Linearisation Strategies

To demonstrate the impact of different linearisation strategies, consider the isiZulu phrase: “abafana bakabhuti banezinja ezinenkani” (my brother’s boys have stubborn dogs). This sentence is simple yet has a complex morphological structure that includes prefixes, noun classes, and verb morphology, making it an ideal candidate for illustrating how each lin-

earisation strategy affects segmentation. Figure 4.6 below illustrates parse tree for this sentence is as follows:

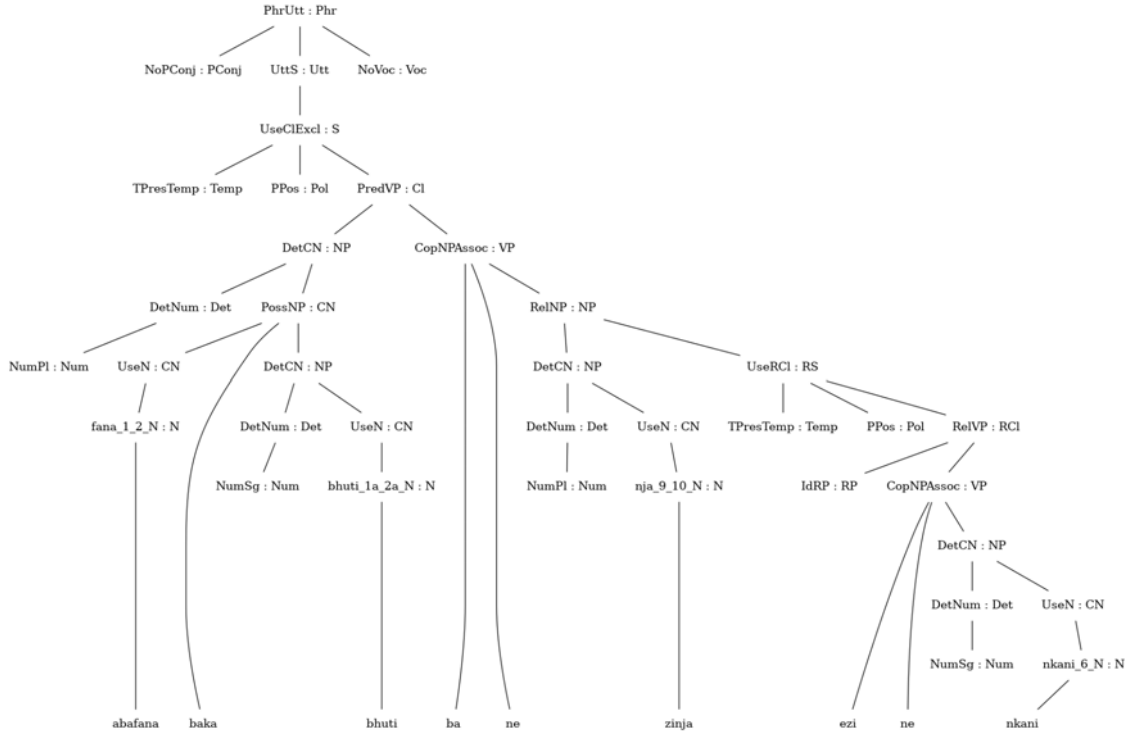


Figure 4.6: The abstract syntax tree for “*abafana bakabhuti banezinja ezinenkani*”.

Table 4.2: Comparison of linearisation strategies for the sentence “*abafana bakabhuti banezinja ezinenkani*”.

Linearisation Strategy	Segmented Sentence
<i>Lin A</i>	<i>a-ba-fana baka-bhuti ba-n-e-zinja ezi-n-e-nkani</i>
<i>Lin B</i>	<i>aba-fana baka-bhuti ba-ne-zinja ezi-ne-nkani</i>
<i>Lin C</i>	<i>abafana baka-bhuti bane-zinja ezine-nkani</i>

Table 4.2 shows how each linearisation strategy segments the sentence, highlighting the differences in the way morphemes are divided. Each linearisation strategy affects the segmentation as follows:

1. *Lin A* segments at the smallest morpheme level, explicitly marking each grammatical element within a word. In Table 4.2, *Lin A* corresponds with: “*a-ba-fana baka-bhuti ba-n-e-zinja ezi-n-e-nkani*”. This strategy aggressively segments words into individual morphemes, such as “a-ba-fana” (the boys), “ba-n-e” (they have), and “ezi-n-e” (which have). Even grammatical morphemes like “a-” in “a-ba-fana” are explicitly segmented. This highly detailed segmentation strategy is useful for linguistic

analysis. The dataset generated by this strategy is used to train a fine-grained segmentation model.

2. *Lin B* applies segmentation at meaningful grammatical boundaries, ensuring that prefixes and stems remain distinct while avoiding unnecessary segmentation of fused morphemes. In Table 4.2, Lin B corresponds with: “*aba-fana baka-bhuti ba-ne-zinja ezi-ne-nkani*”. This strategy maintains some morpheme groupings for better readability, such as “aba-fana” instead of “a-ba-fana” and “ezi-ne” instead of “ezi-n-e”. It provides a structured segmentation that balances granularity and linguistic coherence. The dataset produced using this strategy is used to train a model with moderate segmentation granularity.
3. *Lin C* reduces segmentation to enhance interpretability by keeping morphemes grouped into more natural word structures. As shown in Table 4.2, this strategy results in: “*abafana baka-bhuti bane-zinja ezine-nkani*”. This version minimizes segmentation boundaries, treating morphemes as cohesive units. For instance, “abafana” (the boys) and “bane” (they have) remain unsegmented. The dataset generated using this strategy is employed to train a coarse-grained segmentation model.

4.8 Final Dataset Output

This section provides an overview of the final datasets produced, presenting the numerical breakdown and bringing all processes into perspective.

From the various sources, a total of 93 630 sentences were successfully parsed. However, a significant portion of these parse trees (75 662) contained *MkSymb* components, which required further processing. After extracting useful subtrees that could be successfully linearised, the number of abstract trees was 130 548.

These abstract trees then underwent data augmentation, significantly expanding the dataset to 683 681 trees. The augmented trees were subsequently linearised and processed further. During this stage, instances that took excessively long to linearise were filtered out through post-processing. In addition, in some cases, the augmentation process resulted in combinations of syntactic features that are invalid in isiZulu. In such cases, linearisation failed, and hence, these trees were filtered out as well. As a result, the final number of successfully parsed and linearised phrases or sentences was 511 435.

Since the morphological segmenter is trained at the token level, an additional post-processing step was conducted to convert these phrases or sentences into word tokens. This step resulted in a final dataset of 367 178 unique segmented word tokens, each segmented into three different granularity levels. Similarly, the 130 548 trees that did not

undergo data augmentation were also linearised into phrases or sentences. Further extraction of unique tokens from this subset yielded 76 271 unique tokens.

These segmented tokens provide a suitable dataset for training supervised machine learning models, allowing for the systematic experimentation with different segmentation strategies, and for investigating the impact of data augmentation on segmentation.

4.9 Chapter Summary

This chapter discussed the relevant processes for preparing an annotated dataset for morphological segmentation in isiZulu to be used for this study, addressing challenges related to limited lexicon coverage, complex morphosyntactic structures, and data scarcity and sparsity typical of low-resource languages. Leveraging a diverse set of isiZulu datasets, including *Ukwabelana*, *IsiZulu Bible corpus*, *JW300*, *NCHLT isiZulu Text Corpora*, and *IsiZulu monolingual Autshumato*, thus, allowed for creating of a corpus that captures various registers and text types, thus enhancing both data quality and quantity.

The chapter began with a discussion on existing isiZulu datasets and a reflection on ZRG as a data generation pipeline, highlighting both its strengths and limitations. This was followed by a detailed explanation of the corpora acquisition process and the parsing of isiZulu sentences into abstract syntax trees. Some of these abstract syntax trees contained unsegmented words represented by *MkSymb* components. To address this issue of *MkSymb*, trees were recursively broken down into smaller parts, allowing valuable linguistic information to be retained even when complete segmentation was not possible. These trees then underwent data augmentation through syntactic tree manipulation. By modifying specific lexical items and grammatical structures within the parse trees, the dataset's volume and diversity were expanded without the need for extensive manual annotation.

Three linearisation strategies, the *Lin A* (Fine-grained segmentation), *Lin B* (Middle-grained segmentation), and *Lin C* (Coarse-grained segmentation), were then introduced, each targeting a different level of segmentation granularity. These strategies offer customised configurations that can be applied to various NLP tasks, enhancing segmentation accuracy and adaptability in downstream applications.

With the groundwork for morphological segmentation established, the next chapter will focus on the implementation and evaluation of segmentation models, examining the effectiveness of each segmentation strategy and assessing their impact on model performance across diverse isiZulu linguistic contexts.

Chapter 5

Design of Models

5.1 Introduction

The aim of this chapter is to present the design of the models developed for isiZulu morphological segmentation, directly addressing *Sub-objective 3* and *Sub-objective 4*, which seek to investigate the suitability of different machine learning models for supervised segmentation, and to train those models on the syntactic data of different linearisation strategies (granularity).

In Chapter 3, an overview of various morphological segmentation techniques was provided. Given the breadth of available methods, this study focusses on a subset of techniques tailored for supervised machine learning, with the goal of developing a final supervised model that will serve as the primary system for further investigating the hypotheses formulated in Chapter 1.

This chapter is structured as follows: Section 5.2 provides a discussion on the criteria used for model selection. Section 5.3 outlines the training process of the selected models, detailing dataset preprocessing, word tokenisation or splitting strategies, and an overview of evaluation techniques employed to assess model performance. Additionally, this section covers hyper-parameter optimisation as a means of improving performance. Finally, Section 5.4 summarises the key takeaways from the chapter.

5.2 Models Selection

5.2.1 Criteria for Model Selection

There are several factors that influence the type of system one can implement as a solution to a given task. The type of problem to be solved determines the tools, resources and approaches that can be effectively employed. In this study, the goal is to develop and train a supervised morphological segmentation tool for isiZulu using the annotated synthetic dataset generated with the ZRG. This task is framed as a sequence-to-sequence problem, where character sequences representing a word are transformed into character sequences that represent their segmented morphemes.

As discussed in Chapter 3, various studies have proposed various techniques and strategies to address the task of morphological segmentation. These range from *rule-based approaches* that rely on linguistic expertise and predefined grammatical rules to *statistical approaches*, which incorporate handcrafted features and probabilistic modelling, as well as *neural network-based approaches*, which automatically learn representations from data.

While rule-based and statistical approaches have significantly contributed to the field and continue to play a crucial role in morphological segmentation, they also present notable challenges. Rule-based systems are labour-intensive and struggle to segment tokens that do not conform to predefined rules, limiting their ability to generalise. Similarly, statistical approaches require extensive feature engineering, often demanding domain expertise. This makes them less adaptable to morphologically complex languages such as isiZulu, where the number of possible morphological patterns and features is vast.

On the other hand, neural network-based approaches have revolutionised sequence modelling, offering solutions that are not only highly effective; but also, the ability of these systems to handle large datasets and extend to various tasks without the need for extensive manual intervention render them extremely useful. These systems, which rely on automatic feature extraction, have significantly reduced the reliance on handcrafted features, enabling more streamlined development. However, their data-intensive nature poses challenges for low-resource languages, because these models require large datasets to achieve optimal performance as they extract features in an automatic fashion (Bansal *et al.*, 2021). Additionally, training well-performing models that generalise effectively to unseen data can be computationally expensive, often requiring several days of processing.

In Chapter 4, data preparation strategies were discussed, which involved using ZRG to create a morphological segmentation dataset and employing a naive data augmentation approach to significantly expand the dataset. The availability of a substantial morphologically surface-segmented dataset enabled the training of a supervised morphological segmenter. This aligns with *Sub-objective 3*, as outlined in Section 1.4.2 of Chapter 1.

Given the various types of models available, the following key criteria guided the model selection process.

1. *Capability for Sequence Modelling*: Morphological segmentation inherently involves processing sequences of characters and identifying morpheme boundaries within words. Therefore, the selected model must be able to handle sequential dependencies and contextual relationships in character sequences to accurately segment words.
2. *Supervised Learning Suitability*: Since annotated (labelled) data is available, the model must be compatible with supervised learning approaches. Supervised models can learn explicit mappings from input sequences to their segmented outputs, leading to higher segmentation accuracy compared to unsupervised approaches that rely on statistical heuristics.
3. *Minimal Dependence on Additional Expert Knowledge*: Traditional rule-based and statistical approaches require linguistic expertise to manually craft rules and features. Given the existing ZRG and the goal of reducing further expert intervention, the selected model should automate feature extraction rather than relying on hand-crafted linguistic rules.
4. *Robust Generalisation to Unseen Data*: A good segmentation model must not only perform well on training data, but also generalise effectively to previously unseen words. Given the limited existing isiZulu resources and the complex morphology of the language, the model must be resilient to OOV words, thus ensuring that it can segment novel linguistic constructs correctly.
5. *Computational Feasibility*: The selected model must balance computational cost with segmentation accuracy, ensuring that training and inference remain feasible given available resources.

5.2.2 Models Considered

In Chapter 3, various techniques for morphological segmentation were discussed, with supervised methods consistently found to outperform their counterparts when sufficient annotated data is available. Given the substantial dataset created, comprising approximately 360 000 annotated isiZulu texts, a supervised machine learning approach emerged as the most suitable choice for this task. However, the decision between a statistical or deep learning-based approach required further consideration. As discussed in Chapter 3, Conditional Random Fields (CRFs) emerged as a strong candidate among statistical machine learning methods for morphological surface segmentation, because they are used

more widely for morphological segmentation for isiZulu than its counterparts. This preference stems from CRFs' ability to capture sequential dependencies in morphemes, which are pure derivations of the orthographic word (Moeng *et al.*, 2022).

Despite their advantages, CRFs come with notable limitations. A significant drawback is their reliance on manual feature engineering, which is both labour-intensive and requires extensive linguistic expertise. The performance of CRF models depends heavily on the quality and relevance of these handcrafted features, making them less flexible and scalable compared to newer approaches (Yu & Fan, 2020). However, given that CRFs have been used for morphological surface segmentation, and various feature extraction techniques have already been developed for them, they are a viable choice to use in this study compared to other statistical approaches, and they meet almost all the criteria outlined in Section 5.2. Nevertheless, the availability of a large annotated dataset allows for the training of deep learning models, which fully align with the selection criteria and offer a very robust and fully automated approach.

The section covering deep learning techniques in Chapter 3 first discussed the Recurrent Neural Network (RNN), a type of neural network designed for processing sequences. RNNs maintain a hidden state to capture temporal dependencies (Rumelhart *et al.*, 1986). An updated version known as Bidirectional Recurrent Neural Networks (BiRNNs) extends the RNN's functionality by processing input sequences both forwards and backwards, enabling the model to capture context from surrounding elements on both sides of a given point in the sequence (Schuster & Paliwal, 1997). Long Short-Term Memory networks (LSTMs), another specialised form of RNNs, utilise memory cells and gating mechanisms to control information retention over long sequences, making them particularly useful for tasks with long-range dependencies (Hochreiter, 1997). Furthermore, Gated Recurrent Units (GRUs), a streamlined version of LSTMs, use update gates to control information flow, achieving similar results in capturing long-range dependencies while being computationally more efficient (Cho *et al.*, 2014).

The new variants of RNN, namely the LSTM and GRU, worked well with sequence-to-sequence problems and can even deal well with longer sequences, and also reduces the problem of vanishing gradient as discussed in Chapter 3. However, one issue that remains, as noted by Bahdanau (2014), is how the RNNs squeeze the information to make it fixed into a fixed-length context vector $\mathbf{c}$. In this regard, Bahdanau (2014) introduced the attention mechanism, which allows the model to dynamically focus on different parts of the input sequence at each time step. This mechanism ensures that the context vector $\mathbf{c}$ represents relevant information from the entire input sequence effectively without being constrained to a single fixed-length encoding. This approach was first incorporated in the BiRNN model, and several other models have found this method approach to be better than the normal variants of RNN, which have no attention mechanism.

Taking inspiration from sequence modelling and attention mechanisms, Vaswani *et al.* (2017) introduced the *Transformer* architecture, which fundamentally changed how neural models process sequential data. The Transformer utilises self-attention to dynamically weigh different parts of an input sequence, effectively capturing contextual relationships across long sequences (Vaswani *et al.*, 2017). This work considers the Transformer model for morphological segmentation for several key reasons:

1. *Dynamic Context Representation:* The self-attention mechanism enables the Transformer to assign importance to different segments of an input sequence dynamically. This is particularly beneficial for isiZulu, where words vary significantly in length and morphemes are context dependent. Unlike RNNs, which rely on fixed-length context windows, the Transformer flexibly models variations in morpheme structure without constraints.
2. *Handling Long-Range Dependencies:* The ability to capture both local and global dependencies is crucial in agglutinative languages like isiZulu, where morphemes interact in complex ways. The self-attention mechanism allows the model to consider relationships between distant morphemes, improving segmentation accuracy as compared to models that are constrained by sequential processing (Liu *et al.*, 2020).
3. *Efficient Morpheme Alignment:* The attention mechanism allows the model to focus dynamically on relevant parts of the input sequence, thus improving alignment between input words and their segmented morphemes. This feature is evident in models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin, 2018) and GPT (Generative Pre-trained Transformer) (Radford & Narasimhan, 2018), which excel at learning contextual representations of words and subwords.
4. *Parallel Processing Efficiency:* Unlike RNN-based models, which process sequences sequentially, the Transformer operates on the entire sequence at once. This allows for significantly faster training and inference times, making it well-suited for handling large datasets like the one created for this study.
5. *Empirical Evidence in Complex Morphology:* Apart from the theoretical advantages presented above with regard to the Transformer architecture, the empirical evidence presented in Chapter 3 demonstrates its effectiveness in terms of morphological segmentation when looking at languages with complex morphology such as isiZulu.

Based on these considerations, this study evaluates three models for isiZulu morphological segmentation: CRFs, selected for their established role in sequence labelling and the availability of predefined feature extraction methods for isiZulu surface morphological segmentation in statistical machine learning; LSTMs, chosen over GRUs due to their more extensive use in morphological segmentation literature within the RNN family; and Trans-

formers, selected for their superior ability to capture complex morphological structures and being the most efficient in the Deep learning family.

5.3 Models Design and Training

This section presents a systematic comparison of the different machine learning techniques with a view to a final selection of one technique for the investigation of the feasibility of using the synthetic data developed using a rule-based system for training a supervised machine learning system for morphological segmentation. The aim is also to gain relevant insights and be able to investigate the hypothesis and fulfil the primary subjective and subsequently the secondary one.

In terms of comparison, it should be noted that the CRF and LSTM models were adopted from existing implementations rather than being developed from scratch. However, they were intensely modified to include certain features that would allow systematic comparison around different models. This section provides a brief overview of their respective architectures. Furthermore, the Transformer model, which was implemented from scratch, is discussed in detail to also cover certain features and design choices that were implemented in the other models for systematic comparison. The focus will be on the key components of the Transformer architecture, including aspects that were not fully covered during the discussion of deep learning techniques presented in Chapter 3, but which are crucial to understanding its role in this study.

The key sub-objective addressed in this section is *Sub-objective 4*, which focuses on training segmentation models using different linearisations of the treebanks, each reflecting a different segmentation granularity.

5.3.1 Transformer Models

Chapter 3 provided a brief introduction to the Transformer architecture, with an indication that further discussion would follow in this section. The Transformer architecture was originally introduced by Vaswani *et al.* (2017), which also adopts an encoder-decoder structure similar to those of RNNs. However, unlike RNNs, which rely on a recurrent structure, the Transformer utilises an attention mechanism, enabling the model to weigh different parts of an input sequence and capture contextual relationships more effectively. Figure 5.1 illustrates the standard Transformer architecture as proposed by Vaswani *et al.* (2017). This structure forms the basis for the Transformer models implemented in this study, and will serve as a reference point for explaining the model’s design and experimental setup.

The Transformer architecture consists of two main components: the encoder and the decoder. The encoder, located on the left, processes the input sequence, while the decoder, on the right, generates the output sequence. Each encoder block is structured with two primary modules: a *Multi-Head Self-Attention* (MSA) module and a *Position-wise Feed-Forward Network* (PFFN). To enhance the depth and performance of the encoder, *Layer Normalisation* (Ba, 2016) and *residual connections* (He *et al.*, 2016) are integrated, which will be elaborated upon shortly.

In contrast, the decoder blocks include an additional cross-attention mechanism, situated between the MSA and PFFN, which enables the decoder to focus on specific parts of the input sequence processed by the encoder. This additional mechanism is essential for aligning the input and output sequences effectively. The following sections provide a detailed examination of these components and their implementation in the segmentation Transformer model.

The model implementation utilises the PyTorch library, a robust and high-performance deep learning framework widely used for building, training, and deploying neural networks (Paszke *et al.*, 2019). The PyTorch library was chosen over alternatives like TensorFlow (Abadi *et al.*, 2016), Caffe (Jia *et al.*, 2014), and Theano (Team *et al.*, 2016), which rely on static dataflow graphs. In these frameworks, the entire neural network is predefined as a fixed graph before any computation occurs. While this static approach can optimise computation and enhance scalability, it often sacrifices ease of use, debugging simplicity, and flexibility when representing diverse computations.

In contrast, PyTorch constructs the computation graph dynamically as computations are performed, offering greater flexibility and making it an ideal choice for research and rapid prototyping. This capability aligns well with the iterative nature of this work, where experimentation and adjustments are crucial to achieving optimal results.

Below is an overview of the different components of the Transformer model, explaining their functionality as introduced by Vaswani *et al.* (2017) and detailing how they have been adapted and implemented specifically for the morphological segmentation models in this study.

1. Multi-Head Self-Attention (MSA)

As mentioned earlier, the Transformer model adopts the attention mechanism with *Query-Key-Value* (QKV) model. As such, the inputs $X \in \mathbb{R}^{n \times 1}$ are linearly transformed into these three parts, where the queries $Q \in \mathbb{R}^{n \times d_q}$, keys $K \in \mathbb{R}^{n \times d_k}$ and values $V \in \mathbb{R}^{n \times d_v}$. In this, n is for sequence length, where else d , d_q , d_k , and d_v are dimensions of inputs, queries, values, and keys, respectively. The Transformer then applies a scaled dot product on these vectors in accordance with the formula below.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5.1)$$

The dot product of the query with all keys is scaled by a factor of $\frac{1}{\sqrt{d_k}}$ to prevent the exploding gradient problem in the *softmax* function. Without this scaling, large values of d_k could result in excessively large dot products, leading to very small gradients after applying *softmax*.

In Transformer models, this process is not applied just once, but multiple times in parallel through Multi-Head Self-Attention (MSA). Instead of performing a single attention operation using d -dimensional Q , K , and V , the model linearly projects these inputs into d_q , d_k , and d_v dimensions, respectively, h times using different learned linear transformations. These projections generate multiple “heads”, allowing the model to compute attention functions for each head independently.

Each of the h parallel attention heads outputs a d_v -dimensional vector. These outputs are then concatenated and passed through a final linear layer to produce the aggregated result. This mechanism enables the model to attend to multiple aspects of the input representation simultaneously, capturing richer and more diverse contextual information.

As illustrated in Figure 5.1, the Transformer architecture employs MSA in three distinct areas:

- (a) *Encoder Self-Attention*: In the encoder layer, a process called *self-attention* is employed. Here, all Q , K , and V vectors are derived from the same input layer within the encoder stack. This allows each position in the input sequence to focus on other positions, enabling the creation of a context-aware representation of the input.
- (b) *Decoder Masked Self-Attention*: The decoder uses *masked self-attention*, where self-attention is restricted such that each query at a position can only attend to key-value pairs up to and including that position. This ensures that the decoder does not “look ahead” by masking out future positions, maintaining the autoregressive nature of the decoding process.
- (c) *Cross-Attention*: In this mechanism, K and V are derived from the encoder outputs, while Q is projected from the previous layer of the decoder. This enables the decoder to “attend” to relevant parts of the encoded input sequence, facilitating the generation of each token in the output sequence.

2. Position-Wise Feed-Forward Neural Network

Self-attention ensures that token representations incorporate context from the entire sequence. However, another essential component of the Transformer architecture is the position-wise feed-forward network (PFFN), which is applied independently to

each token representation. This process improves the ability of the model to capture complex patterns by introducing additional transformations at each position. This mathematical equation can be expressed as follows:

$$\text{FFN}(H') = \text{ReLU}(H'W_1 + b_1)W_2 + b_2 \quad (5.2)$$

In this equation: H' represents the token representation of the input at a specific position, where W_1 and W_2 are learnable weight matrices, b_1 and b_2 are learnable bias vectors and $\text{ReLU}(\cdot)$ is the Rectified Linear Unit activation function. As such, PFFN is presented as a function that contains two linear transformations that incorporate the ReLU activation function in between.

The first transformation $(H'W_1 + b_1)$ expands the dimensionality of the token representations from d to d_{ff} , introducing a more extensive feature space where complex relationships can be learned. The second transformation $(\text{ReLU}(H'W_1 + b_1))W_2 + b_2$, reduces the dimensionality back from d_{ff} to d , ensuring that the output matches the input size.

However, it is important to note that these linear transformations tend to differ from layer to layer in terms of parameters. This is the case because $W_1 \in \mathbb{R}^{d \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d}$, $b_1 \in \mathbb{R}^{d_{ff}}$ and $b_2 \in \mathbb{R}^d$ are all trainable parameters. Lastly, the applied ReLU activation function introduces non-linearity, which enables the model to learn complex mappings.

3. Residual Connection and Normalisation

To enhance the depth of the model while maintaining stable training and promoting faster convergence, the Transformer architecture incorporates residual connections and layer normalisation around each layer.

The residual connections (He *et al.*, 2016) act as shortcuts that add the input of a layer directly to its output. This mechanism allows the model to focus on learning incremental changes to the input representation at each layer, rather than relearning the entire representation. By facilitating the flow of gradients during backpropagation, residual connections help to mitigate common issues such as vanishing or exploding gradients, which are often associated with deeper models, particularly in RNNs. This enables the Transformer to effectively utilise deeper architectures without compromising stability.

After applying the residual connections, the activations of each layer are normalised through layer normalisation (Ba, 2016). This process ensures that the inputs remain on a consistent scale, stabilising the training dynamics and further accelerating convergence. Layer normalisation adjusts the input to have a mean of zero and a standard deviation of one, improving numerical stability and making the training process more robust.

4. Positional Encoding

The Transformer sublayers differ fundamentally from traditional sequence-to-sequence models in that they are permutation equivariant, meaning that the relationships between all pairs of input tokens are computed using dot products, without inherent sensitivity to token order. This property arises because the attention weights in the self-attention mechanism depend solely on the content of the tokens, not on their positional order.

Unlike Recurrent Neural Networks (RNNs), which process tokens sequentially and inherently encode positional dependencies through recurrence, or Convolutional Neural Networks (CNNs), which capture local dependencies through convolutional filters, Transformers lack a built-in mechanism to represent positional information. Instead, they rely on positional encodings to introduce order awareness into the model.

The lack of an inherent positional encoding mechanism in Transformers can be advantageous for tasks involving unordered inputs, such as set-based data. However, for tasks where the sequence order is critical, such as language translation, morphological segmentation, or image processing, preserving the positional information of tokens is essential for accurate modelling and output.

To address the lack of inherent positional awareness, the Transformer architecture incorporates a mechanism called positional encoding, which encodes the positional information of each token in the sequence. This encoding can either be fixed or learned. The standard approach involves representing positional information as vectors, which are then added to the token embeddings as additional input to the model. This enables the Transformer to capture the order of tokens, ensuring that positional relationships are preserved during processing.

5.3.1.1 Transformer Models Implementation

The preceding section provided a deconstruction of the Transformer architecture, in order to simplify the explanation of its implementation and training process undertaken in this study. This study adopts a Transformer architecture that is closely aligned with the standard architecture introduced by Vaswani *et al.* (2017), to ensure that the fundamental principles of the original design are preserved.

1. Input Tokenization

The input preparation for the Transformer model involves tokenisation, embedding, and sequence alignment. Tokenisation, adapted from the HuggingFace Tokenizer¹, splits raw text into a sequence of characters. Each token is assigned a unique ID, with special tokens like *[PAD]*, *[SOS]*, and *[EOS]* used for padding, marking the

¹https://huggingface.co/docs/transformers/en/main_classes/tokenizer

start of a sequence, and indicating its end. These IDs ensure compatibility with the numerical processing required by the Transformer.

Sequences are standardised to a fixed length of $n = 50$, where words shorter than this length are padded using *[PAD]*. This approach ensures uniform input dimensions, facilitating batch processing and seamless integration into the model.

2. Token Embedding and Masking

Since the tokens are in numerical form, they are then converted into dense vector representations through an embedding layer and scaled to stabilise training. Since the Transformer lacks inherent positional awareness, positional encodings are added to embeddings to encode token order using sine and cosine functions in a similar fashion as the standard architecture presented by Vaswani *et al.* (2017). This is crucial for capturing sequence dependencies in morphological segmentation, since the position of each character in a sequence has an important function.

Following embedding, input tokens undergo masking. On the encoder side, padding tokens are masked to ensure they are ignored during self-attention computations. On the decoder side, *causal masking* is applied to restrict attention to previous and current tokens, preventing the decoder from accessing future tokens during the generation process.

3. Multi-Head Self-Attention (MSA)

As mentioned, the Transformer architecture employs the attention mechanism in three different ways. In the first process in relation to the model being designed, this happens on the encoder side. In this process, each character in the input sequence attends to all other characters, including itself, to compute a weighted representation of the sequence. This is achieved by transforming the input embeddings generated in the above-described step into Q , K , and V , which are used to calculate attention scores. The attention scores, derived as scaled dot products between the Queries and Keys, determine how much focus each character should place on every other character in the sequence. These scores are then normalised using a softmax function, and the resulting weights are applied to the V to compute the final token representations.

To further enhance the ability to capture intricate patterns within the sequence, the model also employs MSA. Instead of computing a single set of attention scores, the model splits the Q , K , and V vectors into N separate heads. Each head operates independently, calculating its own attention scores and weighted representations. This parallelisation allows the model to attend to different parts of the sequence simultaneously, capturing diverse dependencies.

In the baseline implementation, the number of attention heads was kept consistent with the original Transformer paper, resulting in a stack of six encoder layers $N = 6$. Each encoder layer is followed by a simple position-wise feed-forward network (PFFN), which further processes the token representations. Similarly, the decoder

is implemented with the same number of identical layers as the encoder. However, an additional multi-head self-attention (MSA) module is introduced on the decoder side, positioned between the masked self-attention and the PFFN. The dimensionality of both the input and output of each layer is also kept consistent with the original implementation, with $d = 512$ and the inner feed-forward layer having a dimensionality of $d_{ff} = 2048$.

4. Residual Connection and Normalisation

In this implementation, the residual connections are developed around each of the two modules, i.e., MSA and PFFN. This is possible by ensuring a standardised out dimension of $d = 512$. This is also implemented in accordance with the original Transformer, with the benefit of alleviating the vanishing gradient problem and ensuring simple convergence of the model during the training process.

After the residual connection, the implementation then employs layer normalisation to normalise the activations. This is based on the following equation:

$$\text{Normalized Output} = \alpha \cdot \frac{x - \mu}{\sigma + \epsilon} + \beta$$

where α and β are learnable parameters, μ is the mean, σ is the standard deviation, and ϵ is a small constant added to prevent division by zero. In this implementation, ϵ is set to 1×10^{-5} .

5.3.1.2 Training the Transformer Models

With the implementation steps of the Transformer model described above, which aligns with the architecture depicted in Figure 5.1, the next phase involves detailing the training process for the three models. These models are trained on datasets segmented at varying levels of granularity to evaluate the impact of segmentation granularity on model performance.

5.3.1.2.1 Data Splitting

Before the training process can be discussed, it is necessary to briefly mention the data management process for training this model. The data plays an important role in training, validating, and evaluating machine learning models. This means that how data is collected and split into these sections, and the process for each model, are as important as training the model. It is even more important, because a systematic comparison had to be made to ensure that the models use the same data and follow similar data management strategies.

As detailed in Chapter 4, a large dataset of distinct words (X) and their corresponding surface segmentation outputs (Y) was created. A representative portion of this dataset

was randomly sampled for model development. To ensure robust model evaluation and to mitigate risks of *overfitting* and *underfitting*, a systematic data splitting strategy was employed. Before delving into how these two phenomena are addressed, it is crucial first to explain the data-splitting strategy and its role in influencing model performance.

In the machine learning model, the data is commonly divided into three distinct subsets: training, validation, and evaluation. The training set is used to teach the model, enabling it to learn and generalise underlying data patterns. The validation set is that portion of data that has been set aside to verify the model performance as it trains; this happens through estimating the prediction error (validation loss) of the model and is subsequently used to detect when to stop training (Montesinos López *et al.*, 2022). When the final model has been attained, one uses the test set, which is entirely independent of training and validation, to evaluate the final model’s performance, ensuring unbiased assessment as it remains uninvolved in the model selection process.

5.3.1.2.2 Overfitting and Underfitting

If data is not appropriately split to ensure that subsets are completely distinct yet sampled from the same population, this could lead to biased results or models that suffer from underfitting or overfitting (Hastie, 2009). Underfitting occurs when a model is trained on insufficient, unrepresentative data or when the dataset lacks significant predictive features. This results in an overly simplistic model that fails to capture the underlying patterns in the data. Consequently, the model performs poorly on both the training set and unseen data, indicating an inability to generalise effectively.

In contrast to underfitting, there is overfitting, where a model is overly complex, where it has learned the training data so well that it even fails to generalise well on a test set (Montesinos López *et al.*, 2022). Such a model learns too much about the training set, including the noise and systematic part of the data, instead of the hidden patterns that could help it generalise better on unseen sets. This results in a model that performs exceptionally well on the training set, achieving high accuracy, but fails to generalise, producing significantly lower accuracy on the test set.

Reflecting on the data-splitting process, Chapter 4 detailed the preparation of two datasets for training supervised machine learning models. The first dataset was produced by the ZRG system without undergoing data augmentation, consisting of approximately 76 271 data points. The second dataset, which incorporated data augmentation, contained 367 178 data points. Both datasets comprised isiZulu words and their respective morphological segmentations, following three distinct segmentation strategies: *Lin A*, *Lin B*, and *Lin C*.

Due to computational resource constraints, a subset of 200 000 data points was randomly sampled from the augmented dataset 367 178, while the unaugmented dataset was used

in its entirety. Each dataset was then systematically divided into three subsets: 80% for training, 10% for validation, and 10% for testing. To maintain the integrity of the evaluation process, care was taken to ensure that there was no overlap between these subsets. This process is summarised in Figure 5.1.

Table 5.1: Summary of data splitting for training, validation, and testing.

Dataset	Total Size	Training (80%)	Validation (10%)	Testing (10%)
Unaugmented Dataset	76 271	61 017	7 627	7 627
Augmented Dataset (Subset)	200 000	160 000	20 000	20 000

During model design and training, the primary focus was on the training and validation sets. The training set was used to teach the model, while the validation set monitored performance and guided model optimisation. The final evaluation set, reserved exclusively for testing, will be discussed in detail in the next chapter.

5.3.1.2.3 Default Parameter Configuration

For the training process, the Transformer model architecture was initialised with standardised parameters aligned with the original configuration described by Vaswani *et al.* (2017), as outlined in Table 5.2. While these parameters provide a robust baseline, it is well-documented that the optimal parameter values can vary significantly, depending on the specific task (Hrycej *et al.*, 2022; Mahabadi *et al.*, 2021).

Table 5.2: Transformer model architecture parameters.

Parameter	Value
Learning rate	0.1
Model Dimension (d)	512
Number of Layers (N)	6
Number of Attention Heads (h)	8
Feed-Forward Dimension (d_{ff})	2 048
Dropout Rate	0.1
Maximum Gradient Norm	1.0
Label Smoothing Factor	0.1

5.3.1.2.4 Hyper-Parameter Optimisation (HPO)

Key parameters such as the model dimension ($d = 128, 512, 1024, 2048, \dots$), the number of attention heads $4, 8, 16, \dots$, and the feed-forward network size ($d_{ff} = 512, 1024, 2048, \dots$) can significantly impact model performance. Exploring the different permutations of these parameters to improve accuracy while considering resource constraints, such as hardware

limitations and training time, transforms the configuration process into an optimisation problem.

Furthermore, manually testing combinations of these parameters can be labour-intensive, error-prone, and time-consuming. To address this, automated hyper-parameter optimisation (HPO) techniques, such as grid search, random search, or more advanced methods like Bayesian optimisation, are often employed (Bergstra & Bengio, 2012; Mockus, 1989).

To overcome the inefficiencies of using default parameters or a brute-force method for obtaining optimal parameters, this study employed an HPO strategy to identify the best configuration for the Transformer model. Specifically, this study utilised Ray.Tune (Liaw *et al.*, 2018), an open-source library for hyper-parameter tuning that implements a wide range of state-of-the-art search algorithms. For the present study, OptunaSearch, a Bayesian optimisation algorithm based on the Tree-structured Parzen Estimator (TPE) (Bergstra *et al.*, 2011), was used to explore the hyper-parameter space efficiently. This algorithm was selected for its ability to balance exploration and exploitation, enabling the identification of optimal parameters in fewer iterations compared to grid or random search. The search space, outlined in Table 5.3, was designed with consideration of the computational resources available.

Table 5.3: Hyper-parameter search space configuration.

Hyper-parameter	Search Space
Learning Rate	log-uniform(10^{-5} , 10^{-3})
Model Dimension (d)	{128, 256, 512, 1024}
Number of Layers	{3, 4, 6}
Number of Attention Heads	{4, 8, 16}
Feed-Forward Dimension (d_{ff})	{512, 1024, 2048}
Dropout Rate	uniform(0.1, 0.3)
Label Smoothing	uniform(0.0, 0.2)
Maximum Gradient Norm	uniform(0.5, 2.0)

This study utilised a high-performance system featuring an NVIDIA A100 GPU with 80GB of memory and 64-core processors. Ray.Tune’s capability to execute parallel tuning across multiple computational units, leveraging predefined GPU or CPU cores, substantially expedited the optimisation process. For this implementation, 32 cores were allocated for parallel tuning, allowing for the efficient exploration of the hyper-parameter space while minimising runtime.

By integrating these computational resources with Ray.Tune’s advanced optimisation features, the study implemented a systematic and efficient HPO strategy, well-aligned with the computational environment and the specific task requirements.

Table 5.2 includes several parameters utilised during model training that have not yet been discussed in detail, such as the *dropout rate* (Srivastava *et al.*, 2014), the *maximum*

gradient norm (Mikolov *et al.*, 2012), and the *label smoothing factor* (Szegedy *et al.*, 2016). These parameters were incorporated as part of the regularisation strategy, which involves techniques aimed at reducing the generalisation error rather than the training error (Goodfellow *et al.*, 2016). The following section provides a brief discussion of these regularisation strategies and their role in optimising the performance of the developed Transformer models, which are also features in the LTSM models.

1. *Dropout Rate*: The dropout rate is a probability that determines the random selection of neurons to be deactivated during the training of the model. This technique, known as dropout, is designed to make the neural network less sensitive to specific weights of the neurons, thereby reducing the likelihood of overfitting (Yu & Zhu, 2020). By simplifying the Transformer’s complexity, dropout enhances its generalisation ability.

For HPO, the dropout rate was explored within a range of 10%-30%, allowing for the selection of an optimal rate for models across different segmentation granularities. The choice of the dropout rate balances model complexity and generalisation, with higher rates simplifying the model further. These values were subsequently applied to various Transformer components to optimise performance while mitigating the risk of overfitting.

2. *Maximum Gradient Norm*: The second parameter that was considered is the *maximum gradient norm*, which is critical for gradient clipping, a technique used to address the problem of exploding gradients (Mikolov *et al.*, 2012). This parameter establishes a threshold for the norm (magnitude) of the gradients. If the gradients exceed this value during training, they are clipped or scaled down to remain within the defined limit. This mechanism stabilises the training process, ensuring that the model converges smoothly while reducing the risk of overfitting caused by large gradient updates (Ye *et al.*, 2024).

For HPO, the optimal value for this parameter was explored within the range of 0.5 to 2, as outlined in Table 5.3. This range was chosen to balance the need for stability and effective gradient updates while accommodating the variability of different segmentation granularities.

3. *Label Smoothing Factor*: The third parameter employed for regularisation is the *label smoothing factor*, introduced by Szegedy *et al.* (2016). Typically, when training a model to optimise losses, the observed class is treated as a one-point distribution, assigning the entire probability mass to the label (Lienen & Hüllermeier, 2021). However, such extreme distributions can make the model prone to overconfidence and overfitting. To mitigate this, Szegedy *et al.* (2016) proposed label smoothing, which redistributes a portion of the probability mass from the observed class across other classes. This adjustment ensures the distribution is less extreme, promoting stability and generalisation.

In the context of morphological segmentation, label smoothing proved particularly beneficial. Morphological boundaries often involve ambiguities, making it challenging to determine their precise placement. Label smoothing helps the model to become more robust to such ambiguities, enabling it to generalise better across different words and segmentation patterns. For this study, a smoothing factor of 0.1 was used as a default for all three segmentation models. An optimal value was further explored through HPO within the range of 0 to 0.2.

4. *Learning Rate*: The fourth parameter that is crucial for model optimisation is the *learning rate*, a fundamental hyper-parameter that governs the step size at which the model updates its weights during training. The learning rate directly influences how quickly or effectively a model converges to the optimal solution. A learning rate that is too high can result in the model overshooting the optimal point, leading to instability or divergence. Conversely, a learning rate that is too low can slow convergence or cause the training algorithm to stagnate in local minima (Smith, 2017). Thus, dynamically adjusting the learning rate throughout training is essential for achieving better accuracy, making it one of the most critical parameters to tune (Smith, 2017).

In this study, the learning rate was carefully tuned due to its substantial impact on the training dynamics of the segmentation models. The initial default value was set at 1×10^{-5} , with optimal values explored through HPO within a log-uniform range of 1×10^{-5} to 1×10^{-3} . This range provided a balance between fine-grained adjustments and larger updates, enabling flexibility in finding the best learning rate for each segmentation model. Furthermore, the Adaptive Learning Rate (Adam) (Kingma & Ba, 2014) optimiser was employed to adjust the learning rate dynamically for each parameter. By leveraging historical gradient magnitudes, Adam assigns unique learning rates to individual parameters, ensuring that the rate evolves adaptively as training progresses, ultimately enhancing the model's performance.

Throughout the tuning process, the primary objective was to identify the optimal set of hyper-parameters that minimised the *validation loss*, a key metric indicating the model's ability to generalise to unseen data. This approach ensured that the trained model performed effectively not only on the training set, but also on validation and test data, reducing the risk of overfitting. During training, validation loss was carefully monitored, because an increase in this metric would suggest that the model was failing to generalise well. To mitigate this risk, early stopping was employed both during HPO and the training of the final models.

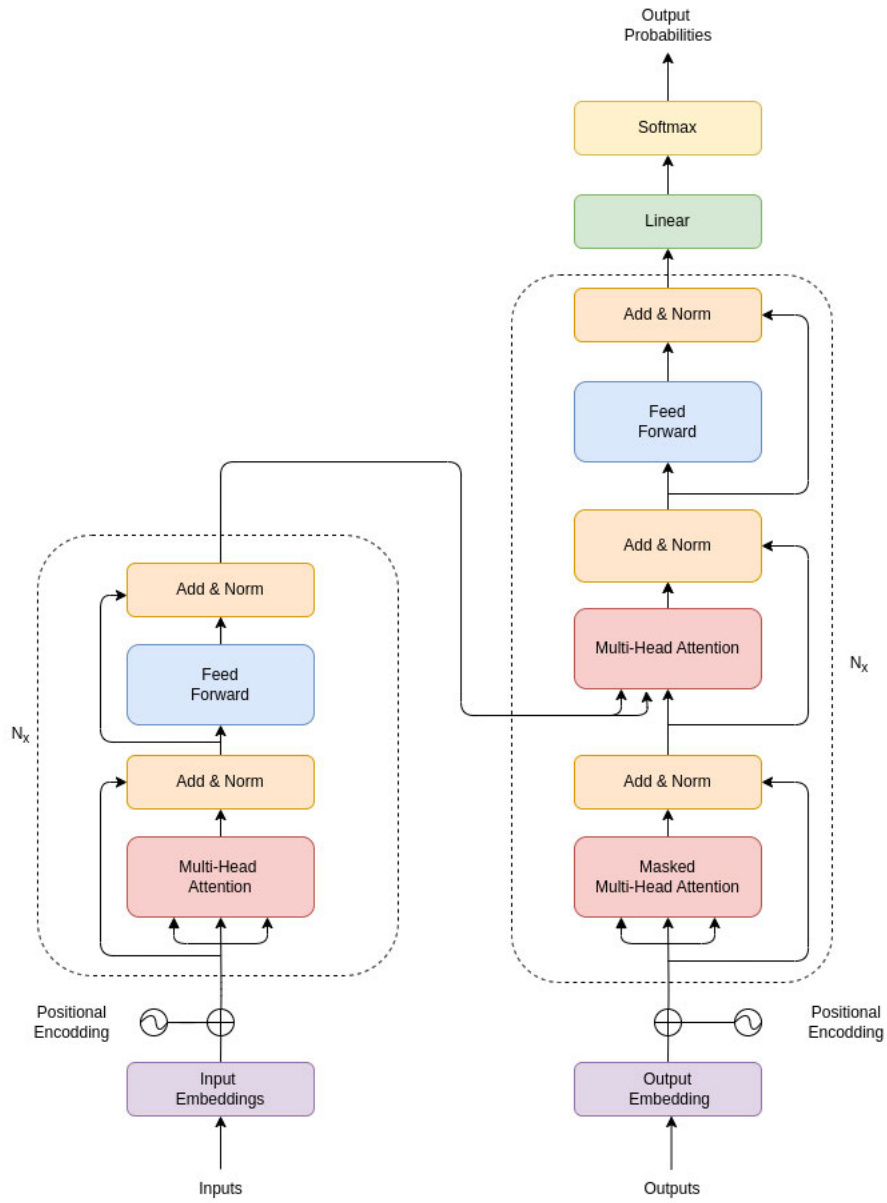


Figure 5.1: An adapted vanilla Transformer architecture.

5.3.1.2.5 Early Stopping

Early stopping or the *stopping criterion* (Morgan & Bourlard, 1989) is a widely used regularisation technique in machine learning to prevent overfitting (Goodfellow *et al.*, 2016). It works by monitoring metrics such as validation loss or accuracy during training and halting the process when validation performance ceases to improve, thereby maintaining the model’s generalisation ability. During HPO, an *Asynchronous Successive Halving Algorithm (ASHA) scheduler* was utilised, which automatically terminated underperforming trials early to save on computational resources. For the actual model training, an early stopping mechanism was implemented to monitor validation loss, with a patience value of five. This means that the training process terminated after five consecutive epochs of validation loss stagnation or incremental increase, ensuring that the model selected was the one with the lowest validation loss. The maximum number of epochs was set to 60 for model training and 5 for HPO, with each search limited to 10 trials.

Table 5.4: Final model configurations for segmenters across different linearisation strategies.

Parameter	Segmenter One	Segmenter Two	Segmenter Three
Learning Rate	4.801×10^{-4}	2.951×10^{-4}	3.105×10^{-4}
Model Dimension (d)	128	1,024	128
Number of Layers	3	6	4
Number of Attention Heads	16	16	8
Feed-Forward Dimension (d_{ff})	1,024	2,048	512
Dropout Rate	0.190	0.246	0.214
Maximum Gradient Norm	0.525	1.043	1.119
Label Smoothing	9.496×10^{-4}	9.885×10^{-3}	8.008×10^{-4}

Table 5.4 presents the optimal hyper-parameters identified through the HPO technique for the *Lin A*, *Lin B*, and *Lin C* datasets, corresponding to Segmenters One, Two, and Three, respectively. The variation in optimal hyper-parameters across these datasets highlights that each segmentation granularity captures distinct underlying morphological patterns, requiring tailored model configurations to achieve optimal performance. The learning rates, model dimensions, dropout rates, and gradient norms vary notably among the three segmenters, indicating the unique training dynamics and complexity associated with each level of segmentation granularity.

The optimal hyper-parameters obtained here were subsequently used to train final models whose validation and test performance results will be comprehensively discussed in Chapter 6.

5.3.2 Long Short-Term Memory Models

Another sequence-to-sequence architecture that was trained in order to select a better-performing model in terms of segmentation accuracy was the LSTM architecture. This architecture is based on LSTM networks, chosen due to their prevalence in related studies. While GRUs could also be tested, as they share similarities in design with LSTMs, there is limited research on implementing GRUs for morphological segmentation. Consequently, this study follows the trend by adopting an LSTM model for the task.

5.3.2.1 Long Short-Term Memory Adaptation

The LSTM implemented for this study was based on the Keras example for LSTM sequence-to-sequence learning². Given this foundation, this section does not provide a deep exploration of the implementation itself, but rather focuses on the modifications made to tailor the LSTM for the specific task of morphological segmentation. Additionally, key components of the model, which have been discussed in previous chapters, are revisited in the context of their relevance to the present discussion.

The LSTM segmentation model implements an encoder-decoder architecture designed for character-level morphological segmentation. The encoder consists of a single LSTM layer with 256 hidden units, processing one-hot encoded character sequences and employing recurrent dropout 0.5 for regularisation. This encoder outputs state vectors (hidden and cell states) that capture the input sequence’s morphological patterns. Since the LSTM architecture is a sequence-to-sequence model, it shares several characteristics with the Transformer model implemented above. For instance, during data preparation and processing, the dataset was split into 80%, 10%, and 10% for training, validation, and testing sets, respectively.

For optimisation, the architecture utilises the Adam optimiser (Kingma & Ba, 2014) with an exponential learning rate decay schedule, starting with an initial learning rate of 1×10^{-5} and a decay rate of 0.9. To enhance training stability and mitigate overfitting, label smoothing with a factor of 0.1 and gradient clipping with a maximum norm of 1.0 are applied. Additionally, early stopping is employed to monitor the validation loss, with a *patience* of five epochs to halt training if no improvement is observed.

The decoder mirrors the encoder’s architecture with a matching LSTM layer initialised using the encoder’s final states. It generates segmented outputs character-by-character through a dense layer with softmax activation, mapping to the vocabulary size. The decoder shares the encoder’s dropout rate, maintaining consistent regularisation across the model.

²https://github.com/keras-team/keras-io/blob/master/examples/nlp/lstm_seq2seq.py

The tokenisation strategy for the LSTM architecture is similar to the one used for the Transformer architecture, employing character-level tokenisation. Each character token is assigned a unique ID, with special tokens such as $[PAD]$, $[SOS]$, and $[EOS]$ incorporated to denote padding, the start of a sequence, and the end of a sequence, respectively. The sequence length is dynamically determined based on the word with the longest sequence length, with an additional two tokens added to account for the $[SOS]$ and $[EOS]$ tokens.

Table 5.5 summarises the initial configurations for the LSTM architecture before HPO. A notable addition in this configuration is the inclusion of the *Decay Steps* and *Decay Rate* parameters. These parameters act as regularisation techniques, working in tandem with the learning rate to dynamically adjust its value during training, thereby improving convergence and model performance. Specifically, the decay step determines the number of training iterations after which the learning rate is updated, while the decay rate defines the factor by which the learning rate is multiplied at each decay step (Albahli *et al.*, 2020).

Table 5.5: LSTM model default configuration.

Parameter	Value
Label Smoothing	0.1
Gradient Clip Norm	1.0
Latent Dimension	256
Initial Learning Rate	1×10^{-4}
Decay Steps	10,000
Decay Rate	0.9
Dropout Rate	0.1

5.3.2.2 Long Short-Term Memory Hyper-Parameter Optimisation

In a similar fashion, the default hyper-parameters were used to initialise the LSTM models during the search for optimal hyper-parameters. As with the Transformer HPO, the Tune toolkit was utilised, requiring the definition of a search space, as presented in Table 5.6. The same arguments were employed for HPO, including a fixed epoch number of 5 and 10 trials, and early stopping to prevent overfitting. For the final training of the LSTM models, an epoch limit of 60 was set, with a patience value of 5 for early stopping.

The results of the HPO process, detailing the optimal hyper-parameters found for the LSTM models across the three different segmentation granularities, are shown in Table 5.7.

Table 5.6: Hyper-parameter search space.

Hyper-parameter	Search Space
Latent Dimension	{128, 256, 512, 1024}
Dropout Rate	uniform(0.1, 0.3)
Initial Learning Rate	log-uniform(10^{-5} , 10^{-3})
Decay Rate	uniform(0.8, 0.99)
Label Smoothing	uniform(0.05, 0.15)
Gradient Clip Norm	uniform(0.5, 2.0)

Table 5.7: LSTM optimal hyper-parameters for Segmenters One, Two, and Three.

Parameter	Segmenter One	Segmenter Two	Segmenter Three
Label Smoothing	0.053	0.080	0.057
Gradient Clip Norm	1.657	1.605	0.525
Latent Dimension	1,024	256	512
Initial Learning Rate	1.602×10^{-4}	5.530×10^{-4}	5.361×10^{-4}
Decay Rate	0.971	0.968	0.805
Dropout Rate	0.260	0.226	0.213

5.3.3 Conditional Random Field Models

Previously, when discussing the statistical-based models that are commonly used for morphological segmentation, the CRF was one of the architecture that was introduced as a graphical sequence labelling model that uses probabilistic approaches to identify and label components of a sequence (Rodrigues *et al.*, 2014). The CRF are designed to model sequential dependencies between adjacent elements. In segmentation, the task is to determine where morpheme boundaries lie within a given sequence of characters (orthographic word).

CRFs excel in modelling these dependencies by assigning labels to each character in a sequence, indicating whether it belongs to a specific morpheme or represents a morpheme boundary. For morphological segmentation, the input to a CRF consists of orthographic words, and the output is a sequence of labels that correspond with the segmentation of the word.

5.3.3.1 Conditional Random Field Adaptation

The CRF implementation used in this study is adapted from the work of Moeng *et al.* (2022), with the codebase publicly available online on GitHub³. In their implementation, the labels are represented as B, M, E, and S, which encode the segmentation of characters as follows:

³https://github.com/DarkPr0digy/MORPH_SEGMENT

- **B**: The character is at the beginning of a new morpheme.
- **M**: The character is part of the current morpheme.
- **E**: The character is at the end of the current morpheme.
- **S**: The character forms a single-length morpheme.

Consider the isiZulu word *ngizobabona*, which can be segmented as *ngi-zo-ba-bona*. The segmentation and corresponding BMES (**B**egin-**M**iddle-**E**nd-**S**ingle) labels are determined as follows:

1. Segment: *ngi*
 Characters: ‘n’ (B), ‘g’ (M), ‘i’ (E)
 BMES Label: BME
2. Segment: *zo*
 Characters: ‘z’ (B), ‘o’ (End - E)
 BMES Label: BE
3. Segment: *ba*
 Characters: ‘b’ (B), ‘a’ (E)
 BMES Label: BE
4. Segment: *bona*
 Characters: ‘b’ (B), ‘o’ (M), ‘n’ (M), ‘a’ (E)
 BMES Label: BMME

Final representation:

- Input Token: *ngizobabona*
- Segmented Form: *ngi-zo-ba-bona*
- BMES Labels: BME-BE-BE-BMME

As previously mentioned, CRFs require an effective feature extraction system to transform raw character sequences into rich feature representations for optimal performance. Specifically, each word is represented as a sequence of feature dictionaries, with one dictionary corresponding with each character. These feature dictionaries provide the CRF model with the necessary input to predict BMES labels for segmentation.

The adapted CRF model employed in this study includes a comprehensive set of features designed to capture various aspects of the morphological structure, ensuring a robust representation of the data for segmentation tasks.

1. *Character-Level Features*: The main function of this feature extractor is simply to look into the basic character-level attributes of a character(s), such as the character itself and its lowercase. This includes the *position-aware feature extractor*, which looks into the character's position within the word, such as whether it starts or ends. Another feature extractor is the *context window extractor* that captures the surrounding characters and their content as either they are a vowel or a consonant. The last feature extractor is *N-gram-based feature extractor*, which looks at the preceding and succeeding features of each character, which comes hand in identifying common affixes. These character-level feature extraction methods enable the model to effectively capture the context and orthographic patterns of characters, which are essential for accurate morphological segmentation.
2. *Pattern-Based Features*: This feature extractor simply looks at the relationship that exists within the characters, such as a *syllable-like pattern* feature extractor that seeks to classify the sequence of characters as a vowels (*V*) or Consonants (*C*). This also describes how characters transition through adjacent character pairs and how vowels are distributed. These features are essential for uncovering potential rules and patterns inherent in isiZulu morphotactics, thereby supporting more accurate segmentation.
3. *Positional Features*: This feature is concerned with the position of characters, including the word boundaries such as the start or end positions, the absolute position within the words, and the word length information.

The primary objective of this process is to format the input data such that each word is paired with its corresponding segmentation labels, transforming them into feature vectors suitable for CRF training. This transformation ensures that the sequential nature of the data is preserved while enriching it with linguistic and structural information. These enriched representations enable the CRF model to effectively learn morpheme boundaries during training, which will be discussed shortly. Consistent with the other models, the dataset was split into 80% for training, 10% for validation, and 10% for testing.

5.3.3.2 Conditional Random Field Training

The training process employs the *sklearn-crfsuite*⁴, a Python library that provides a convenient wrapper around CRFsuite, often referred to as *python-crfsuite*. In adapting the model, the system was configured to accept a dictionary of CRF parameters that govern its behaviour. These parameters include the algorithm type, regularisation settings, and transition constraints, ensuring that the model effectively fits the training data. This

⁴<http://sklearn-crfsuite.readthedocs.io/en/latest>

straightforward training approach is particularly suitable when predefined parameters are sufficient, or when HPO is not required.

5.3.3.3 Conditional Random Field Hyper-Parameter Optimisation

As highlighted in the previous sections, HPO is a crucial step in training models that perform effectively. To this end, an automated HPO process was implemented using Grid Search Cross-Validation (GridSearchCV) from *sklearn* library. GridSearchCV conducts an exhaustive search over specified parameter ranges or grids, employing cross-validation to evaluate various parameter combinations and identify those that yield optimal model performance (Adnan *et al.*, 2022). This approach frames the task of finding the best parameter combination as an optimisation problem aimed at maximising the F1 score. Table 5.8 outlines the basic parameters used in the initial model setup, while Table 5.9 presents the parameter combinations explored during the GridSearchCV process. The full model implementation is available on GitHub⁵.

Table 5.8: CRF model base parameters.

Parameter	Value
Algorithm	LBFGS
L1 regularization (c_1)	0.1
L2 regularization (c_2)	0.1
Min Frequence	2
Maximum iterations	100
All possible transitions	True

Table 5.9: Grid search parameter ranges for GridSearch.

Parameter	Search Values
L1 regularisation (c_1)	[0.1, 0.2, 0.3, 0.4]
L2 regularisation (c_2)	[0.1, 0.2, 0.3, 0.4]
Min Frequency	[2, 3, 4]
Maximum iterations	[100, 300, 500]

To explain the parameters used, the CRF implementation was configured to employ L-BFGS (Liu & Nocedal, 1989), a limited-memory optimisation algorithm specifically designed to solve large-scale nonlinear optimisation problems with constraints on certain variables. L-BFGS is particularly advantageous in morphological segmentation tasks, as Conditional Random Fields (CRFs) rely on a log-linear model that necessitates optimisation across all possible label sequences. Given the high-dimensional and nonlinear nature of this problem, L-BFGS provides an efficient and effective solution.

⁵https://github.com/Sthesha/supervised_surface_segmentation

For a large dataset, such as the one used in this study, and for a language with agglutinative morphology, the number of possible label sequences is inherently vast. This significantly increases the computational complexity and memory requirements for optimisation. The L-BFGS algorithm (Liu & Nocedal, 1989) addresses this challenge by approximating the inverse Hessian matrix, thereby reducing memory usage while maintaining the effective optimisation of the CRF parameters. This makes L-BFGS particularly well-suited for handling the computational demands of morphological segmentation tasks in languages with complex and rich morphological structures.

The second set of parameters optimised for the models includes the coefficients for L_1 and L_2 regularisation methods (Tibshirani, 1996), which are crucial for reducing overfitting and improving generalisation during model training. Regularisation achieves this by adding penalty terms to the loss function, thereby discouraging overly complex models. In the CRF implementation, c_1 and c_2 control the strength of L_1 and L_2 regularisation, respectively.

The c_1 coefficient promotes sparsity in the model by driving certain feature weights to zero, effectively helping the CRF with feature selection by filtering out irrelevant features during training (Goodfellow *et al.*, 2016). This is particularly useful when working with a large feature space, as it focuses the model on meaningful attributes. Conversely, the c_2 coefficient encourages weight shrinkage, ensuring that the weights remain small but non-zero. This helps to stabilise the model and to improve its generalisation, especially when dealing with agglutinative languages like isiZulu, which exhibit high morphological variation and complex feature spaces.

The HPO process searched across various combinations of c_1 and c_2 values to find the optimal balance between feature selection and weight stability, enabling the development of better-performing CRF models tailored to the unique challenges of morphological segmentation in agglutinative languages.

In addition to c_1 , which balances relevant features, the model implementation incorporates a *minimum frequency* parameter. This parameter acts as a threshold, filtering out rare and noisy features by considering only those that occur above a certain frequency. Another critical parameter is the *maximum iterations*, which defines the maximum number of iterations for model training. This parameter ensures convergence to an optimal solution while balancing computational resources and model performance. Larger iteration counts were observed to increase the risk of overfitting or premature termination due to resource constraints, whereas smaller iteration counts resulted in suboptimal models. Therefore, the maximum iteration range was carefully chosen, as shown in Table 5.9.

The optimisation of CRF parameters across different segmentation schemes underscores the importance of fine-tuning to achieve effective segmentation for isiZulu. Table 5.10 provides the optimal parameters for each segmenter, reflecting variations in segmentation granularity and model requirements.

For Segmenter One, c_1 was optimised to 0.10 and c_2 to 0.01, balancing feature sparsity and weight stability while addressing the specific characteristics of finer segmentation granularity. Segmenter Two showed a preference for less sparsity in feature selection, with c_1 at 0.05 and c_2 also at 0.05. In contrast, Segmenter Three required a higher c_1 of 0.20, emphasizing sparsity, while c_2 was adjusted to 0.01 to maintain stability in weights.

Across all segmenters, the *minimum frequency* parameter was consistently set to 1, ensuring that rare features were included to enhance generalisation. The *maximum iterations* parameter reached 2 000 for Segmenters One and Three, indicating their higher complexity and need for longer training. Segmenter Two, however, converged with 1 500 iterations, suggesting a less complex feature space for its segmentation granularity.

These variations in parameters across the segmenters illustrate the adaptability of the CRF model to the nuances of different segmentation granularities. The results emphasise the critical role of hyper-parameter optimisation in tailoring the model to address the linguistic and structural challenges of isiZulu. A detailed evaluation of these models will be presented in the subsequent chapter.

Table 5.10: Optimal CRF parameters across different segmentation schemes.

Parameter	Segmenter One	Segmenter Two	Segmenter Three
c_1	0.10	0.05	0.20
c_2	0.01	0.05	0.01
Max Iterations	2 000	1 500	2 000
Min Frequency	1	1	1

5.4 Chapter Summary

This chapter detailed the methodology and techniques employed to develop and optimise models for morphological segmentation in isiZulu, addressing *Sub-objective 3* as outlined in Chapter 1, namely to investigate the suitability of different machine learning models for the supervised segmentation task.

To fulfil this objective, different strategies were explored, including the adaptation and modification of LSTM and CRF models and the implementation of the Transformer model from scratch. This allows for a systematic comparative analysis, ultimately leading to the selection of the most suitable model. During model implementation and training, HPO strategies were applied across all models to determine the optimal configurations for final

training and evaluation. The models were also adapted to include similar features to ensure a fair comparison.

To maintain consistency, data splitting was standardised across all models, with an 80:10:10 ratio for training, validation, and testing. For the Transformer and LSTM models, the Ray.Tune toolkit was utilised, leveraging OptunaSearch, a Bayesian optimisation algorithm. This approach enabled efficient exploration of complex parameter spaces while balancing computational resources and model performance. Additionally, regularisation techniques such as dropout, gradient clipping, and label smoothing were applied to mitigate overfitting and improve generalisation; thus addressing the unique challenges posed by the agglutinative morphology of isiZulu.

In parallel, feature engineering played a critical role in adapting the CRF model to the segmentation task. Linguistic and structural features were integrated to capture complex morphotactic patterns inherent to isiZulu. Key parameters, such as c_1 and c_2 , were optimised using the GridSearchCV tool from the *sklearn* library. This exhaustive search method enabled the evaluation of various parameter combinations to identify those yielding the highest accuracy while maintaining stability and relevance across diverse segmentation granularities.

Once optimal hyper-parameters were identified, they were used to train different supervised machine learning models, each employing three distinct linearisation strategies based on data generated from ZRG, a rule-based system. This integration of rule-based and machine learning paradigms resulted in a hybrid segmentation framework. At this stage, *Sub-objective 3* was fulfilled, as the segmentation models were successfully trained on annotated data with different granularity levels. The next step is to evaluate their suitability through intrinsic evaluation metrics, including F1-score, BLEU, and chrF, to quantify their segmentation performance. The best-performing model will then be leveraged for further hypothesis investigation and will contribute to fulfilling the study's primary objective and other research goals.

Chapter 6

Experiment Evaluation

This chapter presents the evaluation of the models developed for the morphological segmentation of isiZulu, aligning with *Sub-objective 5*, as outlined in Section 1.4.2: to assess the performance of the created models both intrinsically and extrinsically.

In the previous chapter, three models, the Transformer, LSTM, and CRF-based models, were trained using optimal hyper-parameters identified through hyper-parameter optimisation (HPO). The focus now shifts to evaluating these models systematically, comparing their effectiveness across different segmentation granularities and training strategies.

This chapter serves as the final step in tying together and summarising the primary objective of investigating supervised surface segmentation using synthetic data. Additionally, it provides answers to the research questions presented in Section 1.4.1 and examines the formulated hypotheses, ensuring that the findings contribute meaningfully to the broader discussion on isiZulu morphological segmentation using a hybrid system to perform this task effectively.

This chapter is structured as follows: Section 6.1 begins with an overview of the loss graphs, providing insights into the training dynamics of the models. Section 6.2 presents an intrinsic evaluation of the trained models, comparing their segmentation performance, and ultimately selecting the most suitable machine learning technique among the three. This section also includes a comparison between models trained on augmented data and those trained on unaugmented data. In addition, the robustness and efficiency of the models are evaluated.

Section 6.3 evaluates the performance of the trained translation systems, followed by a discussion of the results. Finally, Section 6.4 provides a summary of the chapter and a brief perusal of the next steps.

6.1 Model Performance Analysis: Training and Validation

This section examines the training and validation dynamics of the Transformer and LSTM models employed during the training process, focusing on the evolution of the loss metrics. This analysis is applicable to deep learning models, because they rely on backpropagation and gradient-based learning to minimise a defined loss function. In contrast, the CRF model employs a probabilistic graphical approach, which maximises the conditional likelihood of the output sequence given the input features, rather than minimising a loss function through iterative weight updates. For the Transformer and LSTM models, validation loss served as a critical metric, providing possible insights into generalisation performance. The training process incorporated an early stopping mechanism, which terminated training if the validation loss did not decrease for five consecutive epochs.

6.1.1 Training Loss

Training loss is a fundamental metric that is used to measure how well a deep learning model is learning during training or fitting the training data (Jaca-Madariaga *et al.*, 2023). It quantifies the discrepancy between the model’s predicted outputs and the actual target values by computing a loss function. The goal of training is to minimise this loss so that the model can make more accurate predictions over time. For this study, the loss function evaluates how closely the predicted segmentations align with the ground-truth segmentations.

Figures 6.1 and 6.2 illustrate the training loss curves for different segmenters trained using the Transformer and LSTM architectures, respectively. The x-axis represents the number of training epochs, while the y-axis represents the loss value. A decreasing loss curve indicates that the model effectively captures morphological segmentation patterns, while high or fluctuating loss values may suggest challenges such as insufficient learning capacity, overfitting, or poor optimisation.

For the Transformer models, Figure 6.1 shows that Segmenter Two converges the fastest, with early stopping triggered at epoch 12 due to validation loss criteria. Segmenters One and Three exhibit a more gradual decline in loss, but ultimately achieve lower final loss values compared to Segmenter Two. This suggests that while Segmenter Two learns quickly, the slower convergence of Segmenters One and Three may allow for better generalisation.

In the LSTM models, shown in Figure 6.2, all segmenters exhibit a slower but consistent decline in training loss. Segmenter Three converges the fastest, around epoch 20, but maintains a slightly higher final loss. Segmenters One and Two take longer to converge (around epoch 30), but achieve comparable loss values, indicating more uniform learning behaviour across segmenters.

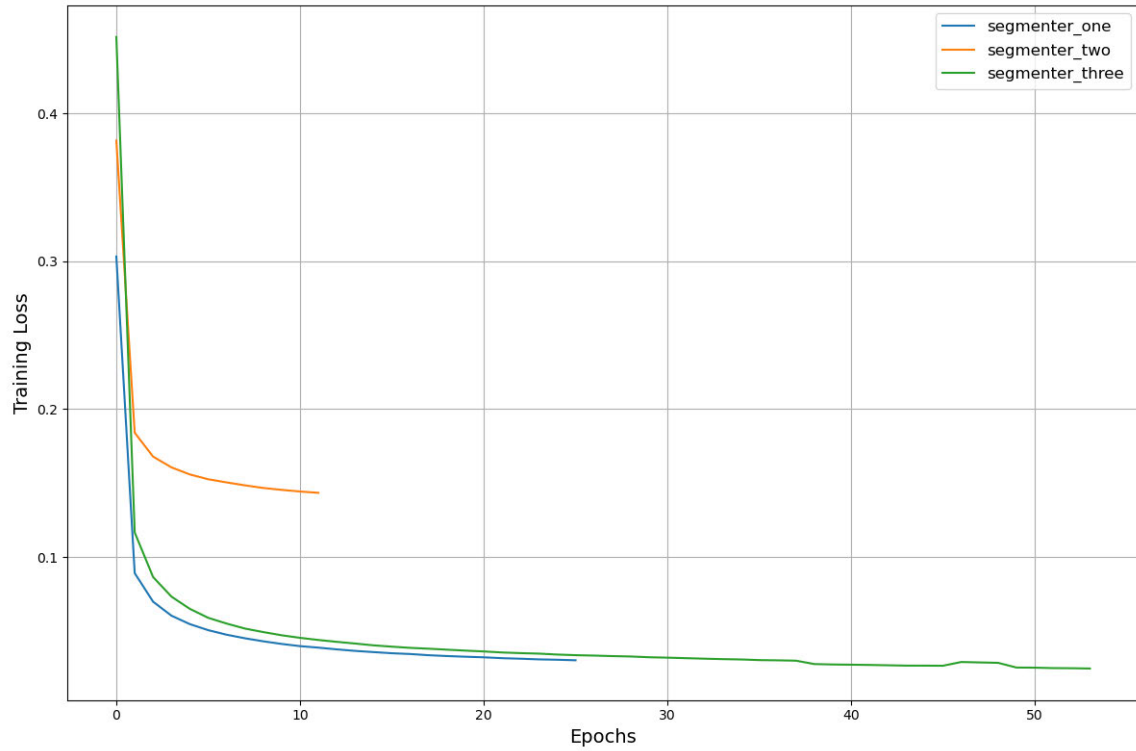


Figure 6.1: Transformer: training loss over epochs for different Segmenters.

Overall, the Transformer Transformer exhibits faster convergence but greater variability in final training loss, with Segmenters One and Three achieving lower loss values than Segmenter Two. In contrast, the LSTM model demonstrates more stable and consistent performance across segmenters, although with a longer training period.

The following sections presents an analysis of how these training dynamics translate into validation and test performance using the validation and training set, providing a more comprehensive assessment of each model's segmentation effectiveness.

6.1.2 Validation Loss

Validation loss is a critical metric used to assess a model's ability to generalise to unseen data. It represents the model error computed on a separate validation dataset that is not used during training (Bahree, 2024). Unlike training loss, which measures how well the model fits the training data, validation loss provides insight into whether the model is learning meaningful patterns or simply memorising the training examples. A consistently low validation loss indicates good generalisation, whereas a growing gap between training and validation loss suggests overfitting, meaning that the model performs well on training data, but struggles with unseen examples.

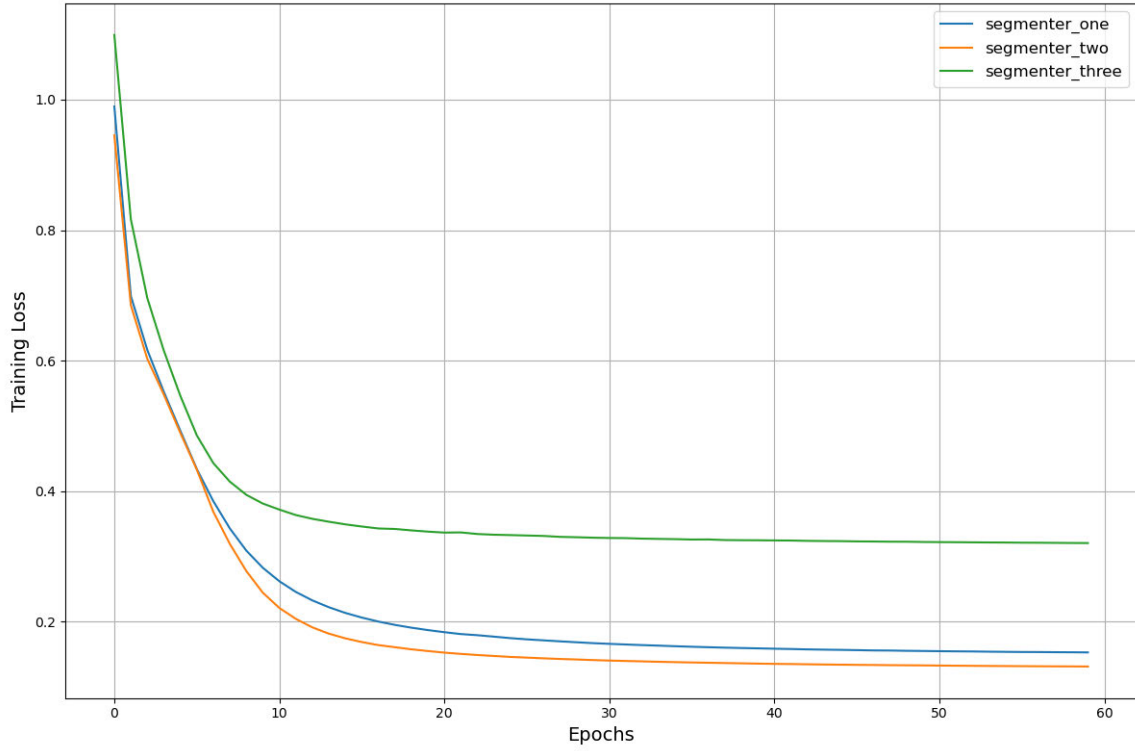


Figure 6.2: LSTM: training loss over epochs for different segmenters.

The trends observed for Transformer and LSTM segmenters in Figures 6.3 and 6.4, respectively, offer valuable information. For the Transformer architecture, Segmenter Two converges the fastest, achieving the lowest validation loss by epoch 10. However, fluctuations in its validation loss post-convergence suggest that it may be more sensitive to variations in the validation data, indicating potential overfitting. Segmenters One and Three, on the other hand, exhibit a more gradual decline in validation loss, with some spikes also, converging at later epochs but reaching comparable final loss values. This suggests that while they take longer to stabilise, they may generalise better across different data points.

For the LSTM model, Segmenter Two again achieves the lowest validation loss within the first 10 epochs, demonstrating strong generalisation on the validation set. Segmenters One and Three converge more gradually, with Segmenter One attaining a slightly better validation loss than Segmenter Three by the end of training. Notably, none of the LSTM models triggered the early stopping mechanism, because their validation loss continued to decrease steadily, although at a slower rate, throughout training. This suggests that while the LSTM models require more training time to reach optimal performance, they maintain a consistent learning trajectory without signs of overfitting.

Overall, the Transformer-based models displayed faster convergence and significantly lower validation loss ($\times 10^{-5}$) compared to the LSTM models, with Segmenter Two consistently

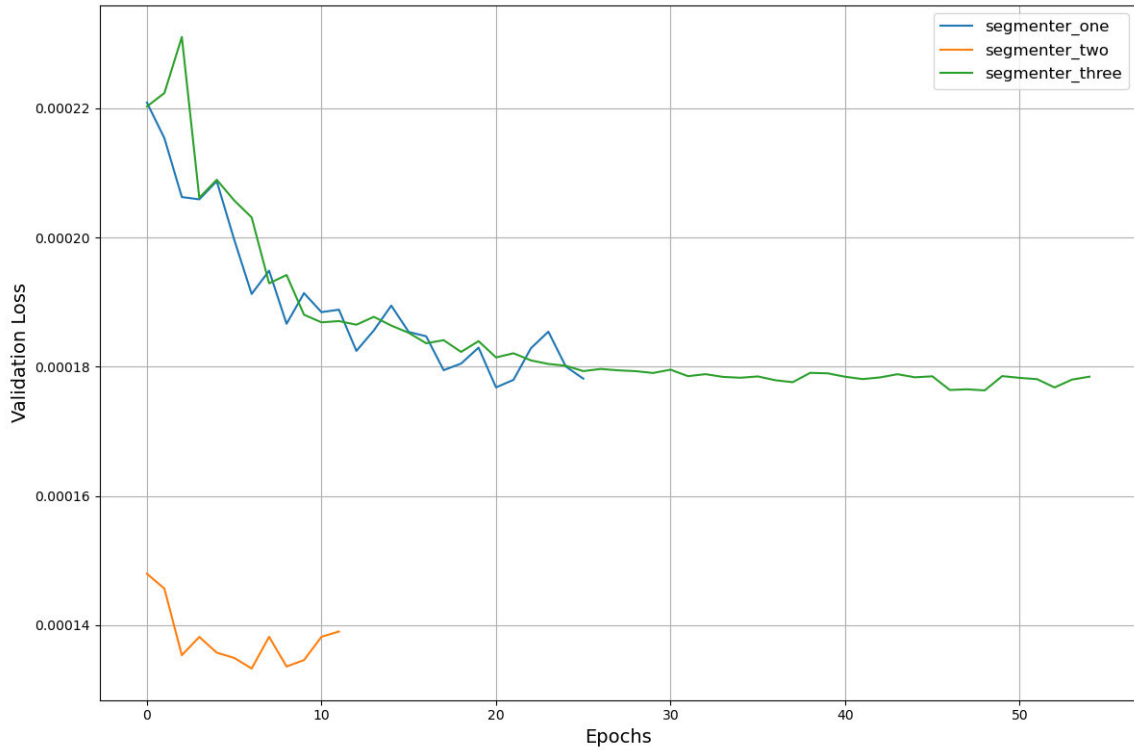


Figure 6.3: Transformer: validation loss over epochs for different segmenters.

performing best across both architectures. However, the fluctuations observed in the Transformer model’s validation loss for Segmenter Two raise concerns about its stability and generalisation capacity. These observations are further evaluated in subsequent sections to understand their implications for test performance and downstream tasks.

6.2 Intrinsic Evaluation

In Chapter 3, intrinsic evaluation was introduced as a method that directly assesses a system’s performance against predefined standards or criteria aligned with its objectives. For morphological segmentation, the primary objective is to accurately segment a given word into its morphemes. Since the system follows a supervised learning approach, the intrinsic evaluation process becomes straightforward. The output consists of morphemes or substrings separated by morpheme boundaries, which can be evaluated by directly comparing them to a pre-generated reference list (ground truth). This evaluation is aimed at investigating how well the machine learning models compare against the ZRG in terms of accuracy, since it is the ZRG that produced the ground truth. Moreover, it is an idealised version of the ZRG that is being compared to, since only inputs for which the ZRG produces a useful segmentation were used for training and testing.

Table 6.1 provides examples of inputs to the model, along with their expected (target)

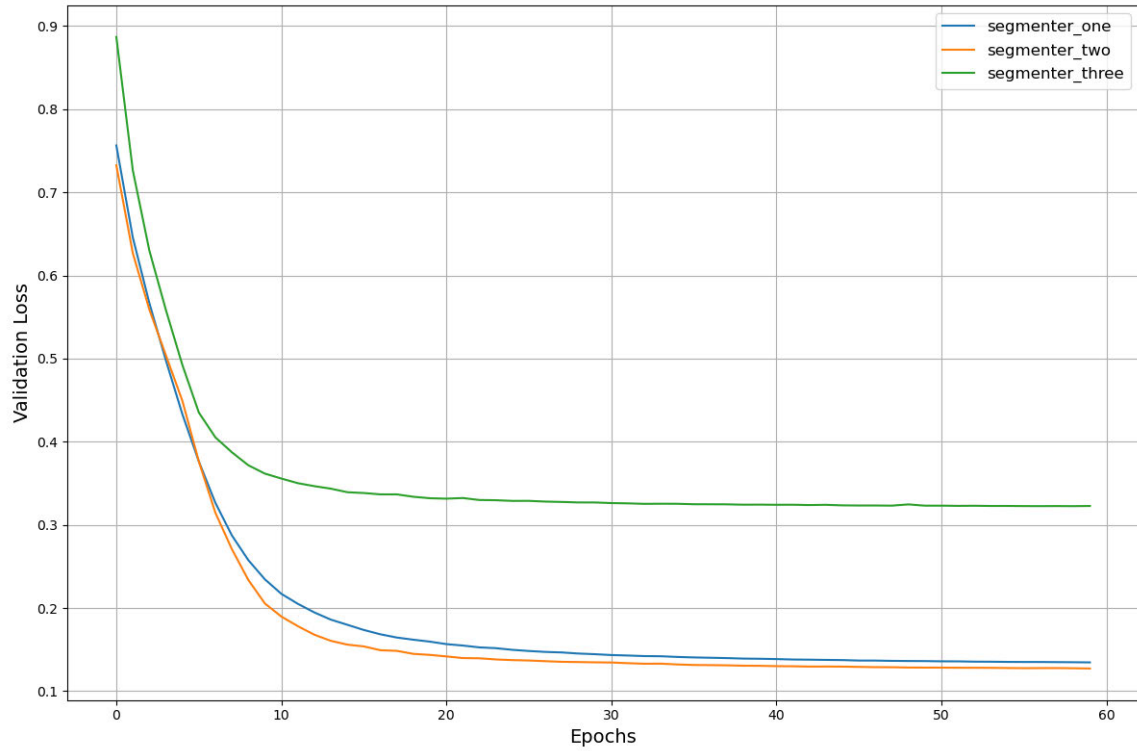


Figure 6.4: LSTM: validation loss over epochs for different segmenters.

segmentations and the segmentations predicted by the Segmenter Three model. The goal of intrinsic evaluation is to measure the similarity between predicted outputs and expected outputs.

Table 6.1: Morphological segmentation examples: input words with target and Segmenter One predicted segmentations.

Input	Target	Predicted
yezimpawu	ye-zimpawu	ye-zimpawu
engaguqulwe	enga-guqulw-e	enga-guqulw-e
ningethuswa	ni-ng-ethusw-a	ni-ngethusw-a
lingcweliseni	li-ngcwelis-e-ni	li-ngcwelis-eni
asikulalelanga	asi-ku-lalel-anga	asi-ku-lalel-anga

To achieve this, three evaluation metrics were employed: F1-Score, BLEU, and chrF, as previously outlined in Chapter 3 under the morphological segmentation metrics subsection. These metrics quantify the model’s ability to generate segmentations that closely match the ground truth. Each metric offers a unique perspective on evaluating segmentation accuracy, ensuring a comprehensive assessment of the models.

The BLEU metric, a precision-based evaluation metric introduced by Papineni *et al.* (2002), uses modified n -grams to assess the similarity between predicted and reference outputs at the morpheme level. In contrast, the chrF metric (Popović, 2015) evaluates

similarity at the character level rather than the token level, making it particularly useful for assessing how well the model generates correctly segmented tokens without introducing new characters or omitting existing ones, including morpheme boundaries. This distinction is especially relevant for morphological segmentation, where minor character-level differences can significantly impact accuracy.

Table 6.2 compares the BLEU scores across the three models (Transformer, LSTM, and CRF) and their respective segmenters. The Transformer model consistently outperforms the LSTM and CRF models across all segmenters, with Segmenter One achieving the highest BLEU score of 0.9561. For the LSTM model, Segmenter Two demonstrates the best performance with a BLEU score of 0.9397, while Segmenter One achieves a comparable score of 0.9372 and slightly surpasses Segmenter Two in the chrF score, achieving 0.9626 compared to 0.9624. The CRF model, although showing reasonable performance, lags behind, with Segmenter One recording its highest BLEU score of 0.8736. Notably, all these models achieve a score of more than 0.8 in these metrics, showcasing their ability to generalise well on unseen data, which is the test set. From these results, again, it can be justified that the Transformer architecture is performing exceptionally well, with none of the metrics yielding a score of lower than 0.9, showing its ability to predict morpheme boundaries accurately and produce correct tokens.

Table 6.2: Model performance comparison across different segmentation approaches (BLEU and chrF scores).

Model	Segmenter One		Segmenter Two		Segmenter Three	
	BLEU	chrF	BLEU	chrF	BLEU	chrF
Transformer	0.9561	0.9726	0.9372	0.9626	0.9326	0.9656
LSTM	0.8497	0.9081	0.9397	0.9624	0.9269	0.9620
CRF	0.8736	0.9356	0.8599	0.9294	0.8033	0.9108

Table 6.3 presents the precision, recall, and F1 scores for the Transformer, LSTM, and CRF models across Segmenters One, Two, and Three. Unlike BLEU, which accounts for n -gram ordering, these metrics evaluate segmentation correctness without considering token order. The Transformer model demonstrates the highest overall performance, with Segmenter One achieving the top F1 score of 0.9477, closely followed by Segmenter Two (0.9298). The LSTM model shows competitive performance, particularly in Segmenter Two with an F1 score of 0.9323, but underperforms in Segmenter One (0.842). The CRF model, while effective in some cases, performs relatively poorly in Segmenter Three, where it records its lowest F1 score of 0.7844. These results, alongside the n -gram-based metrics, highlight the robustness of the Transformer model, particularly Segmenter One, in delivering accurate segmentation. The LSTM model exhibits strong but less consistent results, while the CRF model reveals its limitations compared to the deep learning-based approaches.

Table 6.3: Comparison of precision, recall, and F1 scores across different architectures and segmentation styles.

Model	Metrics	Segmenter One	Segmenter Two	Segmenter Three
Transformer	precision	0.9486	0.9292	0.9295
	recall	0.9467	0.9305	0.9259
	F1	0.9477	0.9298	0.9277
LSTM	precision	0.8396	0.9320	0.9225
	recall	0.8444	0.9326	0.9229
	F1	0.8420	0.9323	0.9227
CRF	precision	0.8505	0.8370	0.7805
	recall	0.8555	0.8422	0.7883
	F1	0.8530	0.8396	0.7844

6.2.1 Model Selection for Downstream Evaluation

Based on the results presented in Table 6.3 and the prior analysis, the Transformer architecture emerges as the most robust architecture for morphological segmentation across all metrics, making it the ideal choice for downstream or extrinsic evaluation. Segmenter One of the Transformer models stands out with the highest F1 score of 0.952, which reflects its ability to accurately and consistently identify morpheme boundaries. Furthermore, the Transformer architecture consistently outperforms the LSTM and CRF architecture across precision, recall, and F1 metrics, highlighting its superior capacity for capturing complex morphological structures, particularly in agglutinative languages like isiZulu.

While the LSTM architecture shows competitive performance, especially in Segmenter Two with an F1 score of 0.9320, its lower performance in Segmenter One and less stable trends across segmenters suggest that it is less reliable for tasks requiring high segmentation accuracy. Similarly, the CRF architecture, although effective to some degree, falls behind both deep learning-based approaches, particularly in Segmenter Three, where its performance significantly drops to an F1 score of 0.7844. This indicates that it struggles to generalise effectively across different granularity levels.

Considering the Transformer architecture’s overall performance, stability, and its ability to achieve superior scores across all segmenters, it has been selected as the main architecture to be used for downstream evaluation. The subsequent extrinsic evaluation examines how incorporating morphological segmentation from the Transformer model impacts the performance of a machine translation system for isiZulu.

6.2.2 Data Augmentation Impact

From the results presented in Table 6.2 and 6.3, it can be seen that the Transformer model significantly outperforms other supervised models, namely the LSTM and CRF. This finding led to an investigation of the impact of data augmentation on segmentation performance using it as an investigation model. The unaugmented dataset generated by the ZRG contains approximately 76000 samples. It has been established in the literature that deep learning models have been found to perform better as the size of the data set increases. Given that deep learning models generally benefit from larger datasets, this experiment investigates how well data augmentation can help mitigate the challenge of data sparsity associated with morphologically complex languages. By increasing the number of training samples, data augmentation has the potential to enhance segmentation performance, particularly when compared to models trained solely on unaugmented data.

To investigate this impact, the Transformer architecture was also trained on the unaugmented dataset, using the same methodologies detailed in the previous sections, which include tokenisation, hyper-parameter optimisation, as well as training and evaluation. The results of this comparison are summarised in Table 6.4, showing key metrics such as precision, recall, F1-Score, BLEU, and chrF. These metrics provide insights into how data augmentation influences the effectiveness of segmentation models, particularly in addressing data sparsity and improving model generalisation. The hyper-parameters used to train the models on the unaugmented data are presented in Appendix A.1 as supplementary resources and can be accessed on the GitHub repository¹.

Table 6.4: Comparison of key metrics scores between Transformer models trained on unaugmented data and augmented data.

Model Name	Data Type	precision	recall	F1-Score	BLEU	chrF
Segmenter One	Unaugmented Data	0.8318	0.8485	0.8400	0.8426	0.9127
Segmenter Two	Unaugmented Data	0.9068	0.9173	0.9120	0.9160	0.9531
Segmenter Three	Unaugmented Data	0.8162	0.8187	0.8174	0.8206	0.9058
Segmenter One	Augmented Data	0.9486	0.9467	0.9477	0.9561	0.9726
Segmenter Two	Augmented Data	0.9292	0.9305	0.9298	0.9372	0.9626
Segmenter Three	Augmented Data	0.9295	0.9259	0.9277	0.9326	0.9656

From the results presented in Table 6.4, it is evident that data augmentation has a positive impact on the performance of the Transformer-based segmenters. Increasing the size of the dataset through augmentation addresses data sparsity issues and improves segmentation performance for a morphologically complex language like isiZulu.

Segmenters trained in augmented data consistently outperform those trained in unaugmented data across all key metrics, including precision, recall, F1-Score, BLEU, and chrF.

¹https://github.com/Sthesha/supervised_surface_segmentation

For instance, Segmenter One’s F1-Score improves from 0.8400 with unaugmented data to 0.9477 with augmented data, representing a substantial enhancement in performance, about a 0.10 increase. Similarly, BLEU and chrF scores also show marked improvements, demonstrating better alignment with ground truth segmentations and more robust generalisation.

This improvement aligns with the expectation that extensive larger and more diverse datasets enable deep learning models, such as Transformers, to learn and generalise more effectively, particularly in low-resource languages with agglutinative morphology. Data augmentation provides additional syntactic variations and examples, which help the model to generalise better to unseen data.

The findings demonstrate that data augmentation indeed helps to create segmentation models that generalise better and deliver superior performance compared to those trained on smaller, unaugmented datasets with limited variations. This highlights the salient role of data augmentation for addressing data sparsity and enhancing the effectiveness of models for morphologically complex, low-resource languages like isiZulu.

6.2.3 Investigating the Robustness of the Transformer Models

In the above section, the Transformer architecture was identified as the most effective among the three supervised methods, demonstrating superior performance. This combination forms a hybrid framework that allows further investigation of the formulated hypothesis, aligning with the hypothetico-deductive methodology outlined in Chapter 2.

As discussed in Chapter 4, the ZRG, a rule-based system, was identified as brittle, as it relies on a predefined lexicon and fails to segment words outside its predefined rules. In contrast, data-driven methods, particularly deep learning models, have demonstrated strong generalisation to OOV tokens, making them a potential remedy for this limitation of the ZRG tool.

The first hypothesis of the present study states: *The hybrid approach will exceed the rule-based approach in robustness.*

To investigate this hypothesis, the robustness of the Transformer model is assessed in comparison to the rule-based ZRG segmentation. Specifically, robustness is examined in terms of the model’s ability to handle unseen words and generalise effectively across different segmentation granularities. Thus, in order to test the first hypothesis, 10 words that ZRG failed to parse and segment were randomly selected. These words were then processed using the three Transformer models developed in this study that are of different granularity. The segmentation results are presented in Table 6.5 below.

Table 6.5: Segmentation results from different segmenters.

Original Word	Segmenter One	Segmenter Two	Segmenter Three
singesokuphumuza	si-nge-soku-phumuz-a	si-nge-so-ku-phumuz-a	singeso-ku-phumuz-a
ngasosebeni	nga-sosebeni	ng-a-soseb-e-ni	ng-as-o-sebeni
nasebumpungeni	na-s-e-bu-mpung-eni	na-s-ebu-mpung-eni	na-s-ebumpungeni
likamtshengiseni	lika-m-tshengiseni	lika-m-tshengiseni	likamtshengiseni
okunyeke	oku-nyek-e	oku-nyek-e	oku-nyek-e
skhathi	s-khathi	s-khathi	skhathi
epulasitiki	e-pulasitiki	e-pulasitiki	e-pulasitiki
ayebhaphithiza	a-ye-bhaphithiz-a	aye-bhaphithiza	aye-bhaphithiz-a
nomprofethikazi	n-o-m-profethikazi	no-m-profethikazi	no-mprofethikazi
osuziqedile	o-su-zi-qed-ile	osu-zi-qed-ile	osuzi-qed-ile

Since the original words shown in Table 6.5 are those which the ZRG could not linearise, in order to facilitate the comparison, manual segmentation was done in consultation with an expert linguist according to the different segmentation strategies of the ZRG. In terms of the labour-intensiveness and highly specialised nature of this task, a very small sample of 10 items was chosen for a preliminary investigation, which was deemed sufficient given its conclusive result. Table 6.6 below details these manually segmented outputs, aligning with the ZRG’s segmentation strategies, alongside a brief analysis of why ZRG initially failed to parse each token.

Table 6.6: Gold standard segmentations and failure analysis.

Original Word	Gold One	Gold Two	Gold Three	Cause of Failure
singesokuphumuza	si-nge-so-ku-phumuz-a	si-nge-so-ku-phumuz-a	si-nge-so-kuphumuz-a	construction missing
ngasosebeni	nga-s-o-sebe-ni	nga-s-o-sebe-ni	nga-s-o-sebe-ni	construction missing
nasebumpungeni	na-s-e-bu-mpung-eni	na-s-e-bu-mpung-eni	na-s-ebumpung-eni	construction missing
likamtshengiseni	lika-m-tshengiseni	lika-m-tshengiseni	lika-mtshengiseni	oov
okunyeke	oku-nyek-e	oku-nyek-e	oku-nyek-e	input error
skhathi	s-khathi	s-khathi	s-khathi	input error
epulasitiki	e-pulasitiki	e-pulasitiki	e-pulasitiki	oov
ayebhaphithiza	a-ye-bhaphithiz-a	aye-bhaphithiz-a	aye-bhaphithiz-a	oov
nomprofethikazi	n-o-m-profethikazi	no-m-profethikazi	no-mprofethikazi	oov
osuziqedile	o-su-zi-qed-ile	osu-zi-qed-ile	osuzi-qed-ile	construction missing

As depicted in Table 6.6, ZRG primarily failed due to missing constructions and encountering OOV tokens. To enable a fair comparison with the Transformer models, an expert linguist provided manual segmentations based on two hypothetical questions. First, for cases of missing constructions or OOV tokens, the linguist answered: *What segmentation would the ZRG have produced if it had perfect lexical and grammatical coverage?* Secondly, when the input was identified as erroneous or grammatically incorrect, the linguist addressed the question: *Given ungrammatical input, what would an ideal system produce according to each segmentation strategy?* This approach allowed for a systematic and fair evaluation of the robustness of the Transformer models compared to the idealised ZRG output.

6.2.3.1 Quantitative Analysis of Robustness

To objectively evaluate the models’ robustness, intrinsic metrics such as BLEU (unigram and bigram), precision, recall, F1 score, and chrF scores were computed, and presented in Table 6.7. From the quantitative analysis perspective, Segmenter One, representing the finest segmentation granularity, clearly outperforms the other segmenters across all evaluation metrics. It achieved the highest scores with metrics averaging around 0.93, clearly indicating superior robustness in handling previously unseen words or tokens outside the lexicon coverage of ZRG. In contrast, Segmenter Two showed moderate robustness, with scores averaging around 0.78, whereas Segmenter Three performed the poorest, averaging approximately 0.60. These findings indicate a trend in declining segmentation performance as granularity decreases, suggesting that finer-grained segmentation provides greater resilience in challenging segmentation scenarios.

Table 6.7: Evaluation metrics for different segmenters.

Segmenter	BLEU	precision	recall	F1 Score	chrF Score
Segmenter One	0.8932	0.9730	0.9000	0.9351	0.9357
Segmenter Two	0.7349	0.7714	0.7297	0.7500	0.8961
Segmenter Three	0.4840	0.6154	0.5161	0.5614	0.8228

These results strongly support the first hypothesis, which states that: *The hybrid approach will exceed the rule-based approach in robustness.* Segmenter One, leveraging both rule-based synthetic data and supervised machine learning, shows significantly better robustness compared to the purely rule-based ZRG approach. Although Segmenters Two and Three performed less effectively, they still demonstrated clear advantages over the original ZRG tool, which failed entirely to produce segmentation for tokens not covered by its predefined rules. The hybrid approach thus proves to be more robust by successfully providing reasonably accurate segmentations on tokens for which the ZRG failed to produce anything.

6.2.4 Investigating Efficiency of the Transformer Model

One of the key limitations of the ZRG, as previously discussed, is its efficiency to generate segmentations through parsing. Since ZRG enables for the morphosyntactic analysis of isiZulu text by generating syntax trees, this process requires deep linguistic analysis, making parsing significantly slower. Consequently, the segmentation process using ZRG is also slower. The hybrid framework proposed in this study aims to address this limitation by leveraging ZRG to generate training data while utilising a faster machine learning-based approach for segmentation.

To investigate this claim, a second hypothesis was formulated: *The hybrid approach will exceed the rule-based approach in efficiency.*

To evaluate whether this hypothesis can be accepted or rejected, a simple computational efficiency test was conducted, measuring the time taken to segment a token using both systems. Specifically, the segmentation speed of the ZRG and Transformer-based models was compared.

For this experiment, 100 random tokens were selected from the evaluation set consisting of 20 000 isiZulu tokens. For the ZRG, a script was developed to facilitate parsing, linearisation, and preprocessing, allowing for a seamless comparison. The time it took to perform segmentation for all tokens was recorded.

A similar approach was followed for the Transformer models, where the same 100 randomly selected tokens were segmented using the three Transformer models corresponding with different segmentation granularities. To ensure a fair comparison, both systems were executed on identical hardware infrastructure, a Docker-based environment running on multiple server cores with 80 cores and 100 GB of memory. Importantly, all models were strictly run on CPUs rather than GPUs to maintain consistency in performance evaluation.

Table 6.8 presents a side-by-side comparison of execution times for the different segmentation models.

Table 6.8: Execution times for segmenters using ZRG and Transformer models.

Model	Segmenter One (s)	Segmenter Two (s)	Segmenter Three (s)
Rule-Based	1183.55	1179.78	1188.91
Transformer	31.72	34.66	65.02

The results in Table 6.8 clearly illustrate that the Transformer-based models significantly outperform the ZRG-based rule-based approach in terms of segmentation speed. While the ZRG system takes approximately 1 180 seconds to process 100 tokens across different segmenters, the Transformer models complete the same task in under 70 seconds, marking a *30x* to *40x* speed improvement. This substantial reduction in execution time demonstrates the efficiency gains achieved through the hybrid approach.

A key reason behind this efficiency difference lies in the architectural design of the two systems. The ZRG relies on an exhaustive parsing mechanism, where each token must be analysed according to a set of predefined grammatical rules, requiring deep syntactic processing and multiple tree evaluations. Given isiZulu’s highly agglutinative nature, a single word may contain numerous morphemes, each with several possible interpretations, further increasing the computational burden.

In contrast, the Transformer model operates as a data-driven system that processes tokens in parallel without requiring explicit rule-based parsing. Instead of generating syntax trees

for each token, the Transformer directly learns segmentation patterns from large datasets, allowing it to make faster predictions once trained. Additionally, the model leverages self-attention mechanisms, which enable efficient context modelling without the sequential dependencies seen in traditional rule-based parsing.

Interestingly, the variations in execution time among the three Transformer segmenters indicate that segmentation granularity affects computational cost, though not in a straightforward manner. While it might be expected that models processing fewer tokens per word (i.e., those with coarser segmentation) would require less computation, the results suggest otherwise. A key contributing factor to this trend is vocabulary size, models trained on coarser segmentations (e.g., Segmenter Three) tend to have larger vocabularies, leading to an increased number of model parameters.

In particular, Segmenter Three, which introduces fewer segmentation points, results in a larger vocabulary size because words remain whole rather than being decomposed into smaller morphemes. This inflates the embedding matrix and increases computational demands, as the model must learn representations for a greater number of unique tokens. In contrast, Segmenter One, which aggressively segments words into smaller morphemes, reduces vocabulary size, thereby lowering computational overhead associated with word representations and improving efficiency in training and inference. Nonetheless, despite these variations, the Transformer model remains significantly faster than the ZRG-based system, reinforcing the efficiency benefits of the hybrid approach.

6.2.4.1 Practical Implications

The dramatic improvement in segmentation speed has significant practical implications. The rule-based ZRG system is unsuitable for real-time or large-scale applications due to its high computational cost. However, by leveraging the hybrid approach, where ZRG is used for dataset generation, but segmentation is performed using a trained Transformer model, one can achieve the best of both worlds: linguistic accuracy from rule-based segmentation and efficiency from machine learning-based segmentation.

These findings confirm the second hypothesis: *The hybrid approach will exceed the rule-based approach in efficiency.*

By demonstrating superior computational efficiency while maintaining high segmentation accuracy, the Transformer model proves to be a viable and scalable solution for isiZulu morphological segmentation.

6.3 Extrinsic Evaluation

The second stage of evaluation conducted is extrinsic evaluation, which indirectly measures the impact of the created system on a downstream task, aligning with *Sub-objective 4* of this study. Morphological segmentation is typically employed to enhance the performance of downstream NLP tasks. Thus, the evaluation of how well the segmentation system improves the effectiveness of a downstream task, in this case, machine translation, is referred to as extrinsic evaluation.

Based on the intrinsic evaluation results, the Transformer model demonstrated superior performance compared to other architectures, with all its segmenters achieving a score of more than 0.9 across all metrics when evaluated on the testing set. Among the three segmenters, Segmenter One emerged as the best-performing model, achieving an F1-score of 0.951 and a BLEU score of 0.9605. These metrics reflect the model’s ability to learn underlying patterns effectively with the defined granularity.

However, strong performance in intrinsic evaluation metrics does not always lead to improved outcomes in downstream tasks (Jones & Galliers, 1995), highlighting the need for further investigation through extrinsic evaluation. Furthermore, the level of segmentation granularity can influence the performance of the downstream language model (Salesky *et al.*, 2020). This finding highlights the importance of evaluating different segmenters to determine which granularity level yields the most effective translation system.

6.3.1 Neural Machine Translation Model

As explained in Chapter 3, Neural Machine Translation (NMT) systems have emerged as a dominant paradigm, significantly advancing the field of machine translation. However, one persistent challenge is their inability to effectively translate new or rare words, a limitation exacerbated in low-resource languages with complex morphological structures like isiZulu. For such languages, data sparsity is a prevalent issue due to the combinatorial explosion of word forms.

To mitigate this, researchers have explored training NMT systems using subword representations instead of whole tokens. Common approaches include Byte Pair Encoding (BPE), syllabification, or morphological segmentation. However, there is no universal consensus on the extent to which these methods address data sparsity and improve NMT performance for morphologically rich languages.

To investigate this further, another hypothesis was formulated: *The hybrid approach will result in one or more segmenters that improve performance in a downstream task such as machine translation.*

To investigate this hypothesis, this study primarily focuses on the translation from isiZulu into English, where isiZulu is tokenised using various hybrid segmentation methods. The main rationale behind this decision is supported by several studies that explicitly discourage investigating morphological segmentation in the reverse direction, that is, translating from a language with isolating morphology into one with complex morphology. For instance, Badr *et al.* (2008:105) argue that “translating to a more morphologically complex language is a more difficult task, where there is a higher chance of translating word inflections incorrectly”. This often necessitates additional desegmentation steps during decoding, which reverse segmentation to produce readable outputs with fewer grammatical errors (El Kholy & Habash, 2012; Huck *et al.*, 2017; Salameh *et al.*, 2015).

These studies implicitly support the strategy of segmenting only the source language if the target language is morphologically simpler, because it improves alignment quality and generalisation without requiring the target language model to recombine morphemes, which is a process that is error-prone and often demands additional evaluation of generated tokens to assess segmentation effectiveness. For these reasons, the current study primarily examines isiZulu-to-English translation. However, it should be noted that experiments involving English-to-isiZulu translations were conducted as supplementary analyses; the results of these experiments will be presented, but details regarding their training and evaluation processes will not be elaborated upon or used directly to investigate the primary hypothesis.

Thus, this section outlines the steps undertaken to explore and validate this primary hypothesis systematically using the isiZulu-to-English downstream task.

6.3.1.1 Bilingual Data Preparation

In building NMT systems, one of the most critical requirements is access to large, high-quality parallel datasets. However, such datasets are often scarce, particularly for low-resource languages. For many of these languages, existing parallel data is primarily derived from web scraping, which often yields insufficient quality and quantity.

For isiZulu, as of this writing and to the best of the author’s knowledge, there is no publicly available high-quality parallel dataset for isiZulu-to-English translation. The most commonly used resource is the Autshumato² project dataset (Groenewald & Fourie, 2009), which contains approximately 230 000 lines. However, this dataset has notable limitations. A manual inspection by the author, a native isiZulu speaker, revealed a notable number of inaccurate translations. For instance, randomly selected lines 9 061, 77 442 and 163 285 contained mismatched translations. These inconsistencies can hinder the training and evaluation of optimal NMT models, which poses a challenge in this study,

²<https://repo.sadilar.org/handle/20.500.12185/575>

as NMT models play a crucial role in the investigation of segmentation strategies and to draw meaningful conclusions. Additionally, much of the dataset originates from domain-specific texts, such as policies and governmental documents, which restricts its usefulness for developing systems that are capable of generalising across diverse text types.

To create a more diverse dataset, this study leverages the No Language Left Behind (NLLB) corpus (Costa-Jussà *et al.*, 2022). The NLLB corpus provides a broader range of domains, enhancing the development of more versatile NMT systems. This dataset integrates several sources, including Autshumato, with content collected from various open sources.

The sentences in the NLLB corpus are ranked using the LASER score, derived from the Language-Agnostic Sentence Representations model (Heffernan *et al.*, 2022). LASER scores are employed in sentence alignment tasks to measure semantic similarities between sentences in different languages, enabling better alignment assurance. However, despite these advancements, challenges persist due to the automatic alignment process. One notable issue is language mismatch occurs in the parallel data, where source and target sentences occasionally belong to languages other than those specified, potentially impacting the dataset’s quality.

To address the issue of language mismatch in the parallel data, this study employed a Language Identification (LID) system, specifically the fastText LID model (Grave *et al.*, 2018). This system was used to verify whether the source sentence was isiZulu and the target sentence was English. Sentences failing this criterion were discarded. The process continued iteratively until 100 000 sentences were collected.

The data selection process followed these steps:

1. *Initial Filtering*: Selecting sentences with the highest LASER scores to ensure semantic similarity between source and target sentences.
2. *Language Verification*: Using the fastText LID model to confirm that the source sentence was isiZulu and the target sentence was English.
3. *Duplication Removal*: Removing any duplicate sentences to ensure unique entries.

Through this process, approximately 118 000 sentences were initially selected, and after final filtering and deduplication, a dataset of 100 000 sentences was compiled. To investigate the impact of different segmentation on NMT systems, the next step involved segmenting the isiZulu sentences using the developed segmentation models of different granularities, namely Segmenters One, Two, and Three. The final dataset comprised English sentences, unsegmented isiZulu sentences, and three sets of segmented isiZulu sentences.

6.3.1.2 Models Development and Training

For the NMT systems, this study employed the Transformer architecture, known for its revolutionary impact on NLP tasks and its role in achieving state-of-the-art results (Rahali & Akhloufi, 2023). The architecture used here aligns with the one described in Chapter 5, with the primary difference being the tokenisation strategy applied for this task.

A custom tokenisation method was implemented for the segmenter-based NMT models, where isiZulu sentences were tokenised at morpheme boundaries and spaces, keeping them as tokens in the final list, according to the segmentation granularity of the segmenters. For the unsegmented isiZulu-to-English model, tokenisation occurred at the word level. The tokenisation process mirrored that of segmenter development, including assigning unique IDs to tokens and incorporating special tokens such as *[PAD]*, *[SOS]*, and *[EOS]* to facilitate sequence processing.

Beyond tokenisation, other processes, such as embedding, model initialisation, and training configurations, were kept consistent across all models. Additionally, since the optimal parameters for this task were not predetermined, an HPO strategy was employed for all models using the Tune toolkit. Table 6.9 presents the optimal hyper-parameters identified for these models.

Table 6.9: Optimal hyper-parameters for translation models across different segmentation approaches.

Parameter	Seg One	Seg Two	Seg Three	No Seg
Learning Rate	4.551×10^{-4}	1.423×10^{-4}	6.525×10^{-5}	5.176×10^{-5}
Sequence Length	310	300	250	100
d_{model}	512	512	512	1024
Num Layers	3	3	4	3
Num Heads	4	4	4	8
d_{ff}	2048	1024	2048	1024
Dropout	0.1951	0.1203	0.1381	0.2113
Max Grad Norm	1.1460	0.6386	1.8166	1.8326
Label Smoothing	0.0433	0.0422	0.0357	0.0482

6.3.2 Neural Machine Learning Evaluation

After training the NMT models, their performance was evaluated using a test dataset consisting of 10 000 sentences, derived from the original 100 000 parallel sentences collected. Before discussing the evaluation metrics and the results, it is important to examine the training and validation processes through the corresponding loss graphs.

6.3.2.1 Translation Models Performance Analysis: Training and Validation

Figures 6.5 and 6.6 depict the training and validation loss trends across different epochs for the segmented NMT systems, offering insights into their learning behaviours and convergence patterns.

In contrast, Figure 6.7 presents the validation loss trends for the NMT system trained on unsegmented isiZulu data. This system exhibited significantly higher validation losses compared to the segmented systems, necessitating a separate plot to avoid stretching the scale and overshadowing the detailed patterns and fluctuations of the segmented models. Additionally, Figure 6.5 includes the training loss trends for the unsegmented NMT model, providing a comprehensive view of its learning progression alongside the segmented models. These visualisations collectively help identify the epoch corresponding to the best-performing models, which were saved for further evaluation.

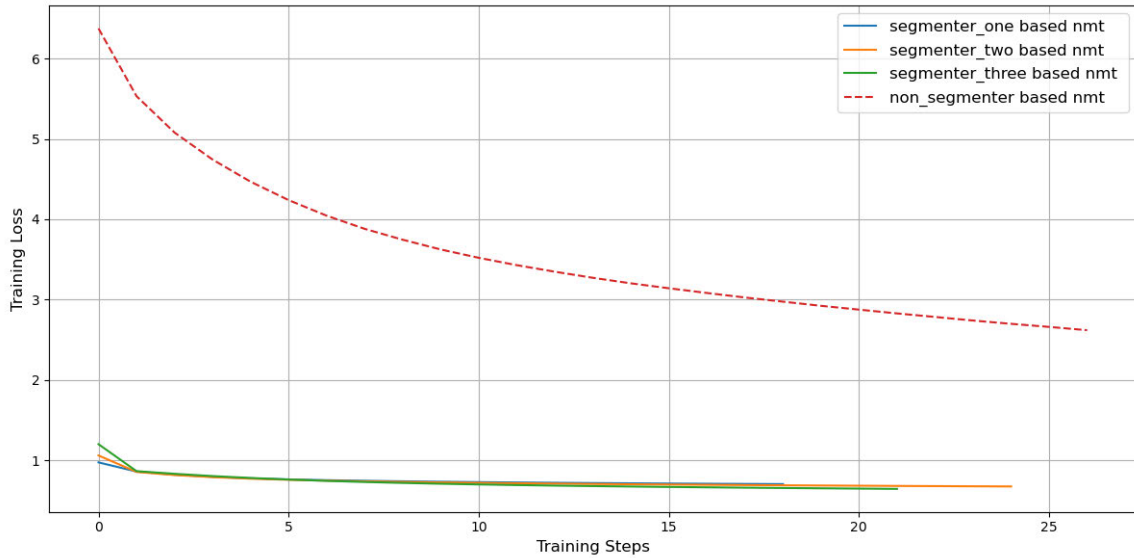


Figure 6.5: Training loss across epochs: Translation models with varying segmentation strategies.

During training, the maximum epoch limit was set to 100 for all four NMT models, with an early stopping technique applied using patience of five epochs, similar to the segmentation models. From the loss graphs, it is clear that the unsegmented model trained for a longer duration, terminating at 27 epochs. However, as depicted in Figure 6.7, its validation loss began to stabilise around epoch 20, after which minor fluctuations with small increases and decreases were observed, ultimately leading to termination. Both the training and validation losses for the unsegmented model were significantly higher than those observed in the segmented models. This outcome aligns with what is commonly reported in literature (Shapiro & Duh, 2018), considering the data sparsity issues associated with token-to-token translation for languages with complex morphologies. In such cases, only a

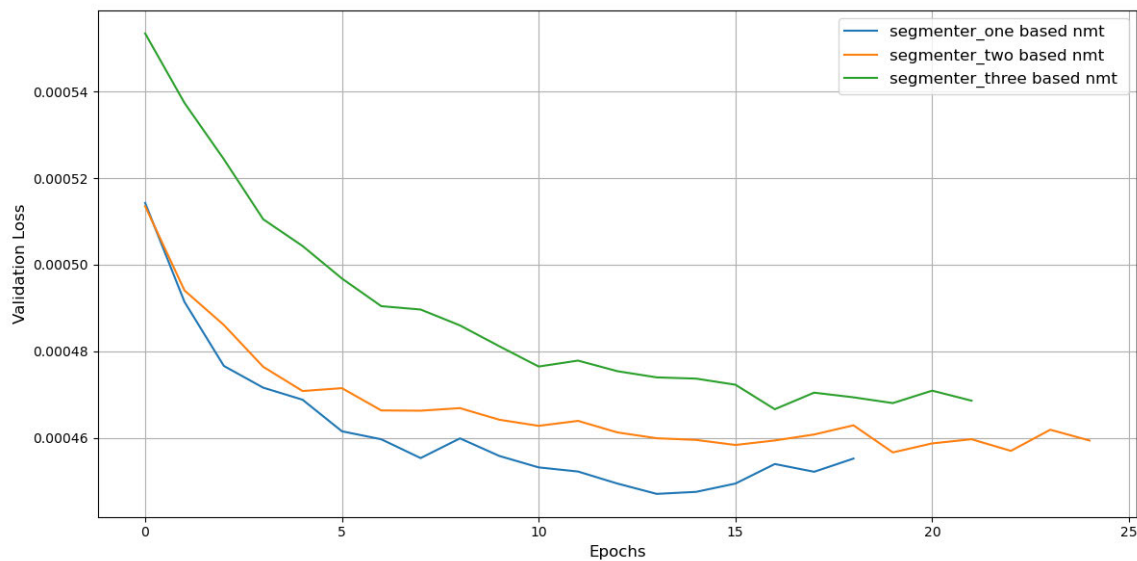


Figure 6.6: Validation loss across epochs: Translation models with varying segmentation strategies.

small portion of words make it into the vocabulary, making it challenging for deep learning models to learn effectively in such environments.

Interestingly, the validation loss trends for the NMT systems using segmenters, as illustrated in Figure 6.6, show that all segmented models achieved significantly lower validation losses compared to the unsegmented model. The validation loss for these systems consistently decreased and stabilised around epoch 10. Among these, the model based on Segmenter One (finer granularity) had the lowest validation loss, followed by Segmenter Two (medium granularity) and Segmenter Three (coarser granularity). Despite Segmenter One’s finer granularity (suggesting easier source-to-target mapping), the models’ best epochs did not strictly align with the validation loss trends. Specifically, the best models were saved at epochs 13, 16, and 19 for Segmenter One, Segmenter Two, and Segmenter Three, respectively.

This discrepancy indicates that validation loss alone is insufficient to determine the best-performing model. As observed in the segmentation analysis, validation loss does not always correlate directly with overall model performance. Consequently, proper evaluation metrics, such as BLEU and chrF scores, are essential for accurately assessing model effectiveness and drawing meaningful conclusions with a view to reporting on the validity of the research hypothesis.

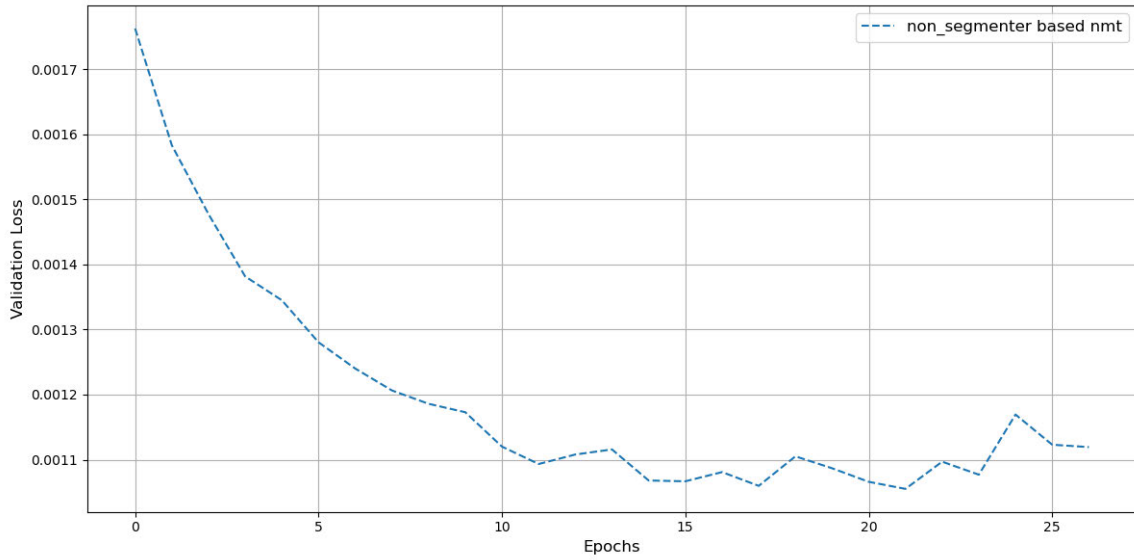


Figure 6.7: Validation loss across epochs: for the Unsegmented model.

6.3.2.2 Intrinsic Evaluation of Translation Models

In this section, hypothesis number two is investigated: *The hybrid approach will result in one or more segmenters that improve performance in a downstream task such as machine translation.* This aligns directly with *Sub-objective 5*, which involves evaluating segmentation models intrinsically and extrinsically using suitable metrics.

Specifically, this hypothesis explores the extent to which morphological segmentation can mitigate data sparsity challenges associated with morphologically complex languages, subsequently enhancing the generalisation capabilities of language models. To examine this hypothesis, the effectiveness of morphological segmentation was evaluated extrinsically using NMT systems as the downstream task. BLEU and chrF scores, two widely recognised metrics for assessing machine translation quality, were used to measure performance improvements.

Table 6.10 presents the results obtained from this evaluation, offering quantitative insights into how different segmentation strategies impact data sparsity reduction and translation quality improvement. These findings directly address the hypothesis formulated, providing evidence to determine the practical benefits of adopting a hybrid morphological segmentation using different models of the different granularity approaches for the downstream task.

The BLEU and chrF scores in Table 6.10 reveal that models trained with segmented isiZulu data outperform the model without segmentation, indicating that morphological segmentation somehow does have a positive impact on the translation quality.

Table 6.10: IsiZulu–English translation metrics for Segmentation models (test set).

Model	BLEU	chrF
Segmenter One	0.227	0.460
Segmenter Two	0.235	0.461
Segmenter Three	0.224	0.459
No Segmentation	0.188	0.420

Among the segmented approaches, the model using Segmenter Two achieved the highest BLEU (0.235) and chrF (0.461) scores, suggesting that the medium granularity segmentation strikes an optimal balance between preserving meaningful linguistic information and reducing the vocabulary size for efficient learning.

While Segmenter One and Segmenter Three also performed well, their slightly lower scores (BLEU: 0.227 and 0.224, chrF: 0.460 and 0.459) suggest that overly fine or coarse segmentation may not fully optimise the trade-off between generalisation and specificity in the translation task. As noted in Chapter 4, Segmenter Two segments according to “segmentation naturally built into the ZRG internal structure”. It is noted that the internal segmentation model of the ZRG seems to correspond to the optimal level for downstream tasks. The unsegmented model demonstrated the weakest performance (BLEU: 0.188, chrF: 0.420), highlighting the challenges of token-to-token translation in languages with complex morphologies due to data sparsity issues.

As mentioned above, similar training steps and evaluation steps were undertaken for English to isiZulu, where four models were trained, with the difference being the direction. The results are presented in Table 6.11, and the configurations used to train these models can be observed in Appendix A.2.

Table 6.11: Translation quality metrics for different Segmentation models in English–isiZulu translation.

Model	BLEU	chrF
Segmenter One	0.1179	0.504
Segmenter Two	0.0872	<i>0.494</i>
Segmenter Three	0.1025	0.497
No Segmentation	0.0901	0.405

In the English-to-isiZulu translation direction, the models achieved significantly lower scores, approximately half of the BLEU scores to those obtained by the NMT systems trained in the opposite direction (isiZulu-to-English). These results align with and further support the arguments presented in the previously discussed studies, emphasising that translating into a morphologically richer language poses additional complexities and typically yields lower translation quality.

Interestingly, the chrF scores for English-to-isiZulu translations remain comparatively close to those of isiZulu-to-English. This similarity in character-level performance suggests that while character matches occur frequently, token-level discrepancies are notably higher, leading to reduced BLEU scores. This further highlights the inherent difficulties in modelling translations at the token-to-morpheme level for morphologically complex target languages.

6.3.2.2.1 Evaluation on External Dataset

The dataset used for evaluation in this study was initially derived from the NLLB corpus. To ensure unbiased and standardised evaluation and to validate the consistency of findings, an external dataset, known as FLORES (Goyal *et al.*, 2022; NLLB Team *et al.*, 2024), was provided for further assessing the performance of models trained on the NLLB dataset. The FLORES dataset, designed for Few-Shot Learning for Zero-Shot Evaluation in Translation, consists of human-translated and verified sentences, establishing a gold standard for evaluating translation quality.

However, for this study, an updated version of FLORES (Abdulmumin *et al.*, 2024) was used for evaluation. This version addressed corrections in isiZulu and three other African languages, ensuring greater accuracy and reliability. The updated dataset was processed in the same way as the NLLB dataset, maintaining consistency in preparation and allowing for meaningful comparisons. The results presented in Table 6.12 highlight the translation performance of different segmentation strategies on the updated FLORES dataset for isiZulu to English.

Table 6.12: Translation scores for Segmentation models on updated FLORES (isiZulu–English).

Model	BLEU	chrF
Segmenter One	0.1056	0.4021
Segmenter Two	0.1019	0.3973
Segmenter Three	0.0891	0.3767
No Segmentation	0.0764	0.3542

The results in Table 6.12 demonstrate that models trained with segmented data consistently outperform the unsegmented model, highlighting the positive impact of morphological segmentation in translating from a morphologically complex language (isiZulu) to a relatively simpler language (English). Among the segmentation strategies, Segmenter One achieves the highest BLEU (0.1056) and chrF (0.4021) scores, indicating that fine-grained segmentation provides the most effective balance for this dataset. However, this trend differs from the findings presented in Table 6.10, where the Segmenter Two-based model yielded the highest scores. This discrepancy may be attributed to the dataset used in the

testing set for Table 6.10, which likely included data similar to or overlapping with the training data, potentially containing shared vocabulary and sentence structures. In contrast, the updated FLORES dataset introduces entirely new data, likely with a significant number of previously unseen words. This provides a distinct advantage to the model using finer-grained segmentation, because its aggressive approach of breaking words into smaller morphological units better addresses data sparsity and allows for improved generalisation.

Segmenter Two demonstrates competitive performance, with BLEU and chrF scores slightly lower than Segmenter One. This suggests that while medium granularity retains meaningful linguistic information, it may not optimise vocabulary size as effectively, particularly when evaluated on a completely new dataset. On the other hand, Segmenter Three shows a notable decline in performance, indicating that overly broad grouping of morphological units reduces the model’s ability to generalise and align effectively. The unsegmented model performs the worst (BLEU: 0.0764, chrF: 0.3542), highlighting the challenges of token-to-token translation for morphologically rich languages, where data sparsity increases vocabulary size and limits the model’s generalisation capabilities.

These results, presented in Tables 6.10 and 6.12, strongly support the acceptance of the second hypothesis, which stated that:

The hybrid approach will result in one or more segmenters that improve performance in a downstream task such as machine translation.

The findings not only confirm that at least one segmented model surpasses the unsegmented model, but clearly demonstrate that all three segmented models consistently outperform the baseline unsegmented model in isiZulu-to-English translation tasks. This finding provides compelling evidence that morphological segmentation effectively reduces data sparsity, enhances translation quality, and improves the generalisation capability of NMT models.

However, while segmentation improves performance, the overall BLEU and chrF scores are lower than those observed with the original testing set. This suggests the need for a more representative and larger parallel dataset to evaluate the models’ potential fully. Nevertheless, this study was intentionally performed in a low-resource setting, where large datasets are unavailable and computational resources are limited, emphasising the practical challenges and potential solutions for such constrained environments.

When the English-to-IsiZulu translation model was evaluated on the FLORES dataset, the results, presented in Table 6.13, show significantly lower scores compared to the isiZulu-to-English direction. This finding aligns with expectations, given the challenges associated with modelling a morphologically richer target language, as well as the complexities introduced by de-segmentation.

Table 6.13: Translation quality metrics scores for different segmentation styles using the updated FLORES dataset (English-isiZulu).

Model	BLEU	chrF
Segmenter One	0.0285	0.2880
Segmenter Two	0.0140	0.4189
Segmenter Three	0.0237	0.4262
No Segmentation	0.0177	0.3629

These results further emphasise that translation into a morphologically rich language (isiZulu) presents greater challenges compared to translation from isiZulu into English. The lower BLEU scores suggest that segmentation alone may not fully address the complexities of English-to-isiZulu translation, and further refinements in model architectures and data preprocessing strategies may be necessary.

6.4 Chapter Summary

This chapter provided a comprehensive evaluation of the morphological segmentation models developed for isiZulu, addressing several research objectives outlined in Chapter 1, specifically *Sub-objective 5*, which aimed to assess the performance of the developed models both intrinsically and extrinsically. Additionally, this evaluation allowed for the investigation and testing of two formulated hypotheses. The first hypothesis examined whether the hybrid segmentation approach (combining rule-based and supervised machine learning methods) would address the weaknesses of the standalone ZRG system, namely its brittleness and slow parsing speed. The second hypothesis investigated whether morphological segmentation would improve the performance of downstream NLP tasks, specifically NMT.

The intrinsic evaluation highlighted clear performance differences among the Transformer, LSTM, and CRF-based architectures, validating the selection of the Transformer as the most effective architecture for the supervised segmentation task. The Transformer demonstrated superior performance across all segmentation granularity levels, consistently achieving higher BLEU, precision, recall, F1, and chrF scores. Given these results, the Transformer architecture was further evaluated for robustness and efficiency. Its robustness, defined by its ability to segment previously unseen or lexically uncovered tokens, was demonstrated to be significantly superior to the rule-based ZRG approach. Additionally, the efficiency of the Transformer was considerably better, validating the first hypothesis that the hybrid system effectively mitigates the shortcomings inherent in purely rule-based systems.

The extrinsic evaluation evaluated the impact of morphological segmentation on NMT performance. Models trained with segmented isiZulu data significantly outperformed the

unsegmented baseline across all evaluation metrics. Segmenter Two achieved the highest BLEU and chrF scores in the original NLLB-based dataset, indicating an optimal balance between vocabulary reduction and linguistic granularity. However, Segmenter One slightly outperformed Segmenter Two on the updated FLORES dataset, which contained more unseen words, indicating its greater resilience to data sparsity.

The evaluations underscored the role of morphological segmentation in addressing data sparsity, particularly for morphologically complex languages such as isiZulu. Furthermore, the results demonstrated the impact of dataset diversity on model generalisation and robustness, emphasising the importance of comprehensive and varied datasets for reliable evaluation.

The next chapter will synthesise these findings, offering final remarks that reflect on the objectives and hypothesis set at the outset of this study. It will critically evaluate whether the study successfully met its goals and validated the hypothesis regarding the impact of segmentation on machine translation performance. Additionally, the chapter will discuss the limitations encountered during the research and provide recommendations for future work, focusing on enhancing segmentation techniques and addressing gaps in low-resource language processing.

Chapter 7

Findings, Conclusion, and Future Research

7.1 Introduction

This study set out to investigate the effectiveness of a hybrid approach towards morphological surface segmentation. Specifically, it explored the use of a rule-based system, the ZRG system, to generate synthetic morphological surface segmented data of varying granularity for isiZulu. This data was then used to train supervised machine learning models for morphological segmentation, and their performance was evaluated both intrinsically and extrinsically.

This chapter is structured as follows: Section 7.2 provides a summary of the salient issues presented in the previous chapters. Section 7.3 revisits the key objectives set in Chapter 1, assessing how they were met, and highlighting key findings. Section 7.4 reflects on how these findings facilitated the investigation of the formulated hypotheses, while Section 7.5 discusses how the study results answer research questions and relate back to the primary objective.

In addition, the challenges and limitations encountered during the study are outlined in Section 7.6, together with recommendations for future research. Finally, Section 7.7 presents the study's overall conclusions.

7.2 Summary of the Chapters

In Chapter 1, the study introduced isiZulu as a morphologically complex language characterised by an agglutinative morphology and conjunctive orthography. Words in isiZulu often consist of multiple morphemes fused into a single token, leading to combinatorial word expansion and data sparsity, which significantly affect language modelling and NLP applications.

To address these challenges, the study motivated for using morphological segmentation as a viable solution, as it breaks words into their constituent morphemes, thereby reducing vocabulary size and improving generalisation in language models. The study distinguished between two primary segmentation approaches: surface segmentation and canonical segmentation. Both segmentation strategies have traditionally been implemented using rule-based approaches, which rely on expert linguistic knowledge, and machine learning approaches, which learn segmentation patterns from data, either in a supervised or unsupervised manner or semi-supervised.

To overcome the limitations of purely rule-based or data-driven methods, this study proposed a hybrid approach to morphological segmentation that leverages the Zulu Resource Grammar (ZRG), a rule-based system in order to generate surface-segmented synthetic data. This data was then used to train supervised machine learning models for segmentation, effectively combining the strengths of both approaches.

Chapter 2 outlined the research methodology adopted to investigate the effectiveness of a hybrid rule-based (ZRG) and supervised machine learning approach for isiZulu morphological segmentation. Grounded in the positivist paradigm, the study follows a hypothetico-deductive methodology, ensuring a structured, hypothesis-driven inquiry. The chapter introduced various research philosophies and paradigms, detailing how ontological, epistemological, axiological, and methodological assumptions inform the study's approach. The positivist paradigm was identified as the guiding framework, emphasising empirical evidence, objectivity, and systematic observation. By aligning the study's research questions, methodology, and objectives, this chapter established a structured framework for investigating the development of a hybrid segmentation system and its impact on NLP tasks, ensuring the validity and reliability of the findings. Subsequent chapters build on this foundation, detailing the implementation, experimentation, and evaluation processes.

Chapter 3 provided an overview of morphological segmentation. The chapter commenced with an exploration of the background and history of isiZulu; it discussed linguistic structure, morphology, and typology and their underlying concepts to be used in the rest of the study. In this chapter, morphological segmentation as a response to the complexity of isiZulu was presented, including different works that have attempted this task using different techniques, ranging from rule-based to data-driven machine learning systems, those

using annotated or unannotated data, either being statistical or deep learning based. The chapter concluded with a discussion of segmentation evaluation methods, covering both intrinsic (direct performance assessment) and extrinsic (impact on downstream tasks) evaluation techniques. This provided a framework for assessing segmentation effectiveness in subsequent chapters.

Chapter 4 detailed the processes involved in preparing a syntactic dataset for morphological segmentation in isiZulu using a rule-based system, namely ZRG. The ZRG pipeline was introduced as a data generation tool, with a discussion on its strengths and limitations. The chapter outlined the need to construct a diverse and representative dataset; hence, various open-source corpora were collected and cleaned to be used for this task of data preparation. Key data preparation steps included parsing isiZulu sentences into abstract syntax trees to capture morphological structures, recursive decomposition of trees to retain linguistic information, and incorporating data augmentation strategy via syntactic tree manipulation to enhance dataset diversity without extensive manual annotation. To systematically investigate segmentation granularity, three linearisation strategies were introduced: Lin A (Fine-grained segmentation), Lin B (Middle-grained segmentation), and Lin C (Coarse-grained segmentation), each tailored for different levels of segmentation complexity and designed to improve adaptability in NLP applications. These strategies facilitate flexibility for various NLP applications, allowing for a more systematic investigation of segmentation approaches.

Chapter 5 presented the methodology and techniques employed to develop and optimise supervised machine learning models for morphological segmentation using the isiZulu synthetic surface segmented text. Various techniques were explored, including the adaptation and modification of Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) models and the implementation of the Transformer model from scratch to allow for a systematic comparative analysis and selection of one machine learning system to use further for evaluation, both intrinsically and extrinsically. During implementation and training, hyper-parameter optimisation (HPO) strategies were applied across all models to determine the best configurations for final training and evaluation. After identifying the optimal hyper-parameters, different supervised machine learning models were trained using three distinct linearisation strategies based on data generated from ZRG, a rule-based system, creating a hybrid segmentation framework that combines rule-based and machine learning paradigms. To ensure a fair comparison, the models were adapted to include similar features, and a consistent 80:10:10 data-splitting ratio was maintained for training, validation, and testing.

Chapter 6 provided a comprehensive evaluation of the morphological segmentation models trained for isiZulu using different granularities. During the evaluation, two forms of evaluation were considered: intrinsic and extrinsic evaluation. In terms of intrinsic evaluation, the different models based on CRF, LSTM, and Transformer architectures created

in Chapter 5 were evaluated in terms of how well they performed the task of segmentation using chrF, BLEU, precision, recall, and F score as evaluation metrics. The intrinsic evaluation revealed clear performance differences among the Transformer, LSTM, and CRF-based architectures, confirming the Transformer architecture as the most effective for supervised segmentation. The Transformer-based models consistently outperformed other models across all segmentation granularities, achieving the highest BLEU, precision, recall, F1 score, and chrF scores. Given these results, Transformer-based models were further evaluated for *robustness* and *efficiency*. These models demonstrated superior handling of previously unseen words or lexicons compared to the ZRG-based rule system, and also significantly improved segmentation speed, confirming that the hybrid system successfully mitigates the weaknesses of purely rule-based approaches.

In terms of extrinsic evaluation, this assessment examined the impact of different segmentation granularities on NMT performance. Models trained on segmented isiZulu data consistently outperformed the unsegmented baseline across all evaluation metrics. Using the NLLB test set, Segmenter Two achieved the highest BLEU and chrF scores, suggesting that moderate segmentation granularity provides a useful balance between vocabulary reduction and preservation of linguistic structure. However, in the external FLORES test set, which contained more unseen words and was not a subset of the original data, Segmenter One performed best, indicating its potential for improved generalisation in low-resource settings. Interestingly, Segmenter Two became the model with the lowest performance in FLORES, slightly underperforming even compared to Segmenter Three. These results suggested that, while moderate segmentation performs well when the test data is similar to the training distribution, finer-grained segmentation may be more beneficial for handling unseen words. In general, these findings highlighted the role of segmentation in addressing data sparsity and improving translation quality, particularly for morphologically complex languages such as isiZulu.

7.3 Findings Regarding Objectives and Research Questions

This section provides an overview of the research questions and hypotheses, and presents the findings in relation to how much of the intended research work was achieved in relation to the objectives and hypotheses. The main objective of this work was that presented in Chapter 1 in Section 1.4.2, namely: “To investigate supervised surface segmentation of isiZulu text using synthetic data generation”, with five sub-objectives:

- *Sub-objective 1:* To utilise the ZRG to parse a large corpus of isiZulu text to obtain a treebank.
- *Sub-objective 2:* To utilise the ZRG to augment the treebank to obtain additional synthetic data.

- *Sub-objective 3*: To investigate the suitability of different machine learning models for the supervised segmentation task.
- *Sub-objective 4*: To train segmentation models on different linearisations of the treebanks.
- *Sub-objective 5*: To evaluate segmentation models intrinsically and extrinsically according to suitable metrics.

These objectives were supporting pillars in terms of answering the two research questions, which were:

1. How can a hybrid approach to supervised surface segmentation be pursued that leverages the strengths of a rule-based approach to overcome the weaknesses of a machine learning approach in the context of isiZulu?
2. How can such a hybrid approach be used to obtain insights into supervised surface segmentation strategies for isiZulu?

In order to address these research questions and to align with the objectives of the study, a clear methodology was required. This involved establishing a structured approach that guided the selection and application of specific methods within a defined research paradigm. This study adopted the hypothetico-deductive method, which aligns with the positivist philosophical framework, emphasising objectivity and empirical evidence in conducting research, based on the formulated hypotheses.

From an epistemological perspective, this methodology assumed that a valid and reliable knowledge about the relationship between segmentation granularity and downstream task performance can be obtained through empirical investigation. This assumption gave rise to the design of comprehensive and thorough experiments with the aim of generating scientifically grounded insights that are reliable and reproducible. This allowed the results obtained to be used as a contribution to the broader scientific discourse, provided that the findings are systematic, testable, and transferable.

Similarly, the axiological stance of this research aligns with positivist principles, particularly in terms of maintaining objectivity. Since this study involves hypothesis testing and comparative evaluation, it is crucial to eliminate bias that could influence the experimental process or outcomes. By adhering to this structured approach, the study ensures that its findings are systematic, testable, and transferable, making them valuable contributions to the broader scientific discourse.

7.3.1 Summary of Achievements in Relation to the Objectives

This section outlines the progress made toward achieving the research objectives, ultimately tying back to the main objective of this study: *“To investigate supervised surface segmentation of isiZulu text using synthetic data generation.”*

Each sub-objective is systematically addressed below, highlighting how it was addressed and achieved throughout the course of the study.

7.3.1.1 Sub-Objective 1

This objective aimed to utilise ZRG to parse a large corpus of isiZulu text and generate a treebank.

In Chapter 4, the study outlined the process of curating, cleaning, and preparing a diverse set of isiZulu datasets from various open-source corpora. These datasets were then parsed using ZRG, which produced abstract syntax trees (ASTs) that structurally represent isiZulu sentences in a hierarchical format. These parsed ASTs were collected and compiled into a treebank, effectively achieving this objective.

7.3.1.2 Sub-Objective 2

Sub-objective 2 entailed utilising the ZRG to augment the treebank to obtain additional synthetic data.

In Chapter 4, the study demonstrated how treebank augmentation was performed by modifying syntactic structures and lexical items within ASTs. This approach facilitated the creation of new ASTs that are different from the ones that were generated by the parser; this process formed part of data augmentation, significantly expanding the dataset without requiring extensive parsing, which has been presented as a slow process. These modified ASTs can be linearised back into isiZulu text, creating more words, phrases, or sentences that are grammatically correct, yet potentially creating semantically incorrect sentences and phrases.

7.3.1.3 Sub-Objective 3

This objective focused on investigating the suitability of various supervised machine learning models for morphological segmentation.

Addressing this objective spanned chapters:

1. Chapter 3, particularly Section 3.4.2.2 and 3.4.3.1, provided a theoretical overview of statistical and deep learning-based segmentation methods, including CRFs, LSTM networks, and Transformers.
2. Chapter 5 detailed the selection and adaptation of these models for isiZulu segmentation, examining their architecture and training methodologies.

By the end of this investigation, CRF, LSTM, and Transformer-based models were chosen for further comparative analysis, setting the stage for subsequent objectives.

7.3.1.4 Sub-Objective 4

This objective required training segmentation models using different granularities of segmentation, determined by varying linearisation strategies.

To achieve this objective, the ZRG tool was adapted into three linearisation strategies, namely: *Lin A* (Fine-grained segmentation), *Lin B* (Middle-grained segmentation), and *Lin C* (Coarse-grained segmentation).

These strategies allowed the treebanks to be linearised into differently segmented text representations, ensuring a diverse range of segmentation styles for model training. Chapter 5 outlined the training process, in which the CRF, LSTM, and Transformer models were trained on these datasets, and hyper-parameter tuning was performed to optimise their performance systematically.

7.3.1.5 Sub-Objective 5

This objective focused on assessing the effectiveness of segmentation models, both in isolation (intrinsic evaluation) and in a downstream application (extrinsic evaluation).

Once segmentation models have been trained, they were evaluated based on performance metrics such as BLEU, chrF, precision, recall, and F1 score, as discussed in Chapter 6. Among the architectures evaluated, the Transformer-based models demonstrated superior effectiveness in morphological segmentation, consistently achieving the highest scores across all metrics.

Additionally, the robustness and efficiency of the Transformer-based models were tested against the rule-based ZRG system, demonstrating that the hybrid approach successfully mitigates the brittleness and inefficiency of ZRG. Furthermore, Chapter 6 examined the

impact of segmentation on Neural Machine Translation (NMT), confirming that segmentation improves translation quality and that Segmenter One (fine-grained segmentation) was the most effective in handling unseen data.

7.4 Evaluating the Hypotheses

Having addressed the sub-objectives outlined earlier, this section presents an evaluation of the formulated hypotheses based on empirical findings. The discussion aims to determine whether the hypotheses are supported or not supported, thereby assessing the overall achievement of this study and providing insights into the research questions posed at the outset.

7.4.1 First Hypothesis: Performance of the Hybrid System

The first hypothesis pertains to the effectiveness of the hybrid system that was developed, which integrates rule-based ZRG and supervised machine learning models for morphological segmentation. This is particularly important, given the low-resource nature of isiZulu, where limited digital and annotated datasets create challenges in terms of NLP development.

The hybrid system was designed with the expectation that it would overcome the limitations of both standalone rule-based and standalone machine learning approaches. The ZRG system, despite its linguistically informed segmentation, suffers from:

1. Lack of robustness (failure to segment words outside its lexicon).
2. Computational inefficiency (slow segmentation due to deep morphosyntactic analysis).

On the other hand, supervised machine learning models are more robust and efficient, but require large annotated datasets to perform satisfactorily, which are scarce for isiZulu. Given that ZRG could generate synthetic data for training supervised models, the hybrid approach was hypothesised to offer an optimal balance between linguistic accuracy and data-driven generalisation.

To test this assumption, the following sub-hypotheses were formulated and empirically evaluated.

7.4.1.1 Testing the First Sub-Hypothesis (H1a: Robustness Improvement)

H1a: *The hybrid approach will exceed the rule-based approach in robustness*

In Chapter 6, an experiment was conducted to evaluate how well the hybrid system handles words that ZRG failed to segment due to lexicon limitations. The segmentation performance of the hybrid system was assessed using intrinsic evaluation metrics, where Segmenter One (Lin A), Segmenter Two (Lin B), and Segmenter Three (Lin C) achieved: 0.90, 0.78 and 0.60 average scores across the different metrics, respectively.

These results indicate that, across all segmentation strategies, the hybrid system consistently outperformed the rule-based ZRG system, which failed to segment the unseen words altogether. This empirical evidence strongly supports H1a, confirming that the hybrid system improves robustness, particularly in handling tokens that fall outside its coverage.

7.4.1.2 Testing the Second Sub-Hypothesis (H1b: Efficiency Improvement)

H1b: *The hybrid approach will exceed the rule-based approach in efficiency.*

To assess efficiency, a comparative experiment was conducted in Chapter 6, measuring the time required for both systems to segment 100 words. The Transformer-based hybrid system completed the segmentation in under 70 seconds for all segmentation, whilst the ZRG took around 1180 seconds. The hybrid system completed the segmentation process nearly 30-40 times faster than ZRG. The inefficiency of the ZRG model can be ascribed to its parsing-based segmentation, which involves complex morphosyntactic analysis, while the hybrid model focuses only on the segmentation task, thus significantly reducing processing time. This finding strongly supports H1b, validating that the hybrid system significantly improves computational efficiency, making it more viable for large-scale applications.

7.4.1.3 Testing the Third Sub-Hypothesis (H1c: Sufficient Performance)

H1c: *The hybrid approach will reach a sufficient level of accuracy in comparison to the rule-based approach.*

The previous sub-objectives established that the hybrid system effectively overcomes two key limitations of the ZRG system: inefficiency and lack of robustness. However, an essential question remains: is the hybrid model accurate enough to serve as a reliable alternative for segmentation?

To evaluate this question, intrinsic metrics were used to compare the hybrid approach with

ZRG in Chapter 6. Since ZRG generated the ground truth, this comparison reflects an *idealised* scenario, assuming that ZRG was absolutely perfect in segmentation. However, ZRG itself is not fully robust, and the datasets used for training and testing were limited to instances in which ZRG could successfully parse and linearise.

Intrinsic evaluation results from Section 6.2 showed that the hybrid model consistently achieved segmentation accuracy comparable to the *idealised* ZRG output, with scores averaging around 0.9 across all metrics.

Although the hybrid system does not exceed ZRG in absolute accuracy – since ZRG itself generated the ground truth – the results nonetheless indicate that it reaches a sufficient level of accuracy to be considered a viable alternative. More importantly, when combined with its superior generalisation to unseen words and significant computational efficiency gains, the hybrid approach proves to be a practical and scalable solution for isiZulu morphological segmentation.

These findings support H1c, affirming that the hybrid system maintains accuracy while addressing the fundamental limitations of ZRG.

7.4.1.4 Final Remarks on the First Hypothesis

The hypotheses H1a (Robustness Improvement), H1b (Efficiency Improvement), and H1c (Sufficient Level of Accuracy) are strongly supported by empirical evidence. The hybrid system effectively addresses the weaknesses of the rule-based ZRG approach and reaches a sufficient level to be considered a viable alternative, demonstrating the following:

1. *Robustness Improvement (H1a)*: The hybrid system successfully segments previously unseen words, overcoming the coverage limitations of ZRG.
2. *Efficiency Improvement (H1b)*: The hybrid model significantly improves the processing speed, making segmentation much faster than ZRG.
3. *Sufficient Level of Accuracy (H1c)*: The hybrid system achieves an accuracy comparable to an idealised ZRG, demonstrating its viability as an alternative.

Although scientific hypotheses are never fully accepted in an absolute sense, the results provide strong empirical support for the claim that the hybrid approach surpasses the rule-based system in terms of robustness and efficiency, and can be considered as a practical alternative for the morphological segmentation of isiZulu.

7.4.2 Second Hypothesis

In Chapter 4, morphological segmentation was motivated as an enabling technology that serves as an important preprocessing step for languages with complex morphology before their data can be used in training language models. It plays an important role in reducing data sparsity by addressing the challenge of OOV words, therefore, improving model performance on unseen words.

However, a dilemma arises when one attempts morphology from the surface segmentation route. A language with complex morphology like isiZulu, which has a high degree of morphophonological alternation, introduces ambiguity in determining where the morpheme boundary should be placed, because there is no single “correct” answer from a linguistic perspective.

Although this flexibility in terms of segmentation can be beneficial, the optimal granularity for training language models remains an open research question. If segmentation is too aggressive, language models may struggle with excessive compositional complexity, making it difficult to capture meaning accurately. In contrast, if segmentation is too conservative, it can not adequately address data sparsity, limiting the model’s ability to generalise to unseen words.

Thus, this study set out to identify which segmentation strategy (Lin A, Lin B, or Lin C) yields the best performance for machine translation, building on the widely accepted claim that morphological segmentation improves language modelling by reducing data sparsity.

To systematically investigate this, the following hypothesis was formulated:

H2: The hybrid approach will result in one or more segmenters that improve performance in a downstream task such as machine translation.

7.4.2.1 Testing the Second Hypothesis

To evaluate H2, a series of extrinsic evaluations were conducted in Chapter 6, focusing on NMT from isiZulu to English. The experiment was carefully controlled, ensuring:

1. Consistent architecture: All models used the same Transformer architecture.
2. Equal training conditions: Models were trained on the same number of data points.
3. Single variable testing: The only variable changed was the dataset segmentation strategy.
4. Robust evaluation: Two testing datasets were used, the original NLLB test set

(Costa-Jussà *et al.*, 2022) and the updated FLORES dataset (Abdulmumin *et al.*, 2024).

The results clearly demonstrated that:

1. All three segmented models (Lin A, Lin B, Lin C) outperformed the unsegmented model across all evaluation metrics (BLEU, chrF).
2. Segmenter Two (Lin B) achieved the highest BLEU and chrF scores on the original NLLB dataset, suggesting an optimal balance between vocabulary size reduction and linguistic granularity.
3. On the updated FLORES dataset, Segmenter One (Lin A) outperformed others, demonstrating slightly better resilience to OOV words and better generalisation in a low-resource setting.

These findings strongly support hypothesis H2, and thus confirm that at least one segmentation strategy improves machine translation performance. Furthermore, the results reinforce the broader claim that morphological segmentation effectively reduces data sparsity and enhances the generalisation capabilities of NMT models.

7.4.2.2 Final Remarks on the Second Hypothesis

While scientific hypotheses are never accepted absolutely, the empirical evidence strongly supports H2, showing that:

1. Morphological segmentation enhances translation performance by improving tokenisation, reducing OOV words, and increasing linguistic compositionality.
2. The optimal segmentation strategy depends on the dataset; Lin B (medium granularity) is ideal for controlled datasets, while Lin A (fine granularity) performs better in more diverse, unseen data settings.
3. All segmentation strategies outperform unsegmented baselines, highlighting the necessity of preprocessing techniques for morphologically rich languages.
4. These findings contribute valuable insights to low-resource language processing, particularly in optimising segmentation for downstream NLP applications.

7.5 Answering the Research Questions

With the evaluation of the hypotheses completed, this section systematically answers the research questions formulated at the outset of the study. These questions were designed to guide the research toward understanding how a hybrid approach to isiZulu morphological segmentation can be developed, evaluated, and applied to downstream NLP tasks.

7.5.1 Research Question 1

Question: How can a hybrid approach to supervised surface segmentation be pursued that leverages the strengths of a rule-based approach to overcome the weaknesses of a machine learning approach in the context of isiZulu?

This study proposed and implemented a hybrid segmentation framework that integrates the ZRG, a rule-based approach, with supervised machine learning models (Transformer, LSTM, and CRF). The framework was designed to leverage the linguistic knowledge embedded in ZRG while harnessing the efficiency and robustness of machine learning models.

The first step towards answering this question involved using ZRG to generate synthetic segmentation data, which served as training data for the supervised models. Three segmentation strategies (Lin A, Lin B, and Lin C) were introduced to explore different segmentation granularities, each reflecting varying degrees of morphological decomposition.

The hybrid system was empirically evaluated, demonstrating that:

1. It surpasses ZRG in robustness, effectively segmenting words that ZRG failed to process due to lexicon limitations.
2. It significantly improves efficiency, reducing segmentation time by a factor of 30-40x compared to ZRG.
3. It maintains a high level of segmentation accuracy, with intrinsic evaluation scores comparable to an idealised ZRG output (0.9 across multiple metrics).

Thus, a hybrid approach to supervised surface segmentation can be pursued by:

1. Using rule-based parsing (ZRG) to generate synthetic training data.
2. Training machine learning models on this data to learn segmentation patterns.
3. Evaluating the trained models using intrinsic and extrinsic metrics to ensure they balance accuracy, efficiency, and robustness.

4. This study successfully implemented and validated such an approach, demonstrating its viability for isiZulu morphological segmentation.

7.5.2 Research Question 2

Question: *How can such a hybrid approach be used to obtain insights into supervised surface segmentation strategies for isiZulu?*

This study systematically investigated how different segmentation strategies impact model performance, both in terms of intrinsic segmentation accuracy and extrinsic effectiveness in Neural Machine Translation (NMT).

To answer this question, three different models were developed and evaluated using three different granularities: Lin A (Fine-grained segmentation), Lin B (Middle-grained segmentation), and Lin C (Coarse-grained segmentation) were evaluated.

The key insights obtained are:

1. Segmentation granularity affects intrinsic performance:
 - (a) Segmenter One (Lin A) consistently achieved the highest scores across the evaluation metrics.
 - (b) Segmenter Two (Lin B) scored slightly lower than Segmenter One, but maintained strong performance across most metrics.
 - (c) Segmenter Two (Lin C) achieved the lowest scores in most metrics compared to Segmenters One and Two, yet still maintained an average score of 0.9 across all metrics.
2. Segmentation improves machine translation quality:
 - (a) Models trained on segmented isiZulu data outperformed the unsegmented baseline in NMT evaluation measures using the BLEU and chrF.
 - (b) Lin A performed the best on the FLORES dataset, demonstrating its resilience to data sparsity.
 - (c) Lin B performed best in the NLLB dataset, highlighting the balance between reducing vocabulary size and preserving linguistics.

These insights suggest that segmentation must be tailored to specific NLP tasks. While fine-grained segmentation improves generalisation, moderate segmentation may be preferable in context-sensitive applications of machine translation.

7.6 Challenges and Limitations

This section outlines the key challenges encountered during the study and the limitations that influenced the findings. Although the research successfully demonstrated the benefits of hybrid morphological segmentation, certain constraints impacted the scalability, evaluation, and general applicability of the approach.

7.6.1 Challenges

7.6.1.1 Limited Availability of Annotated isiZulu Data

A major challenge in this study was the scarcity of isiZulu data, which directly impacted the selection of a suitable downstream task for evaluating the performance of the developed segmentation models. Given the lack of large, high-quality, open-source datasets for various NLP tasks, machine translation (MT) was chosen as the primary downstream task due to the relative availability of parallel isiZulu-English datasets, specifically the Autshumato and NLLB corpora. However, these datasets presented additional challenges, including translation inaccuracies and sentence misalignment issues, which required further preprocessing and filtering to ensure data quality.

This was also the main cause for a limitation to the study, namely that the segmentation models were only evaluated extrinsically on the MT task. While other tasks, such as sentiment analysis, text classification, or named entity recognition (NER), could have provided additional insights into segmentation performance, the lack of sufficient labelled data for these tasks prevented their inclusion. Consequently, the findings of this study are primarily framed within the context of machine translation, leaving room for future research to explore segmentation’s impact on a broader range of NLP applications.

7.6.1.2 Computational Constraints and Training Costs

The study also faced computational constraints, particularly during the training of deep learning models. Transformer-based architectures require substantial computational resources, making it challenging to fully explore hyper-parameter tuning and model optimisation and to use all. Due to hardware limitations, hyper-parameter tuning was restricted to 10 trials per model, which may have limited the discovery of optimal configurations. Bayesian optimisation (OptunaSearch) was used to efficiently explore parameter spaces, mitigating some of these constraints. However, additional fine-tuning with more trials and increased computational resources could have led to even better model performance.

7.6.2 Limitations

The findings of this study have to be seen in the light of the following limitations, which are detailed below.

7.6.2.1 Absence of Human Evaluations for Translation

While the extrinsic evaluation provided valuable insights into the impact of segmentation on NMT performance, a key limitation of this study is the exclusive reliance on automatic evaluation metrics such as BLEU and chrF. Although these metrics offer objective and reproducible measures of translation quality, they do not fully capture aspects such as fluency, adequacy, and semantic correctness.

Furthermore, in the English-to-isiZulu translation task, chrF scores were nearly identical to those observed in the isiZulu-to-English translation. This suggests that BLEU may be penalising models more harshly for morphological variations within tokens, which is particularly relevant in languages with complex morphology.

To address these challenges, future research should incorporate human evaluation to provide a more nuanced understanding of translation performance. Furthermore, the development or integration of evaluation metrics that account for semantic similarity and morphological variations could lead to more accurate assessments of translation quality, particularly for morphologically rich languages like isiZulu.

7.6.2.2 Lack of Direct Comparison with Other Models

One of the limitations of this study is the absence of direct comparisons between the proposed hybrid segmentation approach and other existing segmentation models beyond the ZRG-based approach. While the intrinsic evaluation effectively measured segmentation accuracy within the hybrid system, and the extrinsic evaluation assessed its impact on machine translation performance, the study did not benchmark the segmentation models on the extrinsic evaluation against other widely used segmentation techniques, such as unsupervised morphological segmenters (e.g., Morfessor) or subword-based approaches (e.g., Byte-Pair Encoding [BPE] and SentencePiece).

Future research can consider comparative analyses with alternative segmentation models and their impact on downstream tasks. This would allow for a more comprehensive evaluation of the hybrid approach's strengths and weaknesses, ensuring that the proposed method is benchmarked against existing techniques used in low-resource language processing.

7.6.2.3 Generalisation to Other Low-Resource Languages

A limitation of this study is the extent to which the proposed hybrid segmentation approach can be generalised to other low-resource languages, particularly those within the Bantu language family. While isiZulu serves as a representative case study for morphologically rich and conjunctively written languages, the effectiveness of this approach in terms of application to other languages depends on the availability of structured linguistic resources such as resource grammars.

Currently, resource grammars for Bantu languages remain limited, posing a challenge for directly applying the ZRG-based hybrid segmentation framework to other languages. However, ongoing efforts are addressing this gap. A resource grammar already exists for Siswati (Marais *et al.*, 2024), and isiXhosa grammar is actively being developed. As more grammars become available, it is expected that the methodology proposed in this study can be extended to a broader range of Bantu languages.

Future research should explore how the hybrid segmentation approach adapts to different Bantu languages, considering language-specific morphological and syntactic variations. Additionally, developing strategies to extend the framework to languages without existing resource grammars, potentially through transfer learning or bootstrapping techniques, would further enhance its applicability across diverse low-resource settings.

7.6.2.4 Controlled Research Environment

Lastly, the study was conducted in a controlled research environment, where training datasets, models, and evaluation criteria were carefully curated. However, real-world NLP applications often involve noisy, inconsistent, or domain-specific data, which may affect model performance differently than observed in controlled experiments. In this regard, to gain more practical insights into the robustness and effectiveness of the developed model, one may need to deploy the segmentation models in real-NLP applications.

7.7 Conclusion

This study aimed to investigate the supervised surface segmentation of isiZulu text using synthetic data generation, addressing key challenges associated with data sparsity, morphological complexity, and low-resource NLP. Using a hybrid approach, combining rule-based segmentation using the ZRG and supervised machine learning techniques, this study provided a structured and scalable solution to handle the agglutinative morphology of isiZulu. The study systematically explored different segmentation granularities, evaluating their effectiveness both intrinsically (accuracy, robustness, and efficiency) and

extrinsically (impact on a downstream NLP task, namely Neural Machine Translation).

The intrinsic evaluation of segmentation models revealed that the Transformer-based approach outperformed other supervised models (LSTM, CRF) across all segmentation granularities. Further analyses demonstrated that the hybrid segmentation framework developed was significantly more robust and efficient than the ZRG rule-based approach alone, addressing key limitations such as brittleness in handling OOV words and slow parsing speeds. This confirmed the validity of the hybrid approach, reinforcing its potential as a scalable and adaptable solution for isiZulu segmentation.

Extrinsic evaluation confirmed that morphological segmentation enhances the performance of machine translation systems. Across multiple datasets, including the NLLB and the updated FLORES dataset, NMT models trained on segmented isiZulu data consistently outperformed unsegmented baselines. The results validated the claim that segmenting morphologically rich languages before translation reduces data sparsity, improves vocabulary generalisation, and enhances translation quality. Notably, different segmentation strategies produced varying results, highlighting the trade-offs between segmentation granularity and translation performance.

Despite these successes, several challenges and limitations remain. The availability of high-quality annotated isiZulu data remains a key issue, limiting the scope of this study. Additionally, the impact of segmentation on other NLP applications beyond machine translation, such as word embeddings, text-to-speech synthesis, and information retrieval, requires further exploration. Moreover, the evaluation methodology did not include direct benchmarking with alternative segmentation approaches, human evaluation, or statistical significance testing, all of which could provide a more rigorous assessment of segmentation's impact on model performance. Addressing these gaps in future work could further validate and refine the conclusions drawn from this research.

Ultimately, this study demonstrated the value of hybrid morphological segmentation for isiZulu, offering an empirically validated, scalable approach that improves both segmentation accuracy and downstream NLP performance. By tackling key challenges in low-resource language processing, this work lays the foundation for future advancements in African language NLP, promoting greater linguistic inclusivity and accessibility in language technologies.

Bibliography

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., *et al.* (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Abdulmumin, I., Mkhwanazi, S., Mbooi, M. S., Muhammad, S. H., Ahmad, I. S., Putini, N., Mathebula, M., Shingange, M., Gwadabe, T., & Marivate, V. (2024). Correcting FLORES evaluation dataset for four African languages. *arXiv preprint arXiv:2409.00626*.
- Adnan, M., Alarood, A. A. S., Uddin, M. I., & ur Rehman, I. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Computer Science*, 8, e803.
- Agic, Ž., & Vulic, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages.
- Aguiar, G. J. M. d. (2024). Distinguishing between Method and Methodology in Academic Research. *Available at SSRN 4857923*.
- Ahmed, A. (2008). Ontological, Epistemological and Methodological Assumptions: Qualitative versus Quantitative. *Online Submission*.
- Albahli, S., Alhassan, F., Albattah, W., & Khan, R. U. (2020). Handwritten digit recognition: Hyperparameters-based analysis. *Applied Sciences*, 10(17), 5988.
- Alele, F., & Malau-Aduli, B. (2023). *An introduction to research methods for undergraduate health profession students*. James Cook University Townsville, Qld.
- Alghamdi, A. H., & Li, L. (2013). Adapting design-based research as a research methodology in educational settings. *International Journal of Education and Research*, 1(10), 1–12.
- Ali, M. (2023). Research Philosophies in Social Science and Information Systems Research. In *Information Systems Research: Foundations, Design and Theory* (pp. 23–34). Springer.
- Anand Kumar, M., Dhanalakshmi, V., Soman, K., & Rajendran, S. (2010). A sequence labeling approach to morphological analyzer for tamil language. *International Journal on Computer Science and Engineering*, 2(06), 1944–1951.
- Angelov, K. (2009). *Gf runtime system*. Chalmers University of Technology.

- Angelov, K., Bringert, B., & Ranta, A. (2010). PGF: A portable run-time format for type-theoretical grammars. *Journal of Logic, Language and Information*, 19, 201–228.
- Ansari, E., Žabokrtský, Z., Mahmoudi, M., Haghdoost, H., & Vidra, J. (2019). Supervised morphological segmentation using rich annotated lexicon. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 52–61.
- Asante, M. K., & Mazama, A. (2009). *Encyclopedia of African religion*.
- Ataman, D., & Federico, M. (2018). Compositional representation of morphologically-rich input for neural machine translation. *arXiv preprint arXiv:1805.02036*.
- Ataman, D., Negri, M., Turchi, M., & Federico, M. (2017). Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108(1), 331–342.
- Ba, J. L. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Badr, I., Zbib, R., & Glass, J. (2008). Segmentation for English-to-Arabic statistical machine translation. *Proceedings of ACL-08: HLT, Short Papers*, 153–156.
- Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Bahree, A. (2024). *Generative AI in Action*. Simon and Schuster.
- Banko, M., & Moore, R. C. (2004). Part-of-speech tagging in context. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 556–561.
- Bansal, R., Choudhary, H., Punia, R., Schenk, N., Dahl, J. L., & Pagé-Perron, É. (2021). How low is too low? A computational perspective on extremely low-resource languages. *arXiv preprint arXiv:2105.14515*.
- Bastin, Y., Coupeuz, A., & Mann, M. (1999). Continuity and divergence in the Bantu languages : perspectives from a lexicostatistic study. <https://api.semanticscholar.org/CorpusID:56137712>
- Beesley, K. R. (1996). Arabic finite-state morphological analysis and generation. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
- Beesley, K. R., & Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*, 359–375.
- Bel-Enguix, G., Grando, M. A., & Jiménez-López, M. D. (2006). A grammatical framework for modelling multi-agent dialogues. *Agent Computing and Multi-Agent Systems: 9th Pacific Rim International Workshop on Multi-Agents, PRIMA 2006, Guilin, China, August 7-8, 2006. Proceedings 9*, 10–21.
- Belth, C. (2024). Meaning-Informed Low-Resource Segmentation of Agglutinative Morphology. *Proceedings of the Society for Computation in Linguistics 2024*, 96–106.
- Bergmanis, T., & Goldwater, S. (2017). From segmentation to analyses: a probabilistic model for unsupervised morphology induction. *Proceedings of the 15th Conference*

- of the European Chapter of the Association for Computational Linguistics: Volume 1, *Long Papers*, 337–346.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1), 281–305.
- Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.
- Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society Series D: The Statistician*, 24(3), 179–195.
- Bhaskar, R. (2013). *A Realist Theory of Science*. Taylor & Francis.
- Bhaskar, R. (2014). *The possibility of naturalism: A philosophical critique of the contemporary human sciences*. Routledge.
- Bickel, B., & Nichols, J. (2007). Inflectional morphology. *Language typology and syntactic description*, 3(2), 169–240.
- Bleek, W. (1862). *A Comparative Grammar of South African Languages*. Gregg International Publishers Limited.
- Blevins, J. P. (2014). The morphology of words. *The Oxford handbook of language production*, 152–164.
- Bloomfield, L. (1995). *Language*. Henry Holt Co.
- Bolinger, D., & Sears, D. (1981). *Aspects of Language*. Harcourt Brace Jovanovich. <https://books.google.co.za/books?id=CrscAQAAIAAJ>
- Booij, G. (2012). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.
- Bosch, S. (2020). Computational morphology systems for Zulu—a comparison. *Nordic Journal of African Studies*, 29(3), 28–28.
- Bosch, S., Jones, J., Pretorius, L., & Anderson, W. (2006). Resource development for South African Bantu languages: computational morphological analysers and machine-readable lexicons, 38.
- Bosch, S. E., & Pretorius, L. (2002). The significance of computational morphological for Zulu lexicography. *South African Journal of African Languages*, 22(1), 11–20.
- Bostrom, K., & Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.
- Brent, M. (1993). Minimal generative explanations: A middle ground between neurons and triggers. *Proc. of the 15th Annual Meeting of the Cognitive Science Society*, 28–36.
- Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2), 79–85.
- Browne, A. (2000). The Potential Contributions of Critical Social Theory to Nursing Science. *The Canadian journal of nursing research = Revue canadienne de recherche en sciences infirmières*, 32, 35–55.

- Brun, C. (1998). Terminology finite-state preprocessing for computational LFG. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 196–200.
- Brusciotto, G. (1659). *Grammar of the Congo Language (as spoken two hundred years ago)* (H. G. Guinness, Trans.) [Originally published as *Regulae quaedam pro difficillimi Congensium idiomatis faciliiori captu ad grammaticae normam redactae*, Rome, 1659].
- Buell, L. C. (2005). *Issues in Zulu verbal morphosyntax* [Doctoral dissertation, CiteSeer].
- Burrell, G., & Morgan, G. (2019). *Sociological Paradigms and Organisational Analysis: Elements of the Sociology of Corporate Life*. Taylor & Francis.
- Cambridge Dictionary. (n.d.). Research [Retrieved February 11, 2025, from Cambridge Dictionary website].
- Can, B., & Manandhar, S. (2014). Methods and algorithms for unsupervised learning of morphology. *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part I 15*, 177–205.
- Cao, P. (2022). Distinction and Examples of Morpheme, Morph and Allomorph in English Linguistics Teaching. *2nd International Conference on Education: Current Issues and Digital Technologies (ICECIDT 2022)*, 323–331.
- Catford, J. C. (2001). *A practical introduction to phonetics*. Oxford University Press.
- Celikates, R., & Flynn, J. (2023). Critical theory (Frankfurt school).
- Chanod, J.-P. (1994). *Finite-state composition of French verb morphology*. Rank Xerox Research Centre.
- Chavan, R. S., & Sable, G. S. (2013). An overview of speech recognition using HMM. *International Journal of Computer Science and Mobile Computing*, 2(6), 233–238.
- Chen, C.-T., & Chang, W.-D. (1996). A feedforward neural network with function shape autotuning. *Neural Networks*, 9(4), 627–641.
- Chen, W., & Fazio, B. (2021). Morphologically-guided segmentation for translation of agglutinative low-resource languages. *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, 20–31.
- Chen, X., Qiu, X., Zhu, C., Liu, P., & Huang, X.-J. (2015). Long short-term memory neural networks for chinese word segmentation. *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1197–1206.
- Chiu, B., Korhonen, A., & Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, 1–6.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus: the Bible in 100 languages. *Language resources and evaluation*, 49, 375–395.

- Cing, D. L., & Soe, K. M. (2020). Improving accuracy of part-of-speech (POS) tagging using hidden markov model and morphological analysis for Myanmar Language. *International Journal of Electrical and Computer Engineering*, 10(2), 2023.
- Clavier, P. R., *et al.* (2012). *A service-dominant logic approach to business intelligence* [Doctoral dissertation, University of Pretoria].
- Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002)*, 1–8.
- Collis, J., & Hussey, R. (2014, January). *Business research: A practical guide for undergraduate and postgraduate students*.
- Comrie, B. (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press.
- Cope, A. T. (1966). *Zulu phonology, tonology and tonal grammar*. [Doctoral dissertation, University of Natal].
- Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., *et al.* (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Cotterell, R., Müller, T., Fraser, A., & Schütze, H. (2015). Labeled morphological segmentation with semi-Markov models. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 164–174.
- Cotterell, R., Vieira, T., & Schütze, H. (2016a). A joint model of orthography and morphological segmentation.
- Cotterell, R., Vieira, T., & Schütze, H. (2016b). A Joint Model of Orthography and Morphological Segmentation, 664–669. <https://doi.org/10.18653/v1/N16-1080>
- Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(1), 1–18.
- Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Creutz, M. (2003). Unsupervised segmentation of words using prior distributions of morph length and frequency. *Annual Meeting of the Association for Computational Linguistics*, 280–287.
- Creutz, M. (2006). *Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition*. Helsinki University of Technology.
- Creutz, M., & Lagus, K. (2002a). Unsupervised discovery of morphemes. *arXiv preprint cs/0205057*.
- Creutz, M., & Lagus, K. (2002b). Unsupervised discovery of morphemes [arXiv: cs/0205057], 21–30. <https://doi.org/10.3115/1118647.1118650>

- Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1), 1–34.
- Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pytköinen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., & Stolcke, A. (2007a). Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1), 1–29.
- Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pytköinen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., & Stolcke, A. (2007b). Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1), 1–29.
- Creutz, M. J. P., & Lagus, K. H. (2004). Induction of a simple morphology for highly-inflecting languages. *7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, 43–51.
- Creutz, M. J. P., & Lagus, K. H. (2005). Inducing the morphological lexicon of a natural language from unannotated text. *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, 106–113.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4), 303–314.
- Daelemans, W., Zavrel, J., Van Der Sloot, K., & Van den Bosch, A. (2004). Timbl: Tilburg memory-based learner. *Tilburg University*.
- Dane, F. C. (1990). Research Methods, Brooks. *Cole Pacific Grove, CaUfomia*.
- Dasgupta, S., & Ng, V. (2007). High-performance, language-independent morphological segmentation. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 155–163.
- Davenport, M., & Hannahs, S. (2020). *Introducing Phonetics and Phonology*. Taylor & Francis.
- De Marcken, C. (1995). The unsupervised acquisition of a lexicon from continuous speech. *arXiv preprint cmp-lg/9512002*.
- Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dieronitou, I. (2014). The ontological and epistemological foundations of qualitative and quantitative approaches to research. *International Journal of Economics*, 2(10), 1–17.
- Ding, S., Renduchintala, A., & Duh, K. (2019). A call for prudent choice of subword merge operations in neural machine translation. *arXiv preprint arXiv:1905.10453*.
- Dlamini, S., Jembere, E., & Pillay, A. (2020). Evaluation of word and sub-word embeddings for isizulu on semantic relatedness and word sense disambiguation tasks. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1–6.

- Doke, C. M. (1927a). The significance of class 1 Bantu nouns. In F. Boas (Ed.), *Festschrift Meinhof* (pp. 196–203). Heinrich Augustin.
- Doke, C. M. (1923). A dissertation on the phonetics of the Zulu language. *Bulletin of the School of Oriental and African Studies*, 2(4), 687–729.
- Doke, C. (1927b). *Text Book of Zulu Grammar*. University of the Witwatersrand Press. <https://books.google.co.za/books?id=BK6wvQEACAAJ>
- Du Toit, J. S., & Puttkammer, M. J. (2021). Developing core technologies for resource-scarce nguni languages. *Information*, 12(12), 520.
- Dwivedi, R. K., Nand, P., & Pal, O. (2025). Hybrid NMT model and comparison with existing machine translation approaches. *Multidisciplinary Science Journal*, 7(4), 2025146–2025146.
- Eddington, D. (2012). Morphophonological Alternations. In J. I. Hualde, A. Olarrea, & E. O'Rourke (Eds.), *The Handbook of Hispanic Linguistics* (pp. 193–208). Wiley-Blackwell. <https://doi.org/10.1002/9781118228098.ch10>
- Eiselen, R., & Puttkammer, M. J. (2014a). Developing Text Resources for Ten South African Languages, 3698–3703.
- Eiselen, R., & Puttkammer, M. J. (2014b). Developing Text Resources for Ten South African Languages. *LREC*, 3698–3703.
- El Kholy, A., & Habash, N. (2012). Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*, 26, 25–45.
- Elwell, R. (2008). Finite state methods for Bantu verb morphology. *Proceedings of Texas Linguistic*.
- Engle, E. (2008). Ontology, epistemology, axiology: bases for a comprehensive theory of law. *Appalachian JL*, 8, 103.
- Eskander, R., Rambow, O., & Yang, T. (2016). Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 900–910.
- Eskander, R., Callejas, F., Nichols, E., Klavans, J. L., & Muresan, S. (2020). MorphAGram, evaluation and framework for unsupervised morphological segmentation. *Proceedings of the 12th Language Resources and Evaluation Conference*, 7112–7122.
- Eskander, R., Lowry, C., Khandagale, S., Klavans, J., Polinsky, M., & Muresan, S. (2022). Unsupervised Stem-based Cross-lingual Part-of-Speech Tagging for Morphologically Rich Low-Resource Languages. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Faaß, G., Bosch, S., & Taljard, E. (2012). Towards a Part-of-Speech Ontology: Encoding Morphemic Units of Two South African Bantu Languages. *Nordic Journal of African Studies*, 21(3), 23–23.
- Feng, H., Chen, K., Deng, X., & Zheng, W. (2004). Accessor variety criteria for Chinese word extraction. *Computational linguistics*, 30(1), 75–93.

- Feng, H., Chen, K., Kit, C., & Deng, X. (2005). Unsupervised segmentation of Chinese corpus using accessor variety. *Natural Language Processing-IJCNLP 2004: First International Joint Conference, Hainan Island, China, March 22-24, 2004, Revised Selected Papers 1*, 694–703.
- Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*.
- Flight, C. (1980). Malcolm Guthrie and the reconstruction of Bantu prehistory. *History in Africa*, 7, 81–118.
- Fosl, P. S., & Baggini, J. (2020). *The philosopher's toolkit: a compendium of philosophical concepts and methods*. John Wiley & Sons.
- Fowler, F. (2014). *Survey Research Methods*. SAGE Publications. <https://books.google.co.za/books?id=GcKhAQAAQBAJ>
- Gage, P. (1994). A new algorithm for data compression. *The C Users Journal*, 12(2), 23–38.
- Garg, R. (2016). Methodology for research I. *Indian journal of anaesthesia*, 60(9), 640–645.
- Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings, 2020*, 191.
- Goldkuhl, G. (2012). Pragmatism vs interpretivism in qualitative information systems research. *European journal of information systems*, 21(2), 135–146.
- Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2), 153–198.
- Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1), 21–54.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>
- Goundar, S. (2012). Research methodology and research method. *Victoria University of Wellington*, 1(1), 1–47.
- Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10, 522–538.
- Goyal, P., Pandey, S., & Jain, K. (2018). Deep learning for natural language processing. *New York: Apress*.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.
- Groenewald, H. J., & Fourie, W. (2009). Introducing the autshumato integrated translation environment. *Proceedings of the 13th Annual conference of the European Association for Machine Translation*.

- Grönroos, S.-A., Virpioja, S., Smit, P., & Kurimo, M. (2014). Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1177–1185.
- Grönroos, S.-A., Virpioja, S., & Kurimo, M. (2016). Hybrid morphological segmentation for phrase-based machine translation. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 289–295.
- Grout, L. (1859). *The Isizulu. A grammar of the Zulu language*. Dalcassian Publishing Company.
- Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of qualitative research*, 2(163-194), 105.
- Guo, Z., Zhang, Y., Su, C., & Xu, J. (2012). Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation. *Natural Language Processing and Chinese Computing: First CCF Conference, NLPCC 2012, Beijing, China, October 31-November 5, 2012. Proceedings 1*, 121–131.
- Guthrie, M. (1967). Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages. (*No Title*).
- Halidanmu, A., Yong, C., Yang, L., & Maosong, S. (2017). Uyghur morphological segmentation with bidirectional GRU neural networks. *Journal of Tsinghua University (Science and Technology)*, 57(1), 1–6.
- Hammarström, H., & Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics*, 37(2), 309–350. https://doi.org/10.1162/COLI.a_00050
- Hapke, H., Howard, C., & Lane, H. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Simon; Schuster.
- Harris, Z. S. (1955). From Phoneme to Morpheme. *Language*, 31, 32–67. <https://api.semanticscholar.org/CorpusID:203462101>
- Haspelmath, M. (2023). Defining the word. *Word*, 69(3), 283–297.
- Haspelmath, M., & Sims, A. (2013). *Understanding morphology*. Routledge.
- Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction.
- Hatch, M. J. (2018). *Organization theory: Modern, symbolic, and postmodern perspectives*. Oxford University Press.
- Hawthorne, J. (2004). Inductive Logic (E. N. Zalta, Ed.). *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/logic-inductive/>
- Hazen, K. (2014). *An Introduction to Language*. Wiley.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Heffernan, K., Çelebi, O., & Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:2205.12654*.

- Herbert, R. K. (1977). Morphophonological palatalization in Southern Bantu: A reply to segmental fusion. *Studies in African Linguistics*, 8(2), 143.
- Hochreiter, S. (1997). Long Short-term Memory. *Neural Computation MIT-Press*.
- Holmén, J. (2020). *Navigating Sustainability Transformations: Backcasting, Transdisciplinarity and Social Learning* [Doctoral dissertation, Chalmers University of Technology]. https://research.chalmers.se/publication/520462/file/520462_Fulltext.pdf
- Howell, K. E. (2012). An introduction to the philosophy of methodology.
- Hrycej, T., Bermeitinger, B., & Handschuh, S. (2022). Number of Attention Heads vs Number of Transformer-Encoders in Computer Vision. *arXiv preprint arXiv:2209.07221*.
- Huck, M., Riess, S., & Fraser, A. (2017). Target-side word segmentation strategies for neural machine translation. *Proceedings of the Second Conference on Machine Translation*, 56–67.
- Hurskainen, A. (1992). A two-level computer formalism for the analysis of Bantu morphology. An application to Swahili. *Nordic journal of African studies*, 1(1).
- Irshaidat, R. (2022). Interpretivism vs. positivism in political marketing research. *Journal of Political Marketing*, 21(2), 126–160.
- Jaca-Madariaga, M., Zarrabeitia Bilbao, E., Rio-Belver, R. M., & Ruiz de la Torre, A. (2023). Is the Impact of Management Research Predictable Through the Title?-A BERT Model to Find a Response. *International Conference on Industrial Engineering and Industrial Management (ICIEIM)-Congreso de Ingeniería de Organización*, 379–384.
- Jacobson, A., MC & Traill. (1986). Assessment of speech intelligibility in five South Eastern Bantu languages: Critical considerations. *South African Journal of Communication Disorders*, 33(1), 15–27.
- Jamtsho, Y., & Muneesawang, P. (2020). Dzongkha word segmentation using deep learning. *2020 12th International Conference on Knowledge and Smart Technology (KST)*, 1–5.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678.
- Jin, Z., & Tanaka-Ishii, K. (2006). Unsupervised segmentation of Chinese text by use of branching entropy. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 428–435.
- Johnston, T., & Schembri, A. C. (1999). On defining lexeme in a signed language. *Sign Language & Linguistics*, 2(2), 115–185.
- Jones, K. S., & Galliers, J. R. (1995). Evaluating natural language processing systems: An analysis and review.
- Joseph, J., & Anto, B. (2015). Rule-based morphological analyzer for Malayalam nouns: computational analysis of Malayalam linguistics. *Int. J. Innovative Res. Comput. Commun. Eng.*(2015, October).

- Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd) [Online manuscript released January 12, 2025]. <https://web.stanford.edu/~jurafsky/slp3/>
- Kag, A., & Saligrama, V. (2021). Training recurrent neural networks via forward propagation through time. *International Conference on Machine Learning*, 5189–5200.
- Kairakbay, B. M., & Zaurbekov, D. L. (2013). Finite state approach to the Kazakh nominal paradigm. *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, 108–112.
- Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1700–1709.
- Kambarami, F., McLachlan, S., Bozic, B., Dube, K., & Chimhundu, H. (2021). Computational Modeling of Agglutinative Languages: The Challenge for Southern Bantu Languages. *Arusha Working Papers in African Linguistics*, 3(1), 52–81. https://arushalinguistics.org/publications/Kambarami_et_al_AWPAL_2021.pdf
- Karttunen, L. (1993). *Finite-state lexicon compiler*. Xerox Corporation, Palo Alto Research Center.
- Karttunen, L., & Beesley, K. R. (1992). *Two-level rule compiler*. Xerox Corporation, Palo Alto Research Center.
- Karttunen, L., Koskeniemi, K., Kaplan, R. M., *et al.* (1987). A compiler for two-level phonological rules. *Tools for morphological analysis*, 87, 108.
- Keet, C. M., & Khumalo, L. (2014). Toward verbalizing ontologies in isiZulu. *International Workshop on Controlled Natural Language*, 78–89.
- Khalidi, K. (2017). Quantitative, qualitative or mixed research: Which research paradigm to use. *Journal of Educational and Social Research*, 7(2), 15–24.
- Khan, T., & Kadenge, M. (2016). *IsiZulu adoptives from English and Afrikaans: an optimality theory analysis* [Doctoral dissertation, University of the Witwatersrand, Faculty of Humanities, School of Social Science].
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kituku, B. (2022). *An Approach to Bootstrapping the Development of Multilingual Rule-Based Grammars for Underresourced Languages Using Cross-Linguistic Similarities: A Case Study of a Sub-Set of Kenyan Bantu Languages* [Doctoral dissertation, University of Nairobi]. <https://scholar.google.com/citations?hl=en&user=BHYL5BEAAAAJ>
- Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education*, 6(5), 26–41.
- Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

- Kohonen, O., Virpioja, S., & Klami, M. (2009). Allomorfessor: Towards unsupervised morpheme analysis. *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 9*, 975–982.
- Kohonen, O., Virpioja, S., & Lagus, K. (2010). Semi-supervised learning of concatenative morphology. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, 78–86.
- Kornuta, H. M., & Germaine, R. W. (2019). *A Concise Guide to Writing a Thesis or Dissertation: Educational Research and Beyond* (2nd). Routledge. <https://www.routledge.com/A-Concise-Guide-to-Writing-a-Thesis-or-Dissertation-Educational-Research-and-Beyond/Kornuta-Germaine/p/book/9780367174583>
- Kotzé, G., & Wolff, F. (2014). Experiments with syllable-based English-Zulu alignment. *Proceedings of the SaLTMiL Workshop on Free/open-source Language Resources for the Machine Translation of Less-resourced Languages (at LREC 2014)*, 7–11.
- Kotzé, G., & Wolff, F. (2015). Syllabification and parameter optimisation in Zulu to English machine translation. *South African Computer Journal*, (57).
- Krámský, J. (1969). *The Word as a Linguistic Unit* (Vol. 75). Walter de Gruyter. <https://doi.org/10.1515/9783111353609>
- Krishna, A., Aich, A., Hegde, C., *et al.* (2018). Analysis of customer opinion using machine learning and NLP techniques. *International Journal of Advanced Studies of Scientific Research*, 3(9).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Kubeka, I. S. (1979). *A preliminary survey of Zulu dialects in Natal and Zululand*. [Doctoral dissertation].
- Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer speech & language*, 6(3), 225–242.
- Kurimo, M., Virpioja, S., Turunen, V., & Lagus, K. (2010). Morpho challenge 2005-2010: Evaluations and results. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, 87–95.
- Kurimo, M., Virpioja, S., & Turunen, V. T. (2010). Proceedings of the morpho challenge 2010 workshop.
- Kurimo, M., Virpioja, S., Turunen, V. T., Blackwood, G. W., & Byrne, W. (2010). Overview and results of Morpho Challenge 2009. *Multilingual Information Access Evaluation I. Text Retrieval Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30-October 2, 2009, Revised Selected Papers 10*, 578–597.
- Kurimo, M., Creutz, M. J. P., Varjokallio, M., Arisoy, E., & Saraçlar, M. (2006). Unsupervised segmentation of words into morphemes-Challenge 2005, An Introduction and

- Evaluation Report. *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*.
- Kurimo, M., Creutz, M. J. P., & Turunen, V. (2007). Unsupervised Morpheme Analysis Evaluation by IR Experiments@ Morpho Challenge 2007. *CLEF 2007 Workshop*.
- Kurimo, M., Turunen, V., & Varjokallio, M. (2009). Overview of morpho challenge 2008. *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 9*, 951–966.
- Lafferty, J., McCallum, A., Pereira, F., *et al.* (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, 1(2), 3.
- Lawani, A. (2021). Critical realism: what you should know and how to apply it. *Qualitative research journal*, 21(3), 320–333.
- Le, N. T., Cadotte, A., Boivin, M., Sadat, F., & Terraza, J. (2022). Deep learning-based morphological segmentation for indigenous languages: A study case on innu-aimun. *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 146–151.
- Leben, W. R. (2018). Languages of the World. In *Oxford research encyclopedia of linguistics*.
- Lee, J., Cho, K., & Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5, 365–378.
- Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3, 71–90.
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674.
- Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., & Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.
- Lienen, J., & Hüllermeier, E. (2021). From label smoothing to label relaxation. *Proceedings of the AAAI conference on artificial intelligence*, 35(10), 8583–8591.
- Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1), 503–528.
- Liu, L., Ryan, Z., & Hulden, M. (2021). The usefulness of Bibles in low-resource machine translation. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1, 44–50.
- Liu, N., Su, X., Zhang, H., Gao, G., & Bao, F. (2020). Incorporating inner-word and out-word features for Mongolian morphological segmentation. *Proceedings of the 28th International Conference on Computational Linguistics*, 4638–4648.
- Liu, Z., & Prud’Hommeaux, E. (2022). Data-driven Model Generalizability in Crosslinguistic Low-resource Morphological Segmentation. *Transactions of the Association for Computational Linguistics*, 10, 393–413.

- Loubser, M., & Puttkammer, M. J. (2020). Viability of neural networks for core technologies for resource-scarce languages. *Information*, 11(1), 41.
- Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.
- Luyt, K., et al. (2012). *A managerial framework for the enablement of the performance of virtual knowledge workers* [Doctoral dissertation, University of Pretoria].
- Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge University Press.
- Mafuwane, B. M., et al. (2011). *The contribution of instructional leadership to learner performance* [Doctoral dissertation, University of Pretoria].
- Mager, M., Çetinoğlu, Ö., & Kann, K. (2020). Tackling the low-resource challenge for canonical segmentation. *arXiv preprint arXiv:2010.02804*.
- Mager, M., Oncevay, A., Mager, E., Kann, K., & Vu, N. T. (2022). BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. *arXiv preprint arXiv:2203.08954*.
- Magistry, P., & Sagot, B. (2013). Can MDL Improve Unsupervised Chinese Word Segmentation? *Sixth International Joint Conference on Natural Language Processing: Sighan workshop*, 2.
- Mahabadi, R. K., Ruder, S., Dehghani, M., & Henderson, J. (2021). Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*.
- Maho, J. (2003). A classification of the Bantu languages: An update of Guthrie's referential system. *The Bantu languages*, 639–651.
- Makedonski, P. (2005). Finite state morphology: the Turkish nominal paradigm. *Bachelor thesis, Eberhard-Karls University*.
- Manicas, P. T., & Kruger, A. N. (1976). *Logic: The Essentials*. McGraw-Hill.
- Manners, I. (2020). Critical social theory approaches to European integration. In *The Routledge handbook of critical European studies* (pp. 139–152). Routledge.
- Marais, L., & Pretorius, L. (2021). Exploiting a multilingual semantic Machine translation architecture for knowledge representation of patient information for Covid-19. *Proceedings of the 1st South African Conference for Artificial Intelligence Research*.
- Marais, L., & Pretorius, L. (2023). Parsing IsiZulu text using grammatical framework. *International Symposium on Distributed Computing and Artificial Intelligence*, 167–177.
- Marais, L., Pretorius, L., & Posthumus, L. C. (2024). Bootstrapping syntactic resources from isiZulu to Siswati. *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, 77–85.
- Marks, S. (1967). The Nguni, the Natalians, and their history. *The Journal of African History*, 8(3), 529–540.
- Meinhof, C., Van Warmelo, N. J., & Werner, A. (1932). Introduction to the phonology of the Bantu languages. (*No Title*).

- Mesham, S., Hayward, L., Shapiro, J., & Buys, J. (2021). Low-resource language modelling of South African languages. *arXiv preprint arXiv:2104.00772*.
- Meyer, F., & Buys, J. (2022). Subword Segmental Language Modelling for Nguni Languages. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6636–6649. <https://doi.org/10.18653/v1/2022.findings-emnlp.494>
- Mikolov, T., *et al.* (2012). Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April, 80*(26).
- Mikolov, T., Deoras, A., Povey, D., Burget, L., & Černocký, J. (2011). Strategies for training large scale neural network language models. *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 196–201.
- Mkhwanazi, S., & Marais, L. (2024). Generation of segmented isiZulu text. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).
- Mockus, J. (1989). Global optimization and the Bayesian approach. In *Bayesian Approach to Global Optimization: Theory and Applications* (pp. 1–3). Springer.
- Moeng, T., Reay, S., Daniels, A., & Buys, J. (2022). Canonical and surface morphological segmentation for Nguni languages. *Artificial Intelligence Research: Second Southern African Conference, SACAIR 2021, Durban, South Africa, December 6–10, 2021, Proceedings*, 125–139.
- Montavon, G., Orr, G., & Müller, K.-r. (2012). Neural Networks: Tricks of the Trade. *Lecture Notes in Computer Science*. <https://api.semanticscholar.org/CorpusID:39578794>
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109–139). Springer.
- Morgan, N., & Bourlard, H. (1989). Generalization and parameter estimation in feedforward nets: Some experiments. *Advances in neural information processing systems*, 2.
- Moukangwe, K., Malumba, N., & Suleman, H. (2014). HONOURS PROJECT REPORT.
- Murphy, K. P. (2012a). *Machine learning: a probabilistic perspective*. MIT press.
- Murphy, K. P. (2012b). *Machine learning: a probabilistic perspective*. MIT press.
- Myers, M. (2002). *Qualitative Research in Information Systems: A Reader*. Sage.
- Mzamo, L., Helberg, A., & Bosch, S. (2020). Evaluation of heuristic guided character level word models for morphological segmentation of isiXhosa [[Online]. Available: <https://tinyurl.com/2xybvmda>]. *Proceedings of the SACAIR Conference on Applied AI*, 162–178.
- Mzamo, L. (2021). *Morphological segmentation of isiXhosa using unsupervised machine learning* [Doctoral dissertation, North-West University (South Africa)].
- Mzamo, L., Helberg, A., & Bosch, S. (2019a). Towards an unsupervised morphological segmenter for isiXhosa. *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 166–170.

- Mzamo, L., Helberg, A., & Bosch, S. (2019b). Towards an unsupervised morphological segmenter for isiXhosa. *Proceedings - 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa, SAUPEC/RobMech/PRASA 2019*, 166–170. <https://doi.org/10.1109/RoboMech.2019.8704816>
- Naidoo, Y., Van der Merwe, A., Groenewald, E., & Naudé, E. (2005). Development of speech sounds and syllable structure of words in Zulu-speaking children. *Southern African Linguistics and Applied Language Studies*, 23(1), 59–79.
- Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814.
- Narkhede, M. V., Bartakke, P. P., & Sutaone, M. S. (2022). A review on weight initialization strategies for neural networks. *Artificial Intelligence Review*, 55(1), 291–322.
- Nguyen, V. D. (2018). Natural Language Generation From Ontologies Using Grammatical Framework. *Technical Communications of the 34th International Conference on Logic Programming (ICLP 2018)*.
- Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452, 48–62.
- Nkosi, K. (2017). Analysis of segmentation analysis decision metrics for isiZulu [July 2017, 7 pages]. *Proceedings of ACM Conference*.
- NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., ... Wang, J. (2024). Scaling neural machine translation to 200 languages. *Nature*, 630(8018), 841–846. <https://doi.org/10.1038/s41586-024-07335-x>
- Nurse, D., & Philippson, G. (2006). Zone S Derek Gowlett. In *The Bantu Languages* (pp. 627–656). Routledge.
- Occhipinti, L. (2024, December). Enhancing Lexical Complexity Prediction in Italian through Automatic Morphological Segmentation. In F. Dell’Orletta, A. Lenci, S. Montemagni, & R. Sprugnoli (Eds.), *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)* (pp. 670–678). CEUR Workshop Proceedings. <https://aclanthology.org/2024.clicit-1.75/>
- Oflazer, K. (2003). Lenient morphological analysis. *Natural language engineering*, 9(1), 87–99.
- Önal, E., et al. (2021). MorphLaz:— a finite-state morphological analyzer for Laz.
- Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International journal of social research methodology*, 8(5), 375–387.
- Pan, Y., Li, X., Yang, Y., & Dong, R. (2020). Morphological word segmentation on agglutinative languages for neural machine translation. *arXiv preprint arXiv:2001.01589*.

- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., & Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9, 261–276.
- Park, Y. S., Konge, L., & Artino Jr, A. R. (2020). The positivism paradigm of research. *Academic Medicine*, 95(5), 690–694.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., *et al.* (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Paul, G. (1993). Approaches to abductive reasoning: an overview. *Artificial intelligence review*, 7(2), 109–152.
- Pereira, F. C., & Warren, D. H. (1980). Definite clause grammars for language analysis—a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, 13(3), 231–278.
- Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation*, 57(3), 330–348.
- Poon, H., & Domingos, P. (2008). Joint unsupervised coreference resolution with Markov logic. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language processing*, 650–659.
- Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of the tenth workshop on statistical machine translation*, 392–395.
- Popper, K. (2002). *The Logic of Scientific Discovery*. Routledge.
- Poulos, G., & Msimang, C. (1998). *A Linguistic Analysis of Zulu*. Via Afrika.
- Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.
- Pranjić, M., Robnik-Šikonja, M., & Pollak, S. (2024). LLMSegm: Surface-level Morphological Segmentation Using Large Language Model. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10665–10674.
- Premjith, B., Soman, K., & Kumar, M. A. (2018). A deep learning approach for Malayalam morphological analysis at character level. *Procedia computer science*, 132, 47–54.
- Pretorius, L., & Bosch, S. E. (2003). Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation*, 18, 195–216.
- Pretorius, L., Viljoen, B., Berg, A., & Pretorius, R. (2015). Tswana finite state tokenisation. *Language Resources and Evaluation*, 49, 831–856.
- Pretorius, L. (2024). Demystifying research paradigms: Navigating ontology, epistemology, and axiology in research. *The Qualitative Report*, 29(10), 2698–2715.

- Puttkammer, M., & Du Toit, J. S. (2021). Canonical Segmentation and Syntactic Morpheme Tagging of Four Resource-scarce Nguni Languages. *Journal of the Digital Humanities Association of Southern Africa*, 3(03).
- Quasthoff, U., Goldhahn, D., & Bosch, S. (2016). Morphology Learning for Zulu. *CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity*, 89.
- Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1), 4–16.
- Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. <https://api.semanticscholar.org/CorpusID:49313245>
- Rahali, A., & Akhloufi, M. A. (2023). End-to-end transformer-based models in textual-based NLP. *AI*, 4(1), 54–110.
- Rana, J., Gutierrez, P. L., & Oldroyd, J. C. (2023). Quantitative methods. In *Global Encyclopedia of Public Administration, Public Policy, and Governance* (pp. 11202–11207). Springer.
- Ranta, A. (2004). Grammatical framework. *Journal of Functional Programming*, 14(2), 145–189.
- Ranta, A. (2007). Modular grammar engineering in GF. *Research on Language and Computation*, 5(2), 133–158.
- Ranta, A. (2009). The GF resource grammar library. *Linguistic Issues in Language Technology*, 2.
- Ranta, A., Tian, Y., & Qiao, H. (2015). Chinese in the Grammatical Framework: Grammar, Translation, and Other Applications Anonymous. *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, 100–109.
- Ratnaparkhi, A. (1997). *A Simple Introduction to Maximum Entropy Models for Natural Language Processing* (tech. rep. No. IRCS-97-08). Institute for Research in Cognitive Science, University of Pennsylvania. https://repository.upenn.edu/ircs_reports/81/
- Resnik, P., & Lin, J. (2010). Evaluation of NLP Systems. In A. Clark, C. Fox, & S. Lappin (Eds.), *The Handbook of Computational Linguistics and Natural Language Processing* (pp. 271–295). Wiley-Blackwell.
- Rice, E., Marashian, A., Gessler, L., Palmer, A., & von der Wense, K. (2024). TAMS: Translation-Assisted Morphological Segmentation. *arXiv preprint arXiv:2403.14840*.
- Richardson, I., & Tyers, F. M. (2021). A morphological analyser for K'iche'. *Procesamiento del Lenguaje Natural*, 66, 99–109.
- Ringberg, T., & Reihlen, M. (2008). Towards a socio-cognitive approach to knowledge transfer. *Journal of Management Studies*, 45(5), 912–935.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.
- Rodrigues, F., Pereira, F., & Ribeiro, B. (2014). Sequence labeling with multiple annotators. *Machine learning*, 95, 165–181.

- Rouhe, A., Grönroos, S.-A., Virpioja, S., Creutz, M., & Kurimo, M. (2022). Morfessor-enriched features and multilingual training for canonical morphological segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 144–151.
- Rumelhart, D. E., Hinton, G. E., Williams, R. J., *et al.* (1985). Learning internal representations by error propagation.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.
- Ruokolainen, T., Kohonen, O., Virpioja, S., & Kurimo, M. (2013). Supervised morphological segmentation in a low-resource learning setting using conditional random fields. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 29–37.
- Ruokolainen, T., Kohonen, O., Virpioja, S., & Kurimo, M. (2014). Painless Semi-Supervised Morphological Segmentation using Conditional Random Fields. *EACL 2014 - 14th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 84–89. <https://doi.org/10.3115/v1/e14-4017>
- Ruokolainen, T., Kohonen, O., Sirts, K., Grönroos, S.-A., Kurimo, M., & Virpioja, S. (2016). A comparative study of minimally supervised morphological segmentation. *Computational Linguistics*, 42(1), 91–120.
- Ryan, G. (2018). Introduction to positivism, interpretivism and critical theory. *Nurse Researcher*, 25(4), 41–49.
- Salameh, M., Cherry, C., & Kondrak, G. (2015). What Matters Most in Morphologically Segmented SMT Models? *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 65–73.
- Salesky, E., Runge, A., Coda, A., Niehues, J., & Neubig, G. (2020). Optimizing segmentation granularity for neural machine translation. *Machine Translation*, 34(1), 41–59.
- Sarikaya, R., Afify, M., & Gao, Y. (2007). Joint morphological-lexical language modeling (JMLLM) for Arabic. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 4, IV–181.
- Saunders, M. (2009). Research methods for business students. *Person Education Limited*.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673–2681.
- Schwenk, H. (2012). Continuous space translation models for phrase-based statistical machine translation. *Proceedings of COLING 2012: Posters*, 1071–1080.
- Scotland, J. (2012). Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms. *English language teaching*, 5(9), 9–16.
- Sekaran, U., & Bougie, R. (2016). *Research Methods for Business: A Skill-Building Approach* (7th). Wiley.

- Selvam, M., & Natarajan, A. (2009). Improvement of rule based morphological analysis and pos tagging in Tamil language via projection and induction techniques. *International Journal of Computers*, 3(4), 357–367.
- Sennrich, R. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Shalonova, K., & Golénia, B. (2010). Weakly Supervised Morphology Learning for Agglutinating Languages Using Small Training Sets. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 976–983.
- Shapiro, P., & Duh, K. (2018). BPE and charcnns for translation of morphology: A cross-lingual comparison and analysis. *arXiv preprint arXiv:1809.01301*.
- Sharma, S., Diwakar, M., Singh, P., Singh, V., Kadry, S., & Kim, J. (2023). Machine translation systems based on classical-statistical-deep-learning approaches. *Electronics*, 12(7), 1716.
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306.
- Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data*, 8(1), 101.
- Singh, P. (2018). *Machine learning with PySpark: with natural language processing and recommender systems*. Apress.
- Smith, L. N. (2017). Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 464–472.
- Sobh, R., & Perry, C. (2006). Research design and data analysis in realism research. *European Journal of Marketing*, 40(11/12), 1194–1209.
- Spiegler, S., Golénia, B., & Flach, P. A. (2010). DEAP: Deductive-abductive parsing for morphological analysis. *Proceedings of the Morpho Challenge 2010 Workshop, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland*, 44–48.
- Spiegler, S., & Monson, C. (2010). Emma: A novel evaluation metric for morphological analysis. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 1029–1037.
- Spiegler, S., Van Der Spuy, A., & Flach, P. A. (2010). Ukwabelana-an open-source morphological Zulu corpus. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 1020–1028.
- Spiegler, S., Golénia, B., Shalonova, K., Flach, P., & Tucker, R. (2008). Learning the morphology of Zulu with different degrees of supervision. *2008 IEEE Spoken Language Technology Workshop*, 9–12.
- Spiegler, S. R. (2011). *Machine learning for the analysis of morphologically complex languages* [Doctoral dissertation, University of Bristol].
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of Machine Learning Research*, 15(1), 1929–1958.

- Srivastava, S., Shukla, A., & Tiwari, R. (2018). Machine translation: from statistical to modern deep-learning practices. *arXiv preprint arXiv:1812.04238*.
- StatsSA. (2022a). Statistics South Africa.
- StatsSA. (2022b). Statistics south africa. *Formal census*.
- Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Comput. Linguist.*, 21(2), 165–201.
- Sudoh, K. (2015). A Japanese-to-English Statistical Machine Translation System for Technical Documents.
- Sun, W., & Xu, J. (2011). Enhancing Chinese word segmentation using unlabeled data. *Proceedings of the 2011 Conference on Empirical Methods in natural language Processing*, 970–979.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 3104–3112.
- Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4), 267–373.
- Swart, P. H. (2000). *Prosodic features of imperatives in Xhosa: implications for a text-to-speech system* [Doctoral dissertation, Stellenbosch: Stellenbosch University].
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.
- Taljaard, P., & Snyman, J. (1989). *An Introduction to Zulu Phonetics*. M. Lubbe. <https://books.google.co.za/books?id=WnAHAQAIAAJ>
- Taljar, E., & Bosch, S. E. (2006). A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies*, 15(4).
- Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1, 5–21.
- Tariq, M. U. (2015). Hypothetico-deductive method: A comparative analysis. *Journal of Basic and Applied Research International*, 7(4), 228–231.
- Taye, M. M. (2023). Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5), 91.
- Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., *et al.* (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*.
- Tenny, S., Brannan, J. M., Brannan, G. D., & Sharts-Hopko, N. C. (2017). Qualitative Study. *StatPearls*. <https://www.ncbi.nlm.nih.gov/books/NBK470395/>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.
- Trochim, W. M. K., Donnelly, J. P., & Arora, K. (2016). *Research Methods: The Essential Knowledge Base* (2nd). Cengage Learning.

- Tukeyev, U., Karibayeva, A., & Zhumanov, Z. h. (2020). Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, 7(1), 1856500.
- Tukur, A., Umar, K., & Muhammad, S. (2020). Parts-of-speech tagging of Hausa-based texts using hidden Markov model. *Dutse Journal of Pure and Applied Sciences*, 6, 303–13.
- UNESCO. (2000). *World Languages Report*. United Nations Educational, Scientific; Cultural Organization.
- UNESCO. (2021). Proclamation of World Kiswahili Language Day [Proclaimed at the 41st session of the General Conference]. <https://www.unesco.org/en/days/kiswahili-language>
- Van de Velde, M., Bostoen, K., Nurse, D., & Philippson, G. (2019). *The Bantu Languages*. Routledge.
- Van der Spuy, A. (2001). In defence of the Zulu adjective. *South African Journal of African Languages*, 21(2), 148–162.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information processing systems*, 30.
- Villegas-Ch, W., Gutierrez, R., Navarro, A. M., & Mera-Navarrete, A. (2024). Evaluating Neural Network Models for Word Segmentation in Agglutinative Languages: Comparison with Rule-Based Approaches and Statistical Models. *IEEE Access*.
- Virpioja, S., Kohonen, O., & Lagus, K. (2009). Unsupervised morpheme analysis with Allomorfessor. *Workshop of the Cross-Language Evaluation Forum for European Languages*, 609–616.
- Virpioja, S., Turunen, V. T., Spiegler, S., Kohonen, O., & Kurimo, M. (2011). Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2), 45–90.
- Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., *et al.* (2013). Morfessor 2.0: Python implementation and extensions for Morfessor Baseline.
- Vishwanathan, S., Schraudolph, N. N., Schmidt, M. W., & Murphy, K. P. (2006). Accelerated training of conditional random fields with stochastic gradient methods. *Proceedings of the 23rd International Conference on Machine learning*, 969–976.
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2), 260–269.
- Vuong, Q. (2021). Machine Learning for Robotic Manipulation. *arXiv preprint arXiv:2101.00755*. <https://arxiv.org/abs/2101.00755>
- Wahyuni, D. (2012). The research design maze: Understanding paradigms, cases, methods and methodologies. *Journal of Applied Management Accounting Research*, 10(1), 69–80.
- Wang, C., Cho, K., & Gu, J. (2020). Neural machine translation with byte-level subwords. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 9154–9160.

- Wang, J. (2019). *Formal methods in computer science*. Chapman; Hall/CRC.
- Wang, L., Cao, Z., Xia, Y., & De Melo, G. (2016a). Morphological segmentation with window LSTM neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).
- Wang, L., Cao, Z., Xia, Y., & De Melo, G. (2016b). Morphological Segmentation with Window LSTM Neural Networks. *AAAI*, 30(1). <https://doi.org/10.1609/aaai.v30i1.10363>
- Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10), 1550–1560.
- Westhelle, M., Bencke, L., & Moreira, V. P. (2022). Impact of Morphological Segmentation on Pre-trained Language Models. *Brazilian Conference on Intelligent Systems*, 402–416.
- Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, 14(4), 490–495.
- Wilkes, A. (1985). Words and word division: a study of some orthographical problems in the writing systems of the Nguni and Sotho languages. *South African Journal of African Languages*, 5(4), 148–153.
- Ye, X., Wu, Y., Zhang, W., Li, X., Chen, Y., & Jin, C. (2024). Optimized Gradient Clipping for Noisy Label Learning. *arXiv preprint arXiv:2412.08941*.
- Yong, W. K., Husin, M. M., & Kamarudin, S. (2021). Understanding research paradigms: A scientific guide. *The Journal of Contemporary Issues in Business and Government*, 27(2), 5857–5865.
- Yu, B., & Fan, Z. (2020). A comprehensive review of conditional random fields: variants, hybrids and applications. *Artificial Intelligence Review*, 53(6), 4289–4333.
- Yu, T., & Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*.
- Yusupujiang, Z. (2018). *Using Unsupervised Morphological Segmentation to Improve Dependency Parsing for Morphologically Rich Languages* [Master's thesis, Uppsala University]. <https://www.diva-portal.org/smash/get/diva2:1221345/FULLTEXT01.pdf>
- Zbib, R., Kayser, M., Matsoukas, S., Makhoul, J., Nader, H., Soliman, H., & Safadi, R. (2012). Methods for integrating rule-based and statistical systems for Arabic to English machine translation. *Machine Translation*, 26, 67–83.
- Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*.
- Zhang, T. (2023). Critical realism: A critical evaluation. *Social Epistemology*, 37(1), 15–29.
- Žukauskas, P., Vveinhardt, J., & Andriukaitienė, R. (2018). Philosophy and paradigm of scientific research. *Management Culture and Corporate Social Responsibility*, 121(13), 506–518.

- Zungu, E. M. (1999). *A Comparative Phonological and Morphological Analysis of the North and South Lala Dialects of Tekela Nguni* [Doctoral dissertation, University of South Africa]. <http://hdl.handle.net/10500/17881>

Appendix A

Appendix

A.1 Models Hyper-Parameters

A.1.1 Hyper-Parameters for Unaugmented Models

Table A.1: Model configurations for The Unaugmented Segmentation Models

Parameter	Segmenter One	Segmenter Two	Segmenter Three
Learning Rate	1.861×10^{-4}	2.689×10^{-5}	8.865×10^{-4}
Model Dimension (d)	1024	1024	256
Number of Layers	3	3	4
Number of Attention Heads	4	4	4
Feed-Forward Dimension (d_{ff})	1024	1024	2048
Dropout Rate	0.113	0.103	0.223
Maximum Gradient Norm	1.184	1.483	1.043
Label Smoothing	3.987×10^{-2}	2.306×10^{-2}	2.106×10^{-2}

A.1.2 Hyper-Parameters for English-to-IsiZulu Translation Models

Table A.2: Model Configurations for English-to-IsiZulu Translation With and Without Segmentation

Parameter	No Segmentation	Segmenter One	Segmenter Two	Segmenter Three
Learning Rate	2.061×10^{-4}	2.061×10^{-4}	6.359×10^{-4}	1.331×10^{-4}
Sequence Length	100	100	300	250
Model Dimension (d)	256	256	128	512
Number of Layers	4	4	3	3
Number of Attention Heads	8	8	8	4
Feed-Forward Dimension (d_{ff})	2048	2048	512	512
Dropout Rate	0.155	0.155	0.201	0.168
Maximum Gradient Norm	1.024	1.024	0.862	1.118
Label Smoothing	6.176×10^{-2}	6.176×10^{-2}	4.842×10^{-2}	4.910×10^{-2}
Vocabulary Size	30,000	30,000	30,000	30,000

Declaration

This is to declare that I, Dr L Combrink

Language editor and translator

have language edited the dissertation

with the title

Investigating supervised surface segmentation of isiZulu text using synthetic data generation

by

S. N. Mkhwanazi



orcid.org/0009-0003-3050-4744

A handwritten signature in blue ink that reads 'L. Combrink.'.

Date: March 2025