Generation of segmented isiZulu text

Mkhwanazi, Sthembiso
Voice Computing, CSIR
smkhwanazi2@csir.co.za

Marais, Laurette
Voice Computing, CSIR
lmarais@csir.co.za

Abstract

The complex morphology, conjunctive orthography and widespread occurrence of morphophonological alternation in the Nguni languages have given rise to several efforts towards morphological segmentation of tokens of Nguni languages. For supervised methods, annotated data is required, which currently exists as canonically segmented data in the NCHLT corpus and surface segmented data in the Ukwabelana corpus. In this paper, we present a method and segmentation strategy based on a computational grammar for isiZulu. The grammar, which itself has some limitations in processing speed and robustness to unexpected input, is used to create a new set of segmentations for the tokens of the Ukwabelana corpus.

By training various models with the same architecture but on different datasets, we first show that our approach enables us to match the performance of a model trained on pre-existing data. We also show that our approach provides the flexibility to determine a suitable segmentation strategy and to generate data that reflects this strategy.

Keywords: Nguni languages, agglutinative languages, morphological segmentation, language models.

1 Introduction

IsiZulu is part of the Nguni language family, a group of low-resourced Bantu languages belonging to a larger Niger-Congo language family, and they are widely spoken in Southern Africa (Mesham et al. 2021). In South Africa, out of 12 official languages that currently exist, four of them are Nguni languages (isiXhosa, isiNdebele, isiZulu, and Siswati). This group of languages are agglutinative in their morphology and have a conjunctive orthography (Bosch & Pretorius 2002). As morphologically rich languages, words are typically formed by combining multiple small meaning-carrying units known as morphemes (Bosch & Pretorius 2002).

Since the Nguni languages are considered resource-scarce, this state has hampered progress towards developing technological tools essential to preserving these languages' long-term digital vitality (Loubser & Puttkammer 2020). This effect is especially severe for resource-scarce agglutinative languages, since a given root or stem may appear in hundreds of different morpheme sequences. Moreover, for the Nguni languages, these morpheme sequences are written conjunctively as single tokens. The consequence is a tendency towards data sparsity, where any given corpus is unlikely to contain sufficient forms of all roots and stems. This has led research efforts into morphological segmentation as an effective approach to language modelling for these languages.

2 Overview of linguistic features of isiZulu

In this section, we briefly describe the linguistic features of isiZulu (and its related languages) that require the kind of segmentation described in this paper.

Languages are typically grouped based on their morphological typology, which are commonly distinguished into four types, isolating, agglutinative, fusional and polysynthetic (Pirkola 2001). The majority of the Bantu languages are considered to have an agglutinative morphology and can be further categorized into having conjunctive or disjunctive orthography (Bosch & Pretorius 2002). Out of the nine Southern Bantu languages that exist in South Africa, five are considered to be disjunctively written and the other four (i.e. Nguni languages) are considered to be conjunctively written.

A language is considered disjunctively written when a single linguistic word can result in multiple ortho-

graphic words. Where in contrast, a conjunctively written language maintains a one-to-one correspondence between linguistic words and orthographic words. To illustrate this disjunction, let us consider the phrase "I will tie them", which in Southern Sotho, a disjunctively written language, it is written as, *ke tla mo tlama* 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, *ngizombophela*, with its constituent morphemes *ngi-zo-m-bophel-a*. In this regard to break a conjunctively written word into its respective morphemes, one will need to do a morphological segmentation.

In agglutinative languages such as the Nguni languages, the morpheme combination and order are usually restricted, based on word formation rules known as morphotactics. A central mechanism of morphotactics in the Bantu languages is via the noun classification system, which categorizes nouns into a number of noun classes based on prefixal morphemes (noun prefixes). These noun prefixes further play a pivotal role in linking nouns to other words and govern the grammatical structure of different parts of speech Bosch et al. (2008).

The agglutinating nature of the Nguni languages, coupled with their conjunctive orthography, seems to necessitate morphological segmentation for the purposes of language modelling. The approach described in this paper provides linguistic control over the process of creating suitable training data for such segmentation models.

3 Morphological segmentation for the Nguni languages

Morphological segmentation refers to segmentation of words or tokens done with reference to the morpheme sequence that has given rise to a surface form. When canonical segmentation is in view, it is quite possible for isiZulu to determine the "correct" segmentation, since the canonical forms of morphemes of the language are well understood from a linguistic point of view. However, when surface segmentation is attempted for isiZulu, the high degree of morphophonological alternation makes it less clear what a suitable segmentation should be. For example, in the word *ngomuntu*, the canonical morpheme sequence is *nga+u+mu+ntu*. Both *ng+o+mu+ntu* and *ngo+mu+ntu* could therefore be acceptable surface segmentations of the word, and the suitability of either may be argued for from a linguistic perspective.

The purpose of segmentation, however, is to enable better language modelling on other tasks, and it remains an open question what the ideal granularity of surface segmentation should be (Meyer & Buys 2022). This is why for an unsupervised method such as Byte-pair Encoding, the hyperparameter that fixes granularity must be optimised (Salesky et al. 2020).

In this paper, we present a method for generating surface segmented data according to a specific, but variable, segmentation strategy. This data can then be used for supervised training of a segmentation model. We describe the method for generating the data in the Section 4.1.

3.1 Related work

Despite several efforts involving various techniques, morphological segmentation for the Nguni languages is still considered an active field based on the NO-FREE-LUNCH THEOREM (Wolpert & Macready 1997). According to this theorem, there exists no one-size-fits-all morphological segmentation technique that can perfectly handle all languages, types of words, or linguistic structures. In this regard, different approaches and techniques should be explored since they have differing limitations in different settings. In this section we discuss research efforts in morphological segmentation with particular focus on the Nguni languages.

In computational linguistics, there are usually two approaches that are used to achieve morphological segmentation, namely rule-based or machine learning (Anand Kumar et al. 2010). The rule-based approach utilises a predefined set of rules based on ex-

perts' knowledge and the morphological structure of the given language (Eiselen & Puttkammer 2014). Pretorius & Bosch (2003) developed a morphological analyser for isiZulu utilising the language rules; this was based on the *Xerox Finite State tools* (Beesley & Karttunen 2003). A few years later Bosch et al. (2008) bootstrapped this work into other Nguni languages.

The work of Eiselen & Puttkammer (2014), commonly known as the NCHLT Text project, aimed to develop different text resources for low-resourced South African languages, which included morphological decomposers/segmenters. These decomposers were rule-based, and for the agglutinative and conjunctively written languages, i.e. Nguni languages, their implementations were based on the work of Bosch et al. (2006) on a morphological analyser. The resulting data was canonically segmented. Similarly, *Ukwabelana* project used a semi-automatic process that implemented a partial morphological isiZulu grammar based on definite clause grammar to create surface segmented isiZulu data (Spiegler et al. 2010).

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 makes their development and in some case their maintenance to be an expensive and time-consuming endeavour. Another significant challenge is their robustness, both to linguistic structures not covered by the rules as well as to unseen vocabulary.

In contrast to rule-based approaches, machine learning techniques learn patterns from large annotated data or raw data to perform various tasks (Murphy 2012). 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, semisupervised, and reinforcement learning (Murphy 2012). The first three are the most commonly used in the task of morphological segmentation.

In an unsupervised segmentation approach, the algorithm models from raw texts to produce respective segments. For this kind of technique, Mzamo et al. (2019) as well as Moeng et al. (2021) have investigated various techniques. In the context of supervised methods that draw patterns from labeled data, Moeng et al. (2021) trained a supervised canonical segmenter using the data from the NCHLT project, while du Toit & Puttkammer (2021) created a new linguistically annotated datasets and used them to train a canonical segmenters to be used for morphological analysers.

To date, supervised surface segmentation for isiZulu has primarily relied on the Ukwabelana dataset (Spiegler et al. 2010). Notably, researchers such as Quasthoff et al. (2014) and Cotterell et al. (2015) have leveraged this dataset to develop surface segmentation models. Among these, Cotterell et al. (2015) stand out as they not only developed a system called CHIPMUNK but also rigorously evaluated its performance. In their evaluation, they achieved an impressive F1-score of 87.80% for isiZulu, demonstrating the effectiveness of their segmentation approach.

In comparing various approaches, including semi-supervised learning, Spiegler et al. (2008) conducted an experiment on morphological segmentation with different levels of supervision. This study confirmed supervised techniques as a superior and preferable method of segmentation. Though this is the case, the acquiring of morphological segmented data has been deemed nontrivial and often expensive since it needs experts' knowledge.

4 Methodology

4.1 Datasets

The main contribution of this work is in demonstrating that suitable annotated data can be generated for training supervised surface segmentation models. Our experiment is based on the

Ukwabelana segmented tokens in order to allow for comparison against the most relevant existing dataset.

The Grammatical Framework isiZulu Resource Grammar (ZRG) is an implementation of isiZulu morphosyntax using the GF programming language. The GF runtime can be used to enable parsing of text to obtain syntax trees, as well as linearisation of syntax trees into natural language. The orthography engineering capabilities of the GF runtime allow for custom linearisation (Angelov 2015). An in-depth description of how the ZRG models isiZulu at a subword level is beyond the scope of this paper: the interested reader is referred to Marais & Pretorius (2023). Relevant to this work is the fact that the ZRG systematically uses a specific token for binding subwords together at runtime to form grammatically correct surface tokens. It is therefore possible to parse an isiZulu token in order to obtain a syntax tree, and to use a custom linearisation to produce its segmented version. Figure 1 shows the parse tree for the token *ngizombophela*, where it is clear that the grammar has parsed the token as consisting of a number of subword segments.

It should be noted that the ZRG was not originally implemented as a surface segmentation tool, but as a resource grammar (Ranta et al. 2020), which could either be used as a linguistic software library for application grammar engineering, or as a general purpose parser for isiZulu, albeit a somewhat brittle one. As such, the segmentations produced for this experiment were simply the default segmentations implicitly modelled in the grammar. It would be relatively simple to adapt the core morphology operations of the grammar to reflect a different segmentation strategy, whether more or less fine-grained.

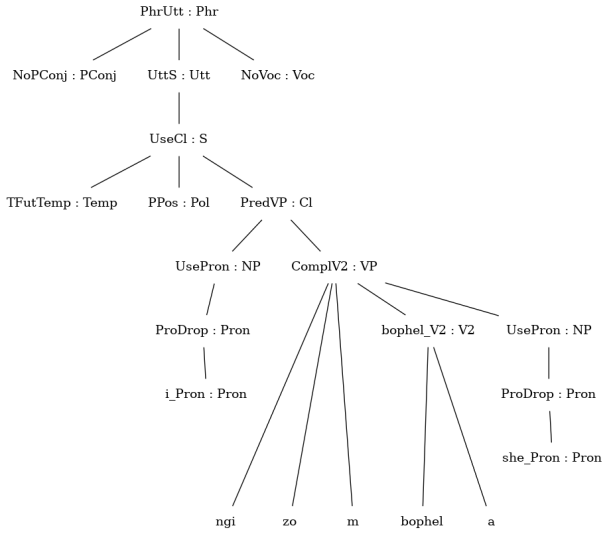

Figure 1: ZRG parse tree for ngizombophela

In order to compare our annotation process with what is already available, the segmented tokens of the Ukwabelana corpus were re-segmented using the ZRG. Due to the presence of non-isiZulu words and loan words that are not included in the lexicon of the ZRG, as well as some missing linguistic constructions in the ZRG, segmentations for only 7992 of the original 10040 could be recreated. To mitigate this, a data augmentation process was followed in which the stem segment of each re-segmented token was "scrambled": the first and last characters were retained, but all characters in between were replaced one-for-one with a randomly chosen character, consonants being replaced with consonants and vowels with vowels.

In this way, three datasets were prepared, namely the original Ukwabelana surface segmented dataset, the GF-based re-segmented dataset, and the GF-based scrambled dataset. The original Ukwabelana dataset was split into training, validation and evaluation sets (at a ratio of 80:10:10), and these were matched token-wise with the GF-based re-segmented annotations where available. In order to evaluate our work in a suitable way, the evaluation set for the GF-based re-segmentations was manually completed to be consistent with what the ZRG should have produced. The resulting split ratio for the GF-based re-segmented data was 78:10:12, since some tokens were left out of the training and validation sets, while the evaluation set retained its original size. The scrambled data was split 90:10 into a training and validation set.

These datasets formed the basis for training four different models using the same architecture, as show

in Table 1.

4.2 Model Training

We used a sequence-to-sequence model to train our segmenters, namely a Transformer model based on the work of Vaswani et al. (2017). Our transformer operates at a character level, taking a word as input and transforming it into the relevant surface morphemes separated by a space.

Training Setup: In terms of implementation, we adapted a ready-available Pytorch implementation of a translation Transformer model, [1]. To train the models, we split the original data set into three non-overlapping parts, namely, training, validation and evaluation sets. We kept the parameters constant for the model training across all four models. The relevant parameters were the number of heads (8), epochs (30), decoder layers (3), encoder layers (3), batch size (128) and embedding size (512).

5 Results and evaluation

To evaluate and compare the performance of our models, we used two metrics, namely, BLEU (Bilingual Evaluation Understudy) and chrF (character n-gram F-score); these are commonly used metrics in NLP, mostly to evaluate machine translation outputs. BLEU is a precision-based evaluation metric that implements a modified version of n-gram that measures the overlapping of words between the candidate set and reference(s) set (Papineni et al. 2002). Similarly, the chrF follows the same logic except that the n-grams are considered at the character level (Popović 2015). Both of these metrics are provided through the NLTK library for implementation. Considering that the BLEU scores here are measured on the segment-level, it is probably more indicative of success for the purposes of language modelling than chrF, since it is the segments that would be used as the vocabulary items of a language model.

In addition to BLEU and chrF, we also incorporated F1-score into our evaluation framework. This metric is commonly used in the context of morphological segmentation to provide a balanced assessment of precision and recall in terms of segmentation. Yet it is worth mentioning that this metric does not evaluate the order of the segments produced by a model, which is a crucial aspect of subword modelling, especially for generative models. Therefore, for this study, we primarily rely on BLEU and chrF as our main evaluation metrics, while we include the F1-score in our evaluation, to give the overall performance of the models and facilitating comparisons with other available segmenters for further experimentation.

In Table 2, each segmenter is evaluated on its own terms, so to speak, where the evaluation set reflects the implicit segmentation strategy of the data itself. As such, the results simply show to what extent the Transformer model was able to learn this segmentation strategy for each dataset. The models for which scrambled (synthetic) data was used during training were evaluated on the real data of the *gf-eval* set. It is clear that, in terms of the chrF score, the addition of the scrambled data to the GF-based re-segmentations allowed *gf-scramble-seg* to slightly outperform the model trained on the original data. In terms of the BLEU score, the results for these models are reversed, and hence we can say that our data annotation process enabled us to match the performance achieved on the existing dataset.

It is worth considering that on the Ukwabelana evaluation, the model trained on the original data (*ukwabelana-seg*), as shown in Table 2, produces a significantly better F1-score than *gf-scramble-seg*. Despite this, their BLEU scores are comparable and the latter's chrF score is significantly better. This seems to show that, despite worse precision and recall, *gf-scramble-seg* seems to have learnt enough about the order of segments to achieve comparable results to *ukwabelana-seg*.

The results for *scramble-segm* are particularly intriguing, since we know that the model could not have seen any of the stems present in the tokens of the evaluation set. Hence, its performance can be ascribed to having learnt surface segmentation of isiZulu somewhat independently of the vocabulary

Table 1: Models and datasets

Model name	Training data	Validation data	Evaluation data
ukwabelana-seg	ukwabelana-train	ukwabelana-val	ukwabelana-eval
gf-seg	gf-train	gf-val	gf-eval
scramble-seg	scramble-train	scramble-val	gf-eval
gf-scramble-seg	gf-train + scramble-train	gf-val + scramble-val	gf-eval

Table 2: Results on designated evaluation set

Model name	chrF	BLEU	F1-score
ukwabelana-seg	0.808	**0.909**	86.906
gf-seg	0.881	0.889	80.543
scramble-seg	0.881	0.851	72.243
gf-scramble-seg	**0.894**	0.901	81.940

Table 3: Results on manually segmented evaluation set

Model name	chrF	BLEU	F1-score
ukwabelana-seg	**0.848**	0.781	58.824
gf-seg	0.833	0.828	64.014
scramble-seg	0.825	0.787	56.373
gf-scramble-seg	**0.846**	**0.838**	61.041

of the corpus it was trained on.

Having evaluated each segmenter on its own terms, we also evaluated the models on a small set of manually segmented tokens from a different domain, namely 100 randomly chosen tokens from the usage examples of the isiZulu African Wordnet. The manual segmentation was performed to reflect a specific segmentation strategy, aimed at a moderate granularity. For example, pre-prefixes and base prefixes of nouns were considered as single segments and verbs were segmented in such a way that the root and its verbal extensions formed a single segment (due to the lexical-semantic effect of the extensions on the root). The strategy bears some natural similarity to what the ZRG produces by default, since both stem from a linguistically oriented approach to isiZulu subwords. However, the specifics of the strategy are at this stage not as important as that it was chosen consciously and applied consistently, since to decide on and implement a different strategy would, as mentioned earlier, require a relatively simple adaptation of the ZRG.

Table 3 shows the results, and here it is clear that the *gf-scramble-seg* outperformed the other models with respect to the BLEU score. We therefore conclude that it is possible to use the ZRG to annotate data according to a specific segmentation strategy, which can be used to outperform models trained on existing, unalterable annotated data.

The advantage of being able to control the segmentation strategy of the annotation process is, of course, that various strategies could be compared, possibly even for different downstream tasks and domains. Furthermore, our process need not be limited to the Ukwabelana dataset. As an automated process, it is relatively cheap, and can be applied to more datasets, which will likely improve its performance further.

6 Conclusion

In this paper, we introduced a method and segmentation strategy based on a computational grammar for isiZulu, namely the isiZulu Grammatical Framework Resource Grammar, to create a morphological surface segmented data. We have evaluated this method in terms of its ability to generate data that can match what is currently available as surface segmented data, namely the Ukwabelana dataset. Moreover, the method allows for flexibility with regards to segmentation strategies, enabling the exploration of optimal strategies based on linguistic approaches to isiZulu subword modelling.

In a resource scarce environment, annotated data is especially costly to obtain, and hence an automated approach to producing such data would enable new

avenues of research in supervised methods, which in turn would contribute to advancing research into subword language modelling for isiZulu.

Notes

[1] `https://pytorch.org/tutorials/beginner/translationtransformer`.

References

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. (2015), Orthography engineering in grammatical framework, *in* 'Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop', pp. 33–40.

Beesley, K. R. & Karttunen, L. (2003), 'Finite-state morphology: Xerox tools and techniques', *CSLI, Stanford* pp. 359–375.

Bosch, S. E. & Pretorius, L. (2002), 'The significance of computational morphological for Zulu lexicography', *South African Journal of African Languages* **22**(1), 11–20.

Bosch, S., Jones, J., Pretorius, L. & Anderson, W. (2006), Resource development for South African Bantu languages: computational morphological analysers and machine-readable lexicons, *in* 'Workshop Organiser (s)', p. 38.

Bosch, S., Pretorius, L. & Fleisch, A. (2008), 'Experimental bootstrapping of morphological analysers for Nguni languages', *Nordic Journal of African Studies* **17**(2), 23–23.

du Toit, J. S. & Puttkammer, M. J. (2021), 'Developing core technologies for resource-scarce Nguni languages', *Information* **12**(12), 520.

Eiselen, R. & Puttkammer, M. (2014), Developing Text Resources for Ten South African Languages, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)', pp. 3698–3703.

Loubser, M. & Puttkammer, M. J. (2020), 'Viability of neural networks for core technologies for resource-scarce languages', *Information* **11**(1), 41.

Marais, L. & Pretorius, L. (2023), Parsing IsiZulu Text Using Grammatical Framework, *in* 'International Symposium on Distributed Computing and Artificial Intelligence', Springer, pp. 167–177.

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, *in* 'Findings of the Association for Computational Linguistics: EMNLP 2022', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 6636–6649.
URL: *https://aclanthology.org/2022.findings-emnlp.494*

Moeng, T., Reay, S., Daniels, A. & Buys, J. (2021), Canonical and surface morphological segmentation for Nguni languages, *in* 'Southern African Conference for Artificial Intelligence Research', Springer, pp. 125–139.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, MIT press.

Mzamo, L., Helberg, A. & Bosch, S. (2019), Towards an unsupervised morphological segmenter for isiXhosa, *in* '2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)', IEEE, pp. 166–170.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: a method for automatic evaluation of machine translation, *in* 'Proceedings of the

40th annual meeting of the Association for Computational Linguistics', pp. 311–318.

Pirkola, A. (2001), 'Morphological typology of languages for IR', *Journal of Documentation* **57**(3), 330–348.

Popović, M. (2015), chrf: character n-gram F-score for automatic MT evaluation, *in* 'Proceedings of the tenth workshop on statistical machine translation', pp. 392–395.

Pretorius, L. & Bosch, S. E. (2003), 'Computational aids for Zulu natural language processing', *Southern African Linguistics and Applied Language Studies* **21**(4), 267–282.

Ranta, A., Angelov, K., Gruzitis, N. & Kolachina, P. (2020), 'Abstract syntax as interlingua:: Scaling up the grammatical framework from controlled languages to robust pipelines', *Computational Linguistics* **46**(2), 425–486.

Salesky, E., Runge, A., Coda, A., Niehues, J. & Neubig, G. (2020), 'Optimizing segmentation granularity for neural machine translation', *Machine Translation* **34**, 41–59.

Spiegler, S., Golénia, B., Shalonova, K., Flach, P. & Tucker, R. (2008), Learning the morphology of Zulu with different degrees of supervision, *in* '2008 IEEE Spoken Language Technology Workshop', IEEE, pp. 9–12.

Spiegler, S., Van Der Spuy, A. & Flach, P. A. (2010), Ukwabelana-an open-source morphological zulu corpus, *in* 'Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)', pp. 1020–1028.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

Wolpert, D. H. & Macready, W. G. (1997), 'No free lunch theorems for optimization', *IEEE transactions on evolutionary computation* **1**(1), 67–82.