Canonical Segmentation and Syntactic Morpheme Tagging of Four Resource-scarce Nguni Languages

Du Toit, Jakobus S.
Puttkammer, Martin J.
CTexT®, North-West University
Jaco.duToit@nwu.ac.za
Martin.Puttkammer@nwu.ac.za

Abstract

Morphological analysis involves investigating the syntactic class of a word but can also extend to the decomposition and syntactic analysis of its underlying morpheme composition. This is especially relevant to languages with an agglutinative writing system where multiple linguistic words are expressed as a single orthographic word. In this paper, we propose a memory-based approach to canonical segmentation using a windowing approach to recover the uncondensed morphemes that differ from the surface form of a word. Additionally, we propose treating the syntactic labelling of morphemes as a sequence labelling task, similar to part of speech tagging. This approach leverages the internal morpheme composition of a word as local context in much the same way that the surrounding sentence of word serves in the disambiguation of its part-of-speech. Both tasks are modelled separately but performed sequentially by cascading the decomposed morphemes of a word into the task of syntactic labelling. When evaluated on four resource-scarce, conjunctively written Nguni languages, the proposed approach achieves an overall accuracy ranging between 82% and 92% which outperforms previously developed rule-based analysers for the same languages.

Keywords: morphological analysis, canonical segmentation, syntactic morpheme tagging

1 Introduction

Access to quality linguistic resources is crucial to any research and development efforts in the field of Natural Language Processing (NLP). Collecting and assembling such resources is generally expensive and time-consuming, especially in the case of resource-scarce languages as is the case for ten of the eleven official languages of South Africa. To address this scarcity and promote research and development efforts in the area of human language technology, the South African government has established several legislative frameworks that promote the use and advancement of these languages and has funded various development projects over the last two decades. One such project was the National Centre for Human Language Technology's Text project (Eiselen and Puttkammer 2014), that collected circa 50,000 tokens, annotated with part of speech (POS), lemma and morphological analysis information for ten of the official South African languages. POS taggers, lemmatisers and rule-based morphological decomposers were also developed. A follow-up project expanded on this work by developing an additional 50,000 annotated tokens and improved technologies for four of these languages, *viz.* isiNdebele (NR), Siswati (SS), isiXhosa (XH) and isiZulu (ZU). In this paper, we focus on the morphological analysers that were developed in the follow-up project.

In the next section we provide a brief overview of the project, followed by a description of morphological decomposition versus analysis. We then (section 4) present our two-tiered approach to morphological analysis and results (section 6). Conclusions and future work are discussed in section 7.

2 Background

The aim of this project was to create corpora annotated with POS, lemma and morphological analysis information for four resource-scarce South African Nguni languages and to develop associated core technologies using these datasets. In this paper the focus is on the development of the morphological analysers. For a detailed description of the development process followed during the annotation of the data, see Gaustad & Puttkammer, 2021. A description of the other core technologies can be found in Du Toit & Puttkammer, 2021.

South Africa is a linguistically diverse country with eleven official languages, nine of which are Southern Bantu languages that follow either a conjunctive or disjunctive writing system. This study only focuses on the conjunctively written Nguni languages namely, isiNdebele, Siswati, isiXhosa, isiZulu. To illustrate the orthographic difference between these writing systems, Prinsloo & De Schryver (2002) use the example phrase "I love him/her". The agglutinative nature of isiZulu produces the phrase as a single word, *ngiyamthanda*, whereas it would be written disjunctively in Sepedi as four separate words, *ke a mo rata*. Although this study is restricted to the four South African Nguni languages, insights gained during development could also be to be applied to languages that share a similar morphology.

3 Morphology

Morphological features are an important aspect in developing language engineering technologies and applications such as machine translation and spelling error correction. Morphological features typically refer to either lexicalized (e.g. lemmas) and non-lexicalized features (e.g. gender, case, number). Most approaches that operate at word-level annotate tokens with lexicalized features using either combined feature sets or by modelling individual features as separate tagging tasks.

A combined feature set expresses multiple aspects as a single composition of non-lexical tags (e.g. Noun+A3sg+Pnon+Nom). This allows for underlying relationships between different features to be modelled explicitly but it does result in a large target space that further increases sparsity for morphologically rich languages. On the other hand, modelling these features separately yields a smaller target space but constrains the capacity of a model to learn any inter-feature dependencies which is crucial for our intended task. When operating at the morpheme-level, the functional role of a morpheme in the Nguni languages is both influenced and constrained by its internal context. Two approaches that utilise morphological features are so called morphological decomposition where morphemes are identified, and morphological analysis, where tags are assigned to each morpheme based on its grammatical function.

Morphological decomposition entails dividing a word into its constituent morphemes, the smallest meaning-bearing units of a word (Ruokolainen *et al.* 2013). However, these morphemes may not be orthographically equal to the corresponding segment of the word in written form when spelling transformations manifest during agglutination. We thus distinguish between two forms of segmentation, surface segmentation and canonical segmentation. The former yields a set of substrings that concatenate to the original word from, whereas the latter yields a sequence of canonical morphemes that are true to the underlying forms of the morphemes but potentially differ in their orthographic representation within the original wordform. Morphological decomposition is beneficial in helping to support further analyses for NLP tasks, especially in resource languages where data sparsity can undermine the quality of a task.

The set of decomposers for conjunctive languages previously developed as part of the NCHLT project were rule-based implementations that follow the work of Bosch *et al.* (2006). These implementations entailed recursively identifying all affixes in a token whereafter the remaining constituent would be verified against a lexicon of roots and stems. Only in instances where a valid stem or root and a valid combination of affixes were confirmed would the decomposition be deemed successful. These rule deductions were based on the collection of 50,000 annotated tokens developed as part of that project. These and the newly developed decomposers split tokens into their constituent morphemes, including each constituent affix, roots in the case of verbs, and stems in other parts of speech. For example, the isiZulu word *ukusebenzisa* ("use") is split into its constituent morphemes as *u-ku-sebenz-is-a*, where each affix boundary is marked in conjunction with the verb root.

Morphological analysis is of particular importance when applied to the Nguni languages since its words are naturally composed of aggregating morphemes that may undergo spelling alterations after their unions. These

languages follow a conjunctive writing system which leads to an agglutinative orthography where morphemes are written unseparated. Yet, the meaning of a word is a function of its morphemes, and it is therefore necessary to isolate these individual morphemes for further syntactic analysis. Full morphological analysis entails both the segmentation of morphemes and the analysis of the interactions among the underlying morphemes of a word by determining their syntactic classes (Van den Bosch and Daelemans 1999). For example, the same isiZulu word used in the previous example, *ukusebenzisa* ("use"), is morphologically decomposed and analyzed as u[NPrePre]-ku[BPre]-sebenz[VRoot]-is[CausExt]-a[VerbTerm], where the syntactic class of each morpheme is assigned. For this example, the syntactic classes consist of a secondary noun prefix (NPrePre), a primary noun prefix (BPre), a verbal root (VRoot), a causative verbal suffix (CausExt), and a verb terminative (VerbTerm).

4 Morphological Analyser Design

This section describes the task of modelling morphological analysis and the level of granularity in analyses that existing solutions support. Our initial investigations into a suitable approach to morphological analysis included NLP pipelines that typically accommodate multiple tagging and morphological tasks, as in the case of UDPipe[1] (Straka 2018) and MarMoT[2] (Müller *et al.* 2013). MarMoT is a generic CRF framework capable of both POS tagging and morphological analysis, similarly UDPipe is a trainable pipeline for tokenization, POS tagging, dependency parsing, and morphological analysis that employs contextualized BERT embeddings. However, the capacity for morphological analysis in these solutions only accommodate non-lexicalised features that distinguish lexical and grammatical properties of words. Attempts to adopt this system to annotate Nguni language tokens with syntactic morpheme classes were expectedly unsuccessful, since its capacity for morphological analysis is limited to only the word-level and context is derived from the encompassing sentence rather than the local composition of morphemes.

Taking into account the required depth of morphological analysis and because canonical segmentation and class annotation at the morpheme level are not usually addressed as a single task, we approached morphological analysis as two separate problems. Both morpheme segmentation and morphological analysis are treated as a sequence tagging task, as was the approach followed by Sorokin & Kravtsova's (2018) for Russian, a comparably agglutinative language.

In their approach, morpheme segmentation was represented using the BMES labelling scheme where the classes account for beginning (B), middle (M), and ending (E) as well as single (S) single letter morphemes. Additionally, morphemes were tagged according to their type namely, root, prefix, suffix, ending, postfix, link, and hyphen. Thus, the task of their system was to predict segmentation and type labels for a sequence of letters. However, morpheme classifications constitute a small target space of only 7 labels thereby ensuring better prediction accuracy. Adopting the approach taken by Sorokin & Kravtsova (2018) to our task resulted in a low tagging accuracy given our larger target space and subsequent increased sparsity as our labelling scheme contains 71 tags for ZU, 70 for NR, 68 for SS and 62 tags for XH.

The before-mentioned NLP pipelines and other approaches to morphological analysis typically model morphological features at the word-level and context representation spans the entirety of the encompassing sentence. In contrast, morpheme segmentation is modelled at the character-level and thus the difference in granularity explains why most approaches approach each task separately unless the target space can be kept relatively small as in the case of Sorokin & Kravtsova (2018). In contrast, Van den Bosch & Daelemans (1999) was however successful at jointly modelling morphological boundaries, syntactic classes, and spelling transformations at the character-level as a single

[1] https://ufal.mff.cuni.cz/udpipe/2

[2] http://cistern.cis.lmu.de/marmot/

task which resulted in a large target space but was supported by a large dataset. This approach is compatible with our intended task but applying a similar method to our limited data sets would introduce sparsity and only diminish the quality of predictions. Nonetheless, inspiration was taken from Van den Bosch & Daelemans (1999) in reducing the task of segmentation to a sequence of classification tasks using a windowing method across the characters of a word. To accomplish tagging each morpheme with its syntactic morphological class we opted to treat the task similar to that of POS tagging and representing individual morphemes as words to realise local context from within the internal morpheme composition of a word. This approach is discussed in greater detail in the next section.

The proposed approach follows a two-tier design for sequential segmentation and analysis. Isolating the two tasks allowed for greater accuracy in segmentation, which suffered the most in prior attempts given the complexity of the languages. A pipeline approach is subject to cascading errors, but the significant accuracy of the morphological decomposer as the initial component in the pipeline does little to impact the overall accuracy.

4.1 Tier 1: Morphological decomposition

The first tier of the approach is a memory-based learning system that models morphological decomposition and spelling transformations as a series of classification tasks inspired by the work of Van den Bosch and Daelemans (1999). Memory-based learning is a class of supervised, inductive machine learning algorithms that learn based on examples of a task stored in memory. The Tilburg Memory-Based Learner (TiMBL) facilitates this function as an open-source software package that supports a selection of k-nearest neighbour classification and feature weighting algorithms (Daelemans *et al.* 2004). When new instances of a learnt task are presented to the TiMBL model, computational effort is invested in finding the best-matching instances from memory as determined by a similarity metric. Once the nearest neighbour (or instance) is identified in memory, the associated class is transferred to the new instance. Memory-based approaches have been successfully applied to other natural language processing tasks such as hyphenation and compounding analysis (Pilon *et al.* 2008).

Morphological decomposition can be treated as a context sensitive mapping problem, similar to most linguistic problems (e.g. source to target language translation, text to speech synthesis etc.). As part of this approach, TiMBL is tasked with learning this particular mapping through a windowing method from the surface form of a word to its canonical segmentation. Windowing in this manner transforms a word into multiple instances where each instance is focused on the boundary between letters, or the start and end boundary of the word. The method is illustrated in table 1, where a sliding context window of 6 letters traverses the length of the word with the point of focus positioned 3 letters left and right of the given boundary. The instance class expresses whether the given boundary maps to a point of segmentation in its decomposed form and if any letters within its right 3 letter window should undergo a spelling transformation in obtaining its canonical segmentation. A window size of 6 letters was found to be sufficient for the task since most conversion-type transformation rules range between 1 and 3 characters. In the end, the canonical segmentation is obtained by constructing the surface form of the word according to the predicted transformation rules.

Table 1: Segmentation rules for the word ngokuphathelene

Instance Number	Left Context			Point of Focus	Right Context			Rule 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	-	-	-		=

Per illustration of the approach, Table 1 contains 16 instances that were generated from the isiZulu word *ngokuphathelene* ("in relation to") with their associated morphological transformation and segmentation classes that produce its canonical segmentation. The generated classes can express five different types of transformations, the first is represented in instance 1 with the "=" class. This indicates that no transformation or segmentation takes place at the current point of focus in the original surface form. These classes denote no difference between the surface form and the canonical segmentation for the right context at the point of focus. The second type of class is represented in instance 4 (o>a*u*) which is indicative of a conversion transformation. These classes are assigned to the letter immediately left of the point of focus but can include letters from the right context when a transformation affects multiple letters like in instance 15 (ne>0). Any asterisks contained within the class represent segmentation points in the canonical form. Instance 6 (*) depicts an asterisk as an independent class which denotes a segmentation point in the canonical form of the word at the current point of focus. The fourth type of classification depicts the insertion of characters at the given boundary as in instance 13 (0>an*il*) where the letters and segmentation points "an*il*" are to be inserted between the letters l and e. The fifth and last class denotes the removal of letters like in the case of instance 15 (ne>0) where the trailing letters "ne" are to be omitted in the canonical segmented wordform.

The classes were derived using diminishing longest string matching between the surface and canonical forms to isolate the differences at character level. This yielded an instance base ranging between 447,605 (SS) and 481,153 (NR) that consist of 98 (NR), 122 (SS), 96 (XH), and 124 (ZU) unique classes, but excluding exceptional classes with an occurrence frequency of less than 3. Across all four languages, the most frequent classes are (=) and (*), making up just over 50% of all class instances in each language.

In order to determine which learning algorithm best served the task of canonical segmentation, TiMBL was trained on the generated instances using each of the five k-nearest neighbour algorithms that it accommodates. A parameter search was used as part of this experiment to determine which hyperparameter adjustments could provide the greatest prediction accuracy. Since the two-tier approach relies on the predicted segmentation, it was important to obtain a reliable morphological decomposition to ensure as few errors as possible that may hamper the second tier's capacity to reliably annotate each morpheme with the related morphological analysis. In the end, IB2 was found to provide the greatest accuracy. IB2 operates similarly to other memory-based learning algorithms by keeping instances in memory that contribute to the potential classification of unseen instances during learning and uses a distance metric to determine class association. IB2 however employs an incremental editing strategy where its instance base is seeded with a certain (typically small) number of instances and only adds to instances in memory when it is misclassified by the k-NN classifier. The intention behind this approach is to construct an instance base that naturally establishes boundaries or key instances within memory with deviating or atypical instances to allow for greater generalisation. To further improve generalisation, each windowed instance was also associated with a seventh feature namely, the actual POS tag of the token. This improved the accuracy of predicted segmentation classes by around 1 to 2 percentage points.

4.2 Tier 2: Morphological Tagging

The second tier of the approach entails adopting a POS tagger to model syntactic morpheme classes of canonically segmented tokens. The chosen candidate for this approach is MarMoT, (Müller *et al.* 2013) given its trainable NLP pipeline and its successful application as a POS tagger on the considered Nguni languages in Du Toit & Puttkammer, 2021. Because morphological classes are context dependent, treating the task of syntactic morphological tagging of morphemes similar to that of POS tagging helps to take advantage of its capacity to learn predictions within the context of a sentence. This is achieved by treating each segmented morpheme as a word and sequentially ordering them to resemble a

sentence, thereby realising the internal morpheme composition of a word as local context in their tagging predictions.

The segmented morphemes are provided to the tagger as words along with the actual POS tag of the word as an additional feature. This improved the quality of its syntactic morphological class predictions by increasing the tagging accuracy by around 1 to 3 percentage points.

5 Results

The annotated data developed as part of this project is divided according to an approximate 90% training and 10% test split. Without including any punctuation, the following token counts make up each dataset split.

Table 2: Token counts per language dataset

	NR	**SS**	**XH**	**ZU**
Train	39,251	37,223	37,926	38,489
Test	4,441	4,084	4,277	4,345

To evaluate the segmentation competency of TiMBL, the before-mentioned test sets were first transformed into windowed instances of 6 letters and segmentation rules were generated. Each instance was then associated with the transformation and segmentation rule as its intended classification. The instances also included the POS tag of the token as a seventh, additional feature alongside its 6 windowed letters. Finally, the four trained TiMBL models were evaluated by presenting the windowed instances for classification. By comparing the class predictions to the intended segmentation rules, a prediction accuracy ranging between 96% and 98% was achieved for each of the language-specific models. These results are listed in Table 3.

Table 3: Segmentation and transformation rule class prediction accuracy

	NR	**SS**	**XH**	**ZU**
Accuracy (%)	97.02	96.92	98.30	96.55

Similarly, to evaluate the syntactic morphological tagging competency of MarMoT, the test set of tokens were segmented into their intended morphemes and associated morphological tags. The test sets were then presented to MarMoT for tagging along with the POS tag of the token as additional feature. By comparing the predicted morphological tags to the intended tags, a prediction accuracy ranging between 91% and 96% was achieved for each of the language-specific models. These results are listed in Table 4.

Table 4: Prediction accuracy for syntactic morpheme tagging

	NR	**SS**	**XH**	**ZU**
Accuracy (%)	92.89	91.27	96.77	94.64

Finally, the combined accuracy of both tiers was evaluated by cascading the resulting canonical segmentation into MarMoT for syntactic morphological tagging. This was performed by applying the predicted transformation and segmentation rules in the first tier to the tokens in the test set to produce the canonically segmented morphemes. These morphemes were then tagged by MarMoT to obtain their syntactic morphological class. The results of the first tier were evaluated by comparing the number of corresponding morphemes between the intended test set of canonical segmentation and the predicted rule-based transformation of the token. Similarly, the morphological tag and morpheme associations were evaluated by comparing the predicted morpheme and tag pairs to the intended test set pairings for each token. These results are listed in Table 5.

Table 5: Canonical segmentation and morphological tagging prediction accuracy

	NR	**SS**	**XH**	**ZU**
Segmentation	86.71	84.94	94.13	86.87
Tagging	83.63	80.61	92.27	83.46

6 Conclusion

To ensure the continued development of human language technologies, it is important that resources be developed and distributed. We have described one such effort funded by the South African government for four resource-scarce Nguni languages. These resources are available as open-source modules from the SADiLaR resource

catalogue[3] and can aid researchers and developers in improving and furthering the reach of language technology. Furthermore, the approaches taken toward morphological analysis at the morpheme level of agglutinative languages may also provide evidence for its viability and applicability to similar languages. The lexical resources provided as part of this project will enable further improvements and alternative approaches in developing related language technologies.

Acknowledgements

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 Innovation (DSI) of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

References

Bosch, S, Jones, J, Pretorius, L & Anderson, W 2006, 'Resource development for South African Bantu languages: computational morphological analysers and machine-readable lexicons' In *Proceedings on the Workshop on Networking the Development of Language Resources for African Languages at the 5th International Conference on Language Resources and Evaluation*, pp. 38-43.

Daelemans, W, Zavrel, J, Van Der Sloot, K & Van den Bosch, A 2004, TiMBL: Tilburg memory-based learner, version 6.4: reference guide, Tilburg, Tilburg University.

Du Toit, JS & Puttkammer, MJ 2021, Developing Core Technologies for Resource-scarce Nguni Languages. Manuscript submitted for publication.

Eiselen, R & Puttkammer, MJ 2014, 'Developing Text Resources for Ten South African Languages', In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 3698-3703.

Gaustad, T & Puttkammer, MJ 2021, Linguistically annotated dataset for four official South African languages with a conjunctive orthography: isiNdebele, isiXhosa, isiZulu, and Siswati. Manuscript submitted for publication.

Müller, T, Schmid, H & Schütze, H 2013, 'Efficient higher-order CRFs for morphological tagging.' In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 322-332.

Pilon, S, Puttkammer, MJ & Van Huyssteen, GB, 2008. 'Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans', *Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, vol. 29 no. 1, pp. 21-41.

Prinsloo, DJ & De Schryver, GM 2002, 'Towards an 11 x 11 array for the degree of conjunctivism/disjunctivism of the South African languages', Nordic Journal of African Studies, vol. 11 no. 2, pp. 249-265.

Ruokolainen, T, Kohonen, O, Virpioja, S & Kurimo, M 2013, 'Supervised morphological segmentation in a low-resource learning setting using conditional random fields' In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 29-37.

Sorokin, A & Kravtsova, A 2018, 'Deep convolutional networks for supervised morpheme segmentation of Russian language' In *Conference on Artificial Intelligence and Natural Language*, pp. 3-10

Straka, M 2018, '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*, pp. 197-207.

Van den Bosch, A & Daelemans, A 1999, 'Memory-based morphological analysis' In *Proceedings of the 37th annual meeting of the association for computational Linguistics,* pp. 285-292.

Zalmout, N & Habash, N 2017, '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,* pp. 704-713.

[3] https://repo.sadilar.org/handle/20.500.12185/7