information

MDPI

*Article*

# Developing Core Technologies for Resource-Scarce Nguni Languages

**Jakobus S. du Toit and Martin J. Puttkammer ***

Centre for Text Technology, North-West University, Potchefstroom 2520, South Africa; Jaco.DuToit@nwu.ac.za
* Correspondence: Martin.Puttkammer@nwu.ac.za; Tel.: +27-82-495-0609

**Abstract:** The creation of linguistic resources is crucial to the continued growth of research and development efforts in the field of natural language processing, especially for resource-scarce languages. In this paper, we describe the curation and annotation of corpora and the development of multiple linguistic technologies for four official South African languages, namely isiNdebele, Siswati, isiXhosa, and isiZulu. Development efforts included sourcing parallel data for these languages and annotating each on token, orthographic, morphological, and morphosyntactic levels. These sets were in turn used to create and evaluate three core technologies, viz. a lemmatizer, part-of-speech tagger, morphological analyzer for each of the languages. We report on the quality of these technologies which improve on previously developed rule-based technologies as part of a similar initiative in 2013. These resources are made publicly accessible through a local resource agency with the intention of fostering further development of both resources and technologies that may benefit the NLP industry in South Africa.

**Keywords:** resource-scarce languages; South African languages; core technologies; part-of-speech tagging; lemmatization; canonical segmentation; morphological analysis



**Citation:** du Toit, J.S.; Puttkammer, M.J. Developing Core Technologies for Resource-Scarce Nguni Languages. *Information* **2021**, *12*, 520. https://doi.org/10.3390/info12120520

Academic Editor: Paulo Quaresma

Received: 2 November 2021
Accepted: 11 December 2021
Published: 14 December 2021



## 1. Introduction

Access to linguistic resources such as annotated data helps to facilitate, or even hinder, research and development efforts based on its quality and availability. Central to these efforts is the notion of lexical semantics, generally defined as the analysis of words and lexical units in terms of their classification, decomposition, and their lexical meaning in relationship to context and syntax. To date, lexical semantic knowledge has been captured through either a knowledge-based approach, where linguistic knowledge is directly recorded in a structured and often rule-based form, and corpus-based approaches where machine-learned semantic knowledge is gained from corpora and represented implicitly [1]. Contemporary research relies on natural language processing (NLP) to investigate usage patterns within large electronic corpora to achieve lexical semantic tasks such as word sense disambiguation and semantic role labelling [2]. In turn, NLP applications rely on these tasks to perform machine translation, information extraction, text classification, among other tasks. For under-resourced languages, this approach suffers due to the scarcity and often lacking quality of available lexical data [3].

Based on their orthographies, the 11 official South African languages are all either Southern Bantu languages or West-Germanic languages, and can be categorized on a conjunctive-disjunctive scale into three groups, i.e., conjunctive languages (four Nguni languages, viz. isiZulu, isiXhosa, isiNdebele, and Siswati), disjunctive languages (Tshivenḓa, Xitsonga, and three Sotho-Tswana languages (Sesotho, Sepedi, and Setswana)), and the middle of the scale, two West-Germanic languages, viz. Afrikaans and English [4]. All 10 the indigenous official languages (English excluded) are considered resource-scarce and attempting to generate lexical resources can be a protracted endeavor. In light of this, the South African government has established several legislative frameworks to help promote

both the use and advancement of its official languages [5]. These measures promote the development of human language technologies (HLT) and as result helps to preserve the cultural significance of each language whilst contributing to the tools and resources that serve its community of language researchers. This study in particular serves as an extension to a previously sponsored NCHLT Text project completed in 2013 [5] that annotated data and developed the associated core technologies for all 10 indigenous South African languages (English was excluded since text resources for English are readily available).

In this follow-up project, we aim at improving on the previously developed rule-based technologies from 2013 and at expanding on the morphological decomposers to also include morphological analysis (see Section 3.3 for a comparison of decomposition vs. analysis) for the four conjunctive languages. The NCHLT project concluded with core technologies for the conjunctive languages that underperformed when compared to those of the disjunctively written languages. This was attributed to the greater morphological complexity of conjunctive languages and the necessity for more data to offset the sparsity of morphological phenomena. The four Nguni languages all exhibit a similar agglutinative morphology that is evident in its conjunctive orthography which allows for the joint development of linguistic resources.

This follow-up project was intended to expand on the initial dataset with an additional 50,000 tokens and had the added goal of introducing morphological analysis for these languages. However, in the time since the NCHLT Text project [5], the protocols and tag sets have been amended and refined and the time-intensive task of amalgamating the two datasets was instead postponed for a later date. Moreover, developments in the NLP field have made the before-mentioned tasks all the more feasible using limited datasets through recent advances that are proven to outperform rule-based approaches. It is for these reasons that new solutions were researched and experimentally applied to the newly generated dataset for the four Nguni languages to potentially produce better performing and more reliable core technologies for these languages.

Lemmatization previously derived lemmata through rule-based normalization for these languages which was susceptible to inaccurate stem identification and consequently unreliable lemmatization. Similarly, the rule-based approach to morpheme segmentation relied on recursive affix identification in a token before verifying these against a lexicon of valid affix combinations and valid roots and stems. Not only do these approaches require expert knowledge to maintain and expand on the rules, but they are also limited to known or expected instances and are thus unable to process or properly analyze any morphological patterns that are not explicitly defined in these rules. Machine learning-based approaches provide for greater flexibility and potentially improved precision when applied to these same tasks. Additionally, POS tagging was performed using a then state-of-the-art trainable POS tagging solution, HunPoS, which relied upon the proficient capabilities of HMM that have since been superseded by novel approaches and algorithms applied to the same task [6,7].

Apart from the writing system, other characteristics of these languages make them complex to deal with computationally. Not only are they tone languages, but they use an elaborate noun classification system of up to 21 classes, and the verbal morphology is highly agglutinative and productive, resulting in a large and ever-growing vocabulary [8]. These factors and the limited availability of quality linguistic resources contribute to the limited NLP research related to these languages and the proposed technologies. Yet, the work performed by Bosch et al. [9] toward the development of morphological analyzers for a subset of South African languages guided how the orthographical and morphological challenges were addressed in the first NCHLT Text project. Their work modelled complex word-formation and morphological alternations for disjunctively written languages and provided insight into how these could be computationally formulated which helped guide establishing the necessary protocols and annotation guidelines used in this project.

The remainder of this paper is organized as follows. The next section provides a brief overview of the lexical resources that were developed as part of the project according to

stipulated guidelines and protocols. Section 3 then describes the development of core technologies with reference to the annotation itself, the adopted technologies, and how the data served in the training of the related models. Evaluation results for each of the core technologies are presented in Section 4. Finally, the paper concludes in Section 5 by acknowledging the valuable gains in performance that machine learning approaches provide over rule-based approaches and considering how linguistic resources are beneficial to the South African language community.

## 2. Data Resources

The goal of the project is two-fold: firstly, to build linguistically annotated datasets for four South African languages and secondly, to develop core technologies based on these generated datasets. The primary focus of this paper is on the development of the derived core technologies, for a detailed description of the data, annotation procedure, tagsets, and protocols, see [10,11]. These resources are intended to benefit corpus linguistic studies and research as well as supporting further development of NLP technologies and applications for NR, SS, XH, and ZU. This section provides an overview of the corpus in terms of the data source, the annotation procedure, and the depth of linguistic information in its annotation. The annotated corpora and tagsets are freely available through the SADiLaR Resource Management Agency (https://repo.sadilar.org/handle/20.500.12185/546, accessed on 10 December 2021).

### *2.1. Corpora*

For the purposes of this study, a collection of unannotated, parallel corpora for the four Nguni selected languages were collected from the South African government domain websites and documents. The textual content of the domain is freely available to the public and is mostly made up of government related topics such as press releases, national address, and government services which are made available in parallel across all official South African languages. Text filtering based on the associated language was performed across the entirety of the collected textual data using an internally developed South African language identifier [12], ensuring only the four relevant languages are included in the dataset. Each of the language datasets contain close to 50,000 tokens.

### *2.2. Protocols*

Each of the four language-specific corpora were annotated for three types of linguistic information, namely lemma, part-of-speech (POS), and morphology. To ensure that consistency was maintained during the annotation process across each of the datasets, we relied upon updated annotation protocols and guidelines that were created as part of the previous NCHLT Text project [5]. These protocols were developed in accordance with existing international standards, mainly the Expert Advisory Group for Language Engineering Standards (EAGLES) [13] and document the procedure and standards for each of the three levels of annotation as well as the set of permissible POS and morphosyntactic tags used for each of the languages.

### *2.3. Annotated Data*

Following the before-mentioned protocols, each annotation level was initially automatically populated to deliver a pre-annotated dataset that could then be presented to linguistic experts for further annotation and correction. This approach helped to improve both the rate and consistency of annotation and made it possible to directly compare the resulting annotations across different languages given the aligned nature of the data. A second measure that ensured consistent and accurate annotations was the use of an in-house developed annotation environment, LARA II (https://repo.sadilar.org/handle/20.500.12185/432, accessed on 10 December 2021), which enables users who have limited or basic computer skills to develop annotated, machine-readable corpora. The tool has shown to increase annotation accuracy while at the same time decreasing annotation time [5]. For the mor-

phological analysis, this allowed the annotator access to a selection of possible analyses for a given token where the correct analysis could then be selected. This is especially useful when dealing with conjunctive languages where a token may have multiple different analyses and up to 12 morpheme split points, each of which are to be identified and assigned the relevant grammatical class. LARA II avoids the need to manually identify and classify each individual morpheme by providing a list of analyses complete with various probable breakpoints and permissible classes, thereby reducing the possibility of human error and fast-tracking the process of annotation.

The lemma and POS of each word were indicated per token in addition to its morphological information which included generic token components such as subject and object concords, roots, and transitivity. This information is represented within the morphological annotation tagset which is applied in the labelling of each morpheme, for example: the [AdjPref] tag indicates an adjective prefix, similarly the [NStem] tag is used to indicate a noun stem. Combining these tags during annotation allows for a complete morphological analysis of a token. Table 1 contains a single sentence as an example of the final annotated data for isiZulu ("The fundraiser needs to raise funds to pay for food parcels."). The data is structured in a four-column text format with each column corresponding to a certain type of information i.e., token, lemma, POS and morphological analysis.

**Table 1.** Example of annotated data for isiZulu.

| Token | Lemma | POS (Full Set) | Morphological Analysis |
|---|---|---|---|
| Umuntu | ntu | N01 | u[NPrePre1]-mu[BPre1]-ntu[NStem] |
| ozosiza | siza | REL | o[RelConc1]-zo[Fut]-siz[VRoot]-a[VerbTerm] |
| ngemali | mali | ADV | nga[AdvPre]-i[NPrePre9]-mali[NStem] |
| kudingeka | dinga | V | ku[SC15]-ding[VRoot]-ek[NeutExt]-a[VerbTerm] |
| aqoqe | qoqa | REL | a[RelConc1]-qoq[VRoot]-e[VerbTerm] |
| izimali | mali | N10 | i[NPrePre10]-zi[BPre10]-mali[NStem] |
| zokukhokhela | khokhela | POSS10 | za[PossConc10]-u[NPrePre15]-ku[BPre15]-khokhel[VRoot]-a[VerbTerm] |
| amaphasela | phasela | N06 | a[NPrePre6]-ma[BPre6]-phasela[NStem] |
| okudla | dla | REL | oku[RelConc15]-dla[NStem] |
| . | . | PUNC | .[Punc] |

## 3. Core Technologies

As part of this project, a set of core technologies associated with each level of annotation was developed based on recent and relevant advances available in the field of HLT. These technologies include lemmatizers, part-of-speech (POS) taggers, and morphological analyzers for each of the four Nguni languages. Only the tokenizers and sentence separators developed as part of the last NCHLT Text project [5] are carried forward without any adaptions since these technologies operate reliably and properly handle abbreviations and contracted forms. As with the annotated data, the core technologies are distributed under the Creative Commons Attribution 4.0 International license (https://creativecommons.org/licenses/by/4.0/, accessed on 10 December 2021) and available from SADiLaR (https://repo.sadilar.org/handle/20.500.12185/548, accessed on 10 December 2021). These technologies benefit language learners and linguists conducting research on the considered Nguni languages and can potentially be implemented in the development of other tools.

### *3.1. Part-of-Speech Taggers*

POS refers to the function that a word fulfils within a sentence (i.e., in a grammatical context) and can also be referred to as the lexical category, word class, or lexical class of a word [14]. The syntactic role of a word in a specific context is what determines its associated POS and can be assigned using a trained tagger. In linguistics, we usually distinguish between open classes (i.e., classes to which new words can be added; usually nouns, verbs

and adverbs), and closed classes (i.e., classes that are generally not expanded through productive processes; e.g., adjectives, pronouns, determiners, etc.) [14,15]. These classes are reflected in the tailored POS tagset used during annotation which is mainly based on the lexical and morphological criteria defined by Taljard et al. [16] and distinguishes 20 main POS categories. The tailored tags consist of up to two elements: an element indicating the word class, and a second specifying functional or syntactic properties. We distinguish between two sets of tags based on these levels where the first is referred to as a simplified tagset that consists of 20 tags and is restricted to only the first element, thereby reducing the complexity of the tags and yielding a smaller target space. This aids generalization during training and helps to obtain a greater POS tagging accuracy. The second set is referred to as the full tagset and consists of 107 possible tags which include the second level of syntactic and functional information. From among the permissible 20 main word classes and 107 unique POS tags, most occurred in the data and their counts are summarized in Table 2:

**Table 2.** POS tag counts per tagset.

| Language | POS Tag Count (Simplified Set) | POS Tag Count (Full Set) |
|---|---|---|
| NR | 16 | 95 |
| SS | 16 | 102 |
| XH | 16 | 105 |
| ZU | 16 | 105 |

Both tagsets are accommodated by training two separate models using the POS component of an existing a morphological tagger, MarMot (http://cistern.cis.lmu.de/marmot/, accessed on 10 December 2021) [6]. Its superior performance and ease of implementation led to its candidacy for the task. Without applying any additional feature engineering, MarMot was found to achieve an average tagging accuracy of 92% using the simplified tagset and 87% for the full tagset when evaluated on annotated test sets across all four languages. These data sets are further described in Section 4 in the evaluation of each of the core technologies. MarMot implements conditional random fields (CRFs) which is a reliable sequence prediction model used in several NLP tasks [17]. In terms of POS disambiguation, the task is often addressed by applying sequence prediction, but CRF's are rarely applied toward this end since the POS tagsets are often too large. This would render first-order dynamic programming too computationally expensive. However, MarMot applies a workaround by approximating a CRF using coarse-to-fine decoding and early updating by pruning the CRF during training. Their experiments showcased fast and accurate tagging results across six languages with different morphological properties and complexities. These results are reflected in this study through the final performance scores detailed in Section 4.

### *3.2. Lemmatizers*

Lemmatization is an important process in many NLP tasks, for example in parsing and machine translation [6]. The process entails establishing the relationship between inflected forms and their lemmata by normalizing all inflected forms of a lexical word to its common headword-form [18], a task that is most significant in morphologically rich languages where unlemmatized words are sparse [19].

In [5], the previous set of Nguni language lemmatizers sought to identify lemmas through the normalization of words using a rule-based approach. The normalization rules were derived from research on morphological analysis performed by Bosch et al. [9] and simulating their approach helped to identify the root or stem before appending the relevant terminative vowel to obtain the lemma. These rules, however, permit multiple possible analyses for a given word, some of which may yield the undesired root or stem and consequently result in an incorrect lemma.

To improve on this approach, we opted to introduce Lemming [6], a language independent statistical lemmatizer that is trainable on annotated corpora. Although Lemming operates at the token level, it allows for additional attributes in its training and execution. Since there is a strong mutual dependency between the POS and lemma of a wordform within its given context, a Lemming model can be provided with POS information to aid in lemma disambiguation. Using the simplified tagset, we included the POS in both the training and test sets. A real-world application of this lemmatizer for unseen data could allow a POS tagger to be executed on incoming tokens prior to lemmatization.

### *3.3. Morphological Analyzers*

The last set of core technologies developed as part of this project consists of morphological analyzers for each of the four Nguni languages which is an important component in language engineering applications such as machine translation and spelling error correction. It can also provide a sufficient starting point for NLP related research, especially when dealing with conjunctive languages [20]. Full morphological analysis generally entails segmenting a word into its individual morphemes, the smallest meaning-bearing units of a word, and obtaining insight into the underlying interaction among them through their syntactic classes [21].

The morphological complexity of words varies for different language families, but it happens to be pronounced in conjunctive languages, which possess words composed of aggregating morphemes that may sometimes undergo spelling transformations during agglutination [22]. The considered Nguni languages follow a comparable conjunctive writing system where the morphemes of its words are written unseparated. Yet, because the meaning of a word is determined by its morpheme composition, it is necessary to isolate them to allow for morphosyntactic analysis [9]. To achieve this, we perform morphological analysis using a two-tier approach where a token is firstly decomposed into its canonical morphemes (similar to the technologies developed during the NCHLT project) before determining their syntactic classes using a pipeline approach. Both these components operate independently but decode sequentially, acting as individual tasks in a two-step series. These two tiers are described in the next sections.

#### 3.3.1. Tier 1: Morphological Decomposition

The first tier entails morpheme segmentation by identifying the constituent morphemes of a word. However, due to spelling transformations that manifest during agglutination, its decomposed form may not always be equal to the word in its written form [23]. We therefore distinguish between two forms of segmentation, namely surface segmentation and canonical segmentation. The former results in a set of substrings that concatenate to their original wordform whereas the latter yields a sequence of substrings that are true to the underlying forms of the morphemes, which can differ in their orthographic representation within the original wordform. For the purpose of this project, the decomposer is tasked with splitting tokens into their canonical morphemes by segmenting affixes, roots in the case of verbs, and stems in other parts of speech. For example, given a permitting context, the isiZulu word ngokuphathelene ("in relation to") can be divided into its canonical morphemes as nga-u-ku-phathelan-il-e.

To facilitate this task, we employ the Tilburg Memory-Based Learner (TiMBL) which is available as an open-source software package that implements a selection of *k*-nearest neighbor classification and feature weighting algorithms [24]. We consider the task of segmentation as a context-sensitive mapping problem similar to many NLP tasks [25] (e.g., machine translation) which allows for a memory-based classifier to learn a mapping between the surface form and canonical form of a word. Training instances are generated using a windowing method applied to the surface form of a word. When applied to our dataset, this method transforms every token into multiple instances that each highlight a unique character boundary as the point of focus. Segmentation and spelling transformations are derived using diminishing longest string matching between the original wordform

and its annotated morphological analysis that isolate differences at the character-level. This process results in a segmentation and transformation rule for the given character boundary. TiMBL learns to associate these rule-based classes based on the morphological context provided by a sliding window of six characters, three preceding and three succeeding the given boundary point. Apart from any instances that exhibit irregular morphology, Van den Bosch and Daelemans [21] demonstrated the successful predictability of spelling variations using this method which is well-suited to morphologically rich languages. Table 3 helps to illustrate this approach for a windowed instance based on six characters for the isiZulu word "ngokuphathelene" ("in relation to").

**Table 3.** Windowed instances and segmentation rules generated for the word "ngokuphathelene".

| Instance Number | Left Context | | | Point of Focus | Right Context | | Class |
|---|---|---|---|---|---|---|---|
| 1 | - | - | - | n | g | o | = |
| 2 | - | - | n | g | o | k | = |
| 3 | - | n | g | o | k | u | = |
| 4 | n | g | o | k | u | p | o > a*u* |
| 5 | g | o | k | u | p | h | = |
| 6 | o | k | u | p | h | a | * |
| 7 | k | u | p | h | a | t | = |
| 8 | u | p | h | a | t | h | = |
| 9 | p | h | a | t | h | e | = |
| 10 | h | a | t | h | e | l | = |
| 11 | a | t | h | e | l | e | = |
| 12 | t | h | e | l | e | n | = |
| 13 | h | e | l | e | n | e | 0 > an*il* |
| 14 | e | l | e | n | e | - | = |
| 15 | l | e | n | e | - | - | ne > 0 |
| 16 | e | n | e | - | - | - | = |

Each of the numbered instances in Table 3 are associated with morphological transformation and segmentation classifications. It is these classes that TiMBL is used to predict and when applied to the original wordform can produce the intended canonical segmentation. The classifications are made up of five different types of rules. Instance 1 represents the first type of class ("="), which denotes that no spelling transformation or segmentation occurs at the current point of focus in the original wordform. This class signifies that there is no difference between the surface and canonically segmented form of the word at that point of focus. Instance 4 represents the second type of class (o > a*u*) which signals a conversion type spelling transformation and is always assigned to the last letter in the left context window. The rule can, however, expand to include letters in the right context window as illustrated by instance 15 (ne > 0). Asterisks within any rule represent segmentation points and when expressed as an independent class, like in instance 6 (*), indicate that the decomposed wordform should be segmented at the current point of focus. Due to agglutination, characters that are omitted when morphemes are aggregated can again manifest in the canonical segmentation. This is dealt with as an insertion of characters like in instance 13 (0 > an*il*) where the letters "anil" are to be reintroduced between the letters l and e. Finally, the fifth type of classification depicts the omission of characters as instance 15 (ne > 0) where the letters "ne" are removed. In the end it was found that a context window size of three characters surrounding the point of focus was sufficient at providing enough local information to allow for adequate predictive segmentation since the majority of conversion-type transformation rules range between one and three characters.

This approach yielded an instance base of between 447,605 (Siswati) and 481,153 (isiNdebele) instances that consist of 98 (isiNdebele), 122 (Siswati), 96 (isiXhosa), and 124 (isiZulu) unique classes for the complete dataset. The most frequent classes across all languages are "=" and "*", occurring in around just over 50% of instances. Exceptional classes with an

occurrence frequency of less than three were however excluded from the training set since these likely constitute unique instances or potentially erroneous annotations.

#### 3.3.2. Tier 2: Morphological Tagging

The second tier in the pipeline approach entails assigning syntactic classes to the canonically segmented morphemes. This is of particular importance in conjunctive languages since the meaning of a word is a function of its underlying and often aggregated morphemes [26]. For true morphological analysis, it is therefore necessary to isolate individual morphemes and identify their syntactic role. Keeping in mind that the functional role of a morpheme in the Nguni languages is both influenced and constrained by its surrounding morphemes, it is thus important to model morphological tagging by accounting for internal context. However, most morphological taggers operate at the word-level where non-lexicalized features (e.g., case, gender, number, POS) are predicted using a context-representation that spans the encompassing sentence rather than the morpheme composition of a word [26]. In addition, these approaches often use a combined feature set which expresses multiple morphological aspects for a single word (e.g., Noun + A3sg + Pnon + Nom) thereby helping to explicitly model underlying relationships between these features. However, our intended task is focused on the morpheme-level and depends on the internal morpheme composition for context and a composite feature set would only further increase sparsity for an already large target space that consists of between 62 and 71 morphosyntactic tags per language. Table 4 summarizes the number of morpheme tags in each of the languages.

**Table 4.** Morpheme tag counts per language.

| Language | Morpheme Tag Count |
|---|---|
| NR | 70 |
| SS | 68 |
| XH | 62 |
| ZU | 71 |

As a solution, we instead opted to treat the task of morphological tagging similar to that of POS tagging to best accommodate a large tagset and focus the context to internal structure of the word itself. Each word is decomposed and presented to the tagger as individual morphemes, thereby representing the internal morpheme composition of a word as context and leveraging the tagger's capacity to learn context-dependent predictions. MarMot is employed to this end given its trainable pipeline and its capacity to be successfully applied in tagging Nguni Languages. We also opted to make use of a lexicon of tokens that have fixed morphological analyses to address any exceptional instances. This lexicon was composed of token instances that express an unchanging morphological analysis throughout the entire annotated dataset and which have an occurrence count of more than 10. In the end, each lexicon consisted of 895 (NR), 752 (SS), 1378 (XH), and 801 (ZU) instances. These closed set of analyses are likely due to irregular morphology or are associated with tokens that exhibit a static morphological analysis [21].

## 4. Evaluation Results

Each of the created language technologies were trained and evaluated on the respective language datasets according to an approximate 90% training and held-out 10% test split and performance was measured in terms of accuracy. Table 5 contains the class prediction accuracies of each of the language-associated core technologies alongside their NCHLT Text project counterparts, with the exception of a simplified POS tagset tagger. In order to compare the NCHLT morphological decomposers, results from our tier 1 approach are provided in Table 5. Full morphological analysis results are provided in Table 6.

**Table 5.** Test set class prediction accuracy scores (%) of each core technology compared to the previous NCHLT Text project.

| Dataset | Lemmatization | POS Tagging | | Morpheme Decomposition |
|---|---|---|---|---|
| | | (Simplified Set) | (Full Set) | |
| NR | 90.35 | 91.54 | 85.28 | 86.71 |
| SS | 90.20 | 91.42 | 87.46 | 84.94 |
| XH | 92.99 | 95.91 | 93.99 | 94.13 |
| ZU | 90.33 | 92.65 | 88.60 | 86.87 |
| NCHLT Text Accuracy | | | | |
| NR | 80.32 | - | 82.57 | 82.26 |
| SS | 81.60 | - | 82.08 | 83.42 |
| XH | 79.82 | - | 84.18 | 84.66 |
| ZU | 81.56 | - | 83.83 | 85.19 |

**Table 6.** Test set accuracy scores (%) for both tiers of morphological analysis, both at instance level and at word level.

| Language | Morpheme Decomposition | | Morpheme Tagging | | Morphological Analysis |
|---|---|---|---|---|---|
| | Instance-Level | Word-Level | Instance-Level | Word-Level | |
| NR | 94.32 | 86.71 | 93.07 | 83.63 | 84.75 |
| SS | 94.21 | 84.94 | 90.70 | 80.61 | 81.48 |
| XH | 97.97 | 94.13 | 96.10 | 92.27 | 93.83 |
| ZU | 94.60 | 86.87 | 91.77 | 83.46 | 84.37 |

These results showcase an improved performance over the rule-based technologies of [5], with the greatest improvements being concentrated in lemmatization and POS tagging. When Eiselen and Puttkammer [5] developed core technologies for 10 of the official South African languages, they concluded that the morphological analysis of disjunctively written languages performed relatively well, while those for conjunctively written languages warranted more research. In this study, we reduced the complexity of the task to a set of morphological transformations which proves to be an effective means to perform morphological decomposition and can be reliably generated from an annotated corpus without requiring expert knowledge on the morphology of each language. The performance of each of the core technologies exceeds previous rule-based accuracy scores across all the languages, thereby demonstrating a clear advantage of the chosen statistical, machine-learned approaches.

Additionally, the cascading accuracy of the two-tier approach on morphological analysis is presented in Table 6 where the accuracies of both tasks are presently separately. The first step in the process involves obtaining the canonical segmentation that is derived from applying the TiMBL predicted segmentation and spelling transformation rules for the test set tokens. Here, the instance-based accuracy refers to the rule class prediction accuracy of TiMBL when evaluated against a gold standard of expected class predictions per character boundary for every token in the test set. Complementary to this is the word-level accuracy, which is the result of applying the predicted rules to the test set tokens to derive their canonical segmentation. This is then evaluated by comparing the number of corresponding morphemes between the decomposed wordform and that of the gold standard morpheme segmentations. Similarly, the accuracy for morpheme tagging was first evaluated on an instance level using a gold standard of correctly segmented morphemes from the test set. The word-level accuracy is the result of tagging the generated canonical segmentations and comparing the number of corresponding morpheme and tag pairs with that of the final morphological analysis of the test set. Additionally, a post-processing step is used to rectify any malformed segmentation predictions from a lexicon of words that exhibit a static morphological analysis. This allowed for a slight increase in accuracy of around 1% and is the resulting final score for morphological analysis as depicted in Table 6.

To gain further insight into the morpheme tagging performance, Table 7 shows the instance-level error rates in relation to the test set tokens and their part-of-speech. Interestingly, open-class instances are not as susceptible to tagging errors as would be expected. This is likely due to not only their generally high degree of representation within the datasets but also because they consist of multiple morphemes and thereby provide more morphological context during tagging. In contrast, a closed class such as conjunctives consists mostly of tokens with a single morpheme which provides no context. In addition, ambiguity exists for some of these highly frequent single-morpheme tokens such as "kanye" (with) and "khona" (there) in Siswati which can take the POS and syntactic morpheme role of either an adverb or a conjunction impacting the statistical probability of its class prediction and thus morphological analysis. This is also true for foreign text which is rare in the data not decomposed in the training or test sets and are kept in its original form which may explain the increased error rate. When considering the demonstrative POS, a similar impact is apparent to a subset of its tokens where a single demonstrative morpheme (e.g., "lo") precedes a typical noun morpheme structure and acts as the sole indicator for its class and thus its own syntactic role (e.g., "lo[Dem][Pos]-m[BPre]-jikeleto[NStem]"). These instances imply that limited morphological context and class ambiguity may cause increased data sparsity and subsequently impact tagging accuracy. This warrants further investigation of the datasets and may help to highlight more idiosyncrasies that may need to be addressed during the annotation of the data.

**Table 7.** Test set instance-level error rate (%) for statistical morpheme tagging in relation to the token POS.

| Part-of-Speech | NR | SS | XH | ZU |
|---|---|---|---|---|
| Abbreviation | 0 | 9.09 | 0 | 0 |
| Adjective | 2.33 | 3.65 | 0 | 3.49 |
| Adverb | 8.77 | 6.12 | 1.36 | 3.69 |
| Class-indicating demonstrative | 16.67 | 16.45 | 2.11 | 12.86 |
| Conjunction | 0.57 | 42.34 | 1.02 | 0 |
| Copulative | 41.38 | 21.13 | 21.04 | 15.28 |
| Foreign | 0.00 | 70.00 | 6.25 | 33.33 |
| Ideophone | 30.00 | 0 | 11.11 | 25.00 |
| Interjection | 11.76 | 24.14 | 5.88 | 0.00 |
| Noun | 4.36 | 4.52 | 0.72 | 4.39 |
| Numerative | 0 | 5.56 | 0 | 0 |
| Possessive | 7.34 | 12.15 | 2.05 | 4.96 |
| Pronoun | 4.80 | 8.00 | 1.94 | 0 |
| Relative | 6.40 | 9.10 | 4.14 | 8.14 |
| Verb | 9.37 | 6.93 | 4.57 | 5.66 |

## 5. Conclusions

In this paper we trained and evaluated machine-learned implementations of core technologies. For lemmatization, the results show that adopting a language independent implementation such as Lemming and MarMot performs comparably better over their rule-based counterparts for the considered conjunctively written Nguni languages: isiNdebele, Siswati, isiXhosa, and isiZulu. In addition, we demonstrated the viability of memory-based learning for morphological analysis by reformulating the problem as a series of classification tasks and leveraging a POS tagger for syntactic morpheme tagging. Relying on statistical learning also realizes a few advantages over traditional rule-based technologies in that no more linguistic knowledge is presupposed than what is already present in the training corpus, it is language independent, and learning is comparably faster and automatic. In addition to the curated and annotated Nguni language datasets, these technologies help to support the continued development and distribution of easily accessible linguistic resources. The deliverables of this project also ensure the sustained progression of human language technology development for these resource-scarce languages. Furthermore, these

resources are intended to aid in the development of other linguistic technologies, such as chunkers, parsers, named entity recognition systems, language identification systems and eventually other language-specific applications such as machine translation, speech recognition, and speech synthesis for these languages.

**Author Contributions:** Conceptualization: M.J.P. and J.S.d.T.; data curation: M.J.P. and J.S.d.T.; formal analysis: J.S.d.T.; funding acquisition: M.J.P.; investigation: J.S.d.T.; methodology: M.J.P. and J.S.d.T.; project administration: M.J.P.; resources: M.J.P.; software: M.J.P. and J.S.d.T.; supervision: M.J.P.; validation: M.J.P. and J.S.d.T.; visualization: M.J.P. and J.S.d.T.; writing—original draft: J.S.d.T.; writing—review and editing: M.J.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR is a research infrastructure established by the Department of Science and Technology of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: https://repo.sadilar.org/handle/20.500.12185/7 (accessed on 10 December 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Gurevych, I.; Eckle-Kohler, J.; Matuschek, M. Linked lexical knowledge bases: Foundations and applications. *Synth. Lect. Hum. Lang. Technol.* **2016**, *9*, 1–146. [CrossRef]
2. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 2nd ed.; Pearson Education: Upper Saddle River, NJ, USA, 2009.
3. Packham, S. Crowdsourcing a Text Corpus for a Low Resource Language. Master's Thesis, University of Cape Town, Cape Town, South Africa, 2016.
4. Loubser, M.; Puttkammer, M. Viability of Neural Networks for Core Technologies for Resource-Scarce Languages. *Information* **2020**, *11*, 41. [CrossRef]
5. Eiselen, R.; Puttkammer, M.J. Developing Text Resources for Ten South African Languages. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 3698–3703.
6. Thomas, M.; Cotterell, R.; Fraser, A.; Schütze, H. Joint lemmatization and morphological tagging with lemming. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2268–2274.
7. Straka, M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, 31 October–1 November 2018; pp. 197–207.
8. Doke, C. Bantu languages, inflexional with a tendency towards agglutination. *Afr. Stud.* **1950**, *9*, 1–19. [CrossRef]
9. Bosch, S.; Jones, J.; Pretorius, L.; Anderson, W. Resource development for South African Bantu languages: Computational morphological analysers and machine-readable lexicons. In Proceedings of the Workshop on Networking the Development of Language Resources for African Languages of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 22–28 May 2006; pp. 38–43.
10. Gaustad, T.; Puttkammer, M.J. Linguistically annotated dataset for four official South African languages with a conjunctive orthography: isiNdebele, isiXhosa, isiZulu, and Siswati. *Data Brief* **2021**. under review.
11. Gaustad, T.; Puttkammer, M.J. Development of linguistically annotated parallel language resources for four South African languages. In Proceedings of the 2nd workshop on Resources for African Indigenous Language (RAIL) at the International Conference of the Digital Humanities Association of Southern Africa (DHASA) 2021, online, 29 November–3 December 2021; pp. 1–8.
12. Hocking, J. Language identification for South African languages. In Proceedings of the Annual Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), Cape Town, South Africa, 27–28 November 2014; p. 307.
13. Expert Advisory Group on Language Engineering Standards (EAGLES). Available online: http://www.ilc.cnr.it/EAGLES/home.html (accessed on 24 October 2021).
14. Voutilainen, A. Part-of-speech tagging. In *The Oxford Handbook of Computational Linguistics*, 1st ed.; Mitkov, R., Ed.; Oxford University Press: New York, NY, USA, 2003; pp. 219–232.
15. Van Rooy, B.; Schäfer, L. The effect of learner errors on POS tag errors during automatic POS tagging. *S. Afr. Linguist. Appl. Lang. Stud.* **2002**, *20*, 325–335. [CrossRef]

16. Taljard, E.; Faaß, G.; Heid, U.; Prinsloo, D.J. On the development of a tagset for Northern Sotho with special reference to the issue of standardisation. *J. Lit. Crit. Comp. Linguist. Lit. Stud.* **2008**, *29*, 111–137. [CrossRef]
17. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 282–289.
18. Plisson, J.; Lavrac, N.; Mladenic, D. A rule-based approach to word lemmatization. In Proceedings of the 7th International Multi-Conference Information Society (IS 2004), Ljubljana, Slovenia, 11–15 October 2004; pp. 83–86.
19. Groenewald, H.J. Automatic Lemmatisation for Afrikaans. Master's Thesis, North-West University, Potchefstroom, South Africa, 2006.
20. Kessikbayeva, G.; Cicekli, I. Rule based morphological analyzer of Kazakh language. In Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, MD, USA, 27 June 2014; pp. 46–54.
21. Van den Bosch, A.; Daelemans, A. Memory-based morphological analysis. In Proceedings of the 37th annual meeting of the association for computational Linguistics, College Park, MD, USA, 20–26 June 1999; pp. 285–292.
22. Van de Velde, M.; Bostoen, K.; Nurse, D.; Philippson, G. *The Bantu Languages*, 2nd ed.; Routledge: New York, NY, USA, 2019.
23. Moeng, T.; Reay, S.; Daniels, A.; Buys, J. Canonical and Surface Morphological Segmentation for Nguni Languages. *arXiv* **2021**, arXiv:2104.00767.
24. Daelemans, W.; Zavrel, J.; van der Sloot, K.; van den Bosch, A. MBT: Memory-Based Tagger, Reference Guide. Technical Report ILK 99-01. In *Induction of Linguistic Knowledge, Computational Linguistics*; Version 2.0; Tilburg University: Tilburg, The Netherlands, 2002.
25. Pilon, S.; Puttkammer, M.J.; Van Huyssteen, G.B. Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans. *J. Lit. Crit. Comp. Linguist. Lit. Stud.* **2008**, *29*, 21–41. [CrossRef]
26. Zalmout, N.; Habash, N. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 704–713.