African Journal of Inter/Multidisciplinary Studies (AJIMS)
Volume 8, Issue 1, pp. 1-12
e-ISSN: 2663-4589

The Efficiency of IsiNdebele Part of Speech Tagger: A Quantitative Analysis

Muzi Matfunjwa and Nomsa Skosana

[1]South African Centre for Digital Language Resources (SADiLaR), North-West University, Potchefstroom, South Africa

Abstract: This study evaluates the performance of the isiNdebele part of speech tagger developed by the National Centre for Human Language Technologies as part of Nguni core technologies. A sample of 522 words from government documents and isiNdebele literary works was randomly selected. A mixed-methods approach was utilised to analyse the data. The raw data were automatically processed using the tagger, and the outputs were compared against the gold standard to calculate the tagger's accuracy. Nouns attained an accuracy of 86%, verbs 66%, adverbs 59%, pronouns 90%, adjectives 14%, conjunctions 33%, copulatives 83%, relatives 50%, possessives 90%, demonstratives 71%, while it was 0% for ideophones, interjections, prepositions, question words and auxiliary verbs. Recall and precision were calculated using Python 3.0, enabling the researchers to determine the F1 score. Nouns achieved a recall of 0.86, precision of 0.55, and F1 score 0.67, verbs 0.66, 0.7 and 0.68, relatives 0.5, 0.46 and 0.48, adverbs 0.63, 0.86 and 0.73, possessives 0.9, 0.56 and 0.69, demonstratives 0.71, 0.86 and 0.78, adjectives 0.14, 0.67 and 0.23, pronouns 0.9, 0.95 and 0.92 copulatives 0.83, 1.0 and 0.91 and conjunctions 0.36, 0.83 and 0.5 respectively. These findings underscore the importance of improving the isiNdebele part of speech tagger.

Keywords: accuracy; F1 score; part of speech tagger; precision; recall

CORRESPONDENCE

Email:
Muzi.Matfunjwa@nwu.ac.za

EDITORIAL DATES

Received: 21 July 2025
Revised: 22 December 2025
Accepted: 09 January 2026
Published: 18 February 2026

DOI: https://doi.org/10.51415/ajims.v8i1.3170

Introduction

IsiNdebele is perceived as the youngest language in academia, having been formally introduced as an academic subject in schools in 1985, and is amongst the least spoken languages in South Africa (Skhosana, 2003; Mnguni, 2004; Mabena, 2020; Mngunı, 2025). This language is also classified as under-resourced in the field of Human Language Technology (HLT) and Natural Language Processing (NLP) because it lacks adequate tools and technologies (Eiselen & Puttkammer, 2014; Mlambo & Matfunjwa, 2024).

Over the years, the South African government has taken initiatives to support and develop indigenous languages by funding various research and technology-driven projects through the Department of Sport, Arts, and Culture, the Department of Science, Innovation and Technology, and the National Research Foundation (Grover et al., 2011). For example, as part of the funded projects, the National

Centre for Human Language Technologies (NCHLT) developed speech and text-processing tools for South African indigenous languages, such as tokenisers, lemmatisers, sentencisers, part of speech (POS) taggers, and morphological analysers (Eiselen & Puttkammer, 2014). However, due to the limited availability of annotated data, the performance of these HLT technologies remains suboptimal compared to other tools developed for widely spoken languages (Koehn & Knowles, 2017; Van Zaanen, et al., 2020; Skosana & Mlambo, 2021). Consequently, an NCHLT follow-up project to enhance and improve core HLT tools for Nguni languages, which underperformed compared to other South African languages in the initial project, was conducted (Du Toit & Puttkammer, 2021). Despite the advancements of the HLT tools, their adoption, practical use and evaluation in indigenous languages remain low, due to a lack of awareness and understanding among users (Mlambo & Matfunjwa, 2025). Therefore, this paper evaluates the performance of the isiNdebele POS tagger (n.d.) (NDE POS tagger), using the metrics accuracy, recall, precision and F1 score. The NDE POS tagger being evaluated is part of the core technologies for Nguni languages.

Literature review

Kumar and Josan (2010) evaluated POS taggers for morphologically rich languages, including Hindi, Punjabi, Malayalam, Bengali and Telugu. Accuracy was used as a primary metric for evaluation. These authors noted that approaches such as Rule-based, Maximum Entropy (ME) model, Conditional Random Field (CRF), Brill transformation rule-based Learning (TBL) and Hidden Markov Model (HMM), as well as Support Vector Model (SVM), were used in the languages' POS tagging. The HMM was employed on Malayalam, with 1400 words tagged, and it achieved 90% accuracy, while for SVM based tagging accuracy for 20 000 words was 63%, for 100 000 words, it was 86%, and for 180 000 words, it was 94%, indicating that it outperformed the HMM. In Bengali, the HMM, ME, SVM and CRF were utilised. Supervised, semi-supervised and semi-supervised with a Morphological analyser were suggested for the HMM and ME. The accuracy levels attained by the Supervised HMM utilising MA and Suffix Information, the Semi-supervised HMM employing MA and Suffix Information, and the ME model with MA and Suffix Information were recorded at 88.75%, 87.95%, and 88.41%, respectively. The SVM based POS taggers achieved an accuracy of 86.94%, while CRF reached 90.3% accuracy. A total of 66 900 words in Hindi were utilised as a training and testing corpus. Four taggers were developed based on HMM, ME, CRF, and a morphology-driven approach, resulting in aggregated accuracies of 93.05%, 89.34%, 82.67%, and 93.45%. In Punjabi, a Rule-based POS tagging was used, and an accuracy of 80.29% was achieved, including unknown words and 88.86% excluding unknown words. In Telugu, the Rule-based approach, TBL and ME were used in the POS tagging, achieving an accuracy of 98%, 90% and 81.78 %, respectively.

Eiselen and Puttkammer (2014), and Du Toit & Eiselen (2017) evaluated the accuracy of POS taggers for ten South African official languages, excluding English and South African Sign Language. Eiselen and Puttkammer (2014) focused on developing linguistic tools such as tokenisers, sentencisers, lemmatisers, POS taggers, and morphological decomposers for these languages. In the development stage, they collected unannotated monolingual corpora for each language exceeding one million tokens. These corpora were sourced from South African government websites, official documents, scientific articles, magazines, literary works, and news publications, with an addition of 5,000 annotated tokens prepared for each language to serve as testing data. At the beginning of the project, Eiselen and Puttkammer (2014) set an accuracy benchmark of 80% for POS tagging. Their results revealed that all ten languages surpassed this threshold, with disjunctively written languages performing better than conjunctively written ones. For instance, Setswana and Sepedi achieved the highest accuracy at 96.02% and 96.00%, respectively. In comparison, Siswati had the lowest accuracy of 82.08%. Similarly, Du Toit and Eiselen (2017) assessed the accuracy and speed of five POS taggers: HunPos, Mate Tools, NLP4J, OpenNLP, and the Stanford POS tagger, for the ten South African official languages. They determined which tagger offered the best balance between accuracy and processing speed for large data annotation. To assess accuracy, they used monolingual POS annotated corpora for training and evaluation. The speed of each tagger was tested using unannotated monolingual NCHLT corpora. The findings indicated that Mate POS was the most accurate, achieving 88.48% accuracy across all ten languages, with the least efficiency, as its average speed ranked among the lowest. Du Toit and Eiselen (2017) concluded that a high accuracy of a POS does not automatically translate to its efficiency. On the other hand, HunPos had the lowest accuracy of 80.67% but was the most efficient. The NLP4J tagger, despite having an average accuracy of 85.95%, outperformed the others in terms of processing speed. To further enhance NLP4J's accuracy, Du Toit and Eiselen (2017) trained it using the Adagrad Mini-batch algorithm for 40 epochs, incorporating L1 regularisation and a feature cut-off value of two. These refinements improved its performance, making it the optimal choice when considering both speed and accuracy. Both studies

highlighted the continued need for evaluating POS taggers for South African languages and the importance of further research to enhance these tools, ensuring their optimal performance.

Jahara et al. (2020) conducted a comprehensive evaluation of 16 POS tagging techniques for Bengali, comprising 8 stochastic methods and 8 transformation-based methods. The stochastic approaches included bigram, unigram, trigram, and combined n-gram models, unigram with bigram, and unigram with bigram and trigram. They also used Hidden Markov Models, Conditional Random Fields (CRF), and Trigrams 'n' Tags. The transformation-based approaches were based on Brill's tagger, implemented in combination with the stochastic models. They utilised a tagged corpus developed by the Linguistic Data Consortium, consisting of 7390 annotated sentences, approximately 115 000 tokens, and 22 330 unique word forms. A comparative analysis of tagging techniques was carried out using two tagsets, namely a fine-grained 30-tagset and a reduced 11-tagset, with performance evaluated in terms of tagging accuracy. The results indicated that the Brill + CRF model achieved the highest accuracy, recording 84.5% for the 30-tagset and 91.83% for the 11-tagset. In contrast, trigram-based methods demonstrated relatively poor performance across both tagsets. Furthermore, the findings showed a consistent increase in tagging accuracy when the size of the tagset was reduced. Overall, the study concludes that POS tagging approaches that integrate both statistical techniques and linguistic rule-based knowledge tend to yield higher performance for Bengali.

Mathe and Eiselen (2021) conducted a quantitative analysis of the Sesotho sa Leboa NCHLT POS annotated dataset and compared the tagging accuracy of this tagger with the Centre for Text Technology (CTexT) POS tagger. This study used datasets of 65,857 tokens for training the tagger and 7,153 tokens for evaluation. The distribution of POS in both the training and evaluation datasets showed that verbs consisted of 12.15% of the training data and 11.79% of the evaluation data. Nouns were the most common category, with 19.84% of the training data and 17.94% of the evaluation data. In evaluating the performance of the two taggers, the CTexT POS tagger had higher accuracy, achieving 94.18%, compared to the NCHLT POS tagger, which attained 88.40%. A major source of errors in the NCHLT tagger was the mistagging of class 10 possessive concords as verbs, with an error rate of 96.51%, where 83 instances out of 86 were tagged incorrectly. Errors in noun classification were also prominent in noun classes 1a, 3, 4, 6, 7, 9, and 10, which collectively accounted for 33.85% of the tagging inaccuracies. Auxiliary verbs exhibited the fewest errors, with a misclassification rate of 3.25%. The CTexT POS tagger encountered challenges in accurately tagging nouns in class 1a, which were frequently mislabeled as foreign words, leading to 50% of the errors in this POS tagger. The class 9 demonstrative concord had the fewest misclassification errors, with a rate of 2.64%. Mathe and Eiselen (2021) concluded that improvements in POS tagging accuracy could be achieved by incorporating n-gram probabilities and gazetteers to enhance the recognition of foreign words and proper names. Expanding the training dataset would also improve the classification of noun classes and overall tagging performance. These scholars stated that these enhancements could contribute to refining the accuracy and reliability of Sesotho sa Leboa POS taggers.

Pannach et al. (2021) and Du Toit and Puttkammer (2021) conducted evaluations of POS taggers for Nguni languages, recognising their complex morphological structures, conjunctive writing systems, and agglutinative nature. Pannach et al. (2021), through the Nguni Languages POS Tagging ($NLAPOST_{2021}$) Shared Task, investigated the performance of POS tagging models trained on limited annotated data. Du Toit and Puttkammer (2021) focused on creating, annotating, and evaluating NLP technologies, including POS taggers for Nguni languages. Their objective was to improve the rule-based NLP technologies, which were initially developed in the NCHLT 2013 project, where Nguni languages had underperformed compared to other South African official languages. These improvements included increasing the dataset by 50,000 tokens from South African government websites and documents, which were automatically filtered, annotated, and verified by linguistic experts. The annotation process classified POS tags into 20 simplified tagsets, which covered basic word classes and 107 full tagsets, which included syntactic and functional details of the data. They used 90% of the dataset for training and 10% for evaluation; their study demonstrated increased accuracy in the full POS tagset tagger across Nguni languages, with isiNdebele improving from 82.57–85.25%, Siswati from 82.08–87.46%, isiXhosa from 84.18–93.99%, and isiZulu from 83.83–88.60%. Pannach et al. (2021) used Long Short-Term Memory (LSTM) networks to process sequences bidirectionally, capturing contextual information from both forward and backward directions. These authors also integrated Conditional Random Fields (CRFs) to improve the POS tagger's predictive accuracy, while the Hidden Markov Model (HMM) served as the baseline model. Datasets from the CTexT were split into 90% training and 10% testing sets. The baseline model had the lowest accuracy among the tested models, with the accuracy of 75% for isiNdebele, 77% for isiXhosa, 76% for isiZulu, and 72% for Siswati. The CRF model improved the accuracy to 86%, 90%, 87%, and 85%, respectively. The LSTM

model had the highest accuracy, with isiNdebele at 91%, isiXhosa at 95%, isiZulu at 92%, and Siswati at 90%. Both studies acknowledged the linguistic complexity of Nguni languages and demonstrated that, despite being low-resource languages, their POS tagging performance can be improved when appropriate models and well-structured datasets are used.

Malema et al. (2017) and Dibitso et al. (2019) developed and evaluated POS tagger models for Setswana, each using different approaches to improve their accuracy. The scholars posited that they developed these models upon recognising the challenges posed by the disjunctive writing system, lexical ambiguity, and limited linguistic resources to improve POS tagging for this language. Malema et al. (2017) implemented a finite state approach to POS tagging, focusing on identifying relative and possessive concords. Their model was developed in Java and incorporated 2D arrays, a dictionary, and a morphological analyser to facilitate state transitions. They used 10 pages of the Setswana document as testing data, and the tagger correctly identified 65 out of 77 manually tagged relatives, achieving an 84% accuracy for this category. The POS tagger was also able to correctly identify 89 out of 111 possessive concords, resulting in an 80% accuracy for possessives. Overall, the POS tagger achieved 97% accuracy in detecting relatives and 82% accuracy for possessives. However, it had challenges in determining where both relative and possessive constructions end. Dibitso et al. (2019) used supervised machine learning, the Support Vector Machine (SVM) Tool, to develop their POS tagger. They trained their model with a corpus of 65,784 annotated Setswana tokens, which were categorised into 128 morphosyntactic tags. To ensure a reliable evaluation, they used 60% of the data for training and 40% for testing. The SVM approach addressed key challenges in classification and regression, improving the robustness, flexibility, and efficiency of POS tagging. This SVM was built using four models, 0, 1, 2, and 4, that were tested using six tagging strategies. Among these, strategy 6 with model 4, which performed sentence-level tagging, achieved the highest accuracy of 92.16%, while strategy 2 with models 0 and 2, which focused on tagging words in unseen contexts, had the lowest accuracy of 90.37%. Both these studies stress the need for high-quality annotated corpora, well-defined tagsets, and effective preprocessing tools to improve POS tagging in under-resourced, disjunctively written languages like Setswana.

The consulted literature demonstrates that POS taggers have been evaluated for accuracy in South African languages and in other morphologically complex languages abroad. However, to the researchers' knowledge, no study has assessed the performance of the NDE POS tagger from the core technologies for Nguni languages using the metrics accuracy, recall, precision, and F1 score. Chiche and Yitagesu (2022) posit that accuracy, recall, precision, and F1 score are the most utilised metrics for assessing and validating POS tagging. This assertion indicates that accuracy alone is insufficient for a holistic evaluation of a POS tagger's performance. Therefore, the combination of these metrics in evaluating the NDE POS tagger provides a comprehensive and objective assessment, eliminating bias. As a result, this study addresses the gap in evaluating the NDE POS tagger using these metrics.

Methodology

This study applied mixed methods to evaluate the efficiency of the NDE POS tagger. Mixed methods research combines quantitative and qualitative methodologies in one investigation to offer a comprehensive understanding of a research problem compared to using a single method (Creswell & Plano Clark, 2011; Almeida, 2018). This study employed calculations of accuracy, recall, precision, and F1 score, and provided their interpretation, allowing the mixed method approach to address the research objectives. Random sampling was used to collect data from the *Presidency Republic of South Africa (n.d.)* website available at https://www.gov.za/nr/speeches/2025StateOfTheNation, the novel *Mbala Ngubaba* by Skhosana (1994), and the book *Ukukhamba Kubona* by Nyamunda (2019), resulting in a sample of 522 words. This sampling method was used to ensure the validity of the data, representing parts of speech found in isiNdebele. This data, although small, is suitable for evaluating the performance of the NDE POS tagger across all word categories in this language. The data (words) were manually analysed to determine their word categories, and each word was labelled with the accurate part of speech to create a gold standard. The NDE POS tagger was downloaded from the South Africa Centre for Digital Language Resources (SADiLaR) website at https://repo.sadilar.org/handle/20.500.12185/548 and was used to automatically analyse the words from the novel. A notepad was used to create a text format (.txt) document containing the words to be analysed, as the NDE POS tagger can only process documents in this format. This was then uploaded to the NDE POS tagger presented in Figure 1, and the results, in which POS labels known as tagsets were assigned to each word processed. The tagsets presented in Table 1 were used in the output of this tagger. These results from the tagger were then compared manually against the gold standard to calculate the accuracy of

each POS and to find the POS taggers' overall accuracy. Manual tagging has been fundamental to POS tagging initiatives and continues to serve as a crucial reference point, which is a gold standard for assessing automated systems (Palmer et al., 2005). Therefore, the gold standard used in this study was developed by two language annotators to triangulate it, assuring reliability and eliminating bias. Employing investigator triangulation to validate and interpret data enhances the credibility of research studies (Polit & Hungler, 1995; Thurmond, 2001).

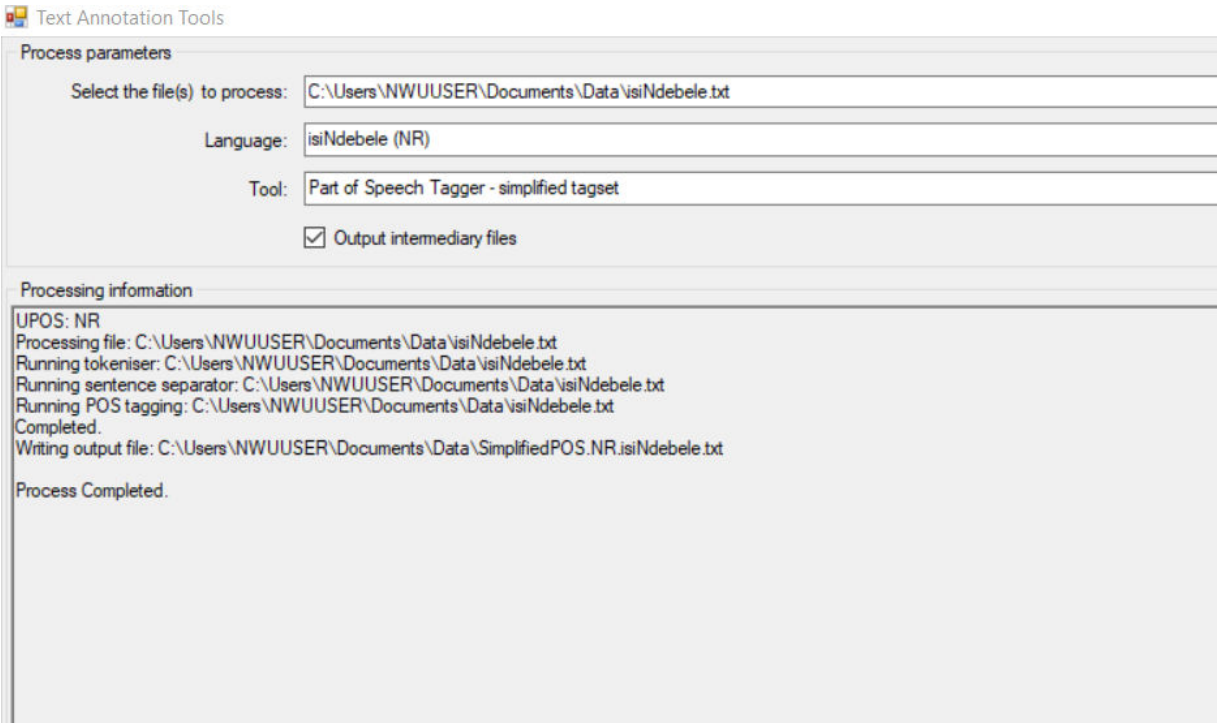

Figure 1. NDE POS Tagger

Table 1. Tagsets used in the output of the NDE POS tagger

PUNC = Punctuation	**PROEMP = Emphatic pronoun**
ABBR = Abbreviation (incl. acronyms)	**PROQUANT** = Quantitative pronoun
ADJ = Adjective (incl. enumerative)	**REL** = Relative
ADV = Adverb	**V** = Verbal
CDEM = Class-indicating demonstrative	**VAUX** = Auxiliary verb
CONJ = Conjunction	**N** = Noun
COP = Copulative	**INTER** = Question word
FOR = Foreign	**NPP** = Place and brand name
IDEO = Ideophone	**POSS** = Possessive
INT = Interjection	**NUM** = Numerative

Source: (Puttkammer & Gaustand, 2021)

The data was then prepared for Python analysis to calculate recall and precision. It was put in an Excel spreadsheet to create a CSV file, in three columns, namely, column A with the head name POS, column B with the predicted word category by the NDE POS tagger (prediction) and column 3 with the gold standard POS (truth). The CSV file was then imported into Python using the Pandas package, in which recall and precision were calculated as presented in Figure 2. Thereafter, the F1 scores were calculated.

a

```
calculate recall

[10]: # function for calculating recall
      def calculate_recall (truth_tag, prediction_tag, pos_tag):
          """
          Calculate the recall for a specific part of speech tag.
          Parameters:
          truth_tag (list): The ground truth POS tags.
          prediction_tag (list): The predicted POS tags by the tagger.
          pos_tag (str): The part of speech tag to calculate recall for.

          Returns:
          float: The recall value for the specified POS tag.
          """

          tp = sum(1 for truth, pred in zip(truth_tag, prediction_tag) if truth == pos_tag and pred == pos_tag)
          fn = sum(1 for truth, pred in zip(truth_tag, prediction_tag) if truth == pos_tag and pred != pos_tag)

          if tp + fn == 0:
              return 0.0

          recall = tp/(tp + fn)
          return recall
```

b

```
Calculating precision

[12]: # function for calculating precision
      def calculate_precision (truth_tag, prediction_tag, pos_tag):
          """
          Calculate the precision for a specific part of speech tag.
          Parameters:
          truth_tag (list): The ground truth POS tags.
          prediction_tag (list): The predicted POS tags by the tagger.
          pos_tag (str): The part of speech tag to calculate recall for.

          Returns:
          float: The precision value for the specified POS tag.
          """

          tp = sum(1 for truth, pred in zip(truth_tag, prediction_tag) if truth == pos_tag and pred == pos_tag)
          fp = sum(1 for truth, pred in zip(truth_tag, prediction_tag) if truth != pos_tag and pred == pos_tag)

          if tp + fp == 0:
              return 0.0

          precision = tp/(tp + fp)
          return precision
```

Figure 2. (a)Recall calculation (b)Precision calculation

Results and discussion

Accuracy

Accuracy is the percentage of correctly tagged parts of speech compared to the gold standard of annotated words. The accuracy for the NDE POS was calculated using this formula: $\text{accuracy} = \frac{\text{Number of correctly tagged words}}{\text{Total number of words}} \times 100$. Figure 3 show a glimpse of the NDE POS tagger results, in which some parts of speech were not tagged accurately by this tagger. The analysed data contained 9 ideophones, 8 interjections, 1 preposition, 2 question words and 3 auxiliary verbs, with zero tagged accurately. However, the NDE POS was able to tag accurately 86 out of 130 verbs which is 66%, 86 out of 100 nouns which is 86%, 73 of 123 adverbs which is 59%, 18 of 20 pronouns which is 90%, 2 of 14 adjectives which is 14%, 5 of 14 conjunctions which is 33%, and 5 of 6 copulatives which is 83%. Moreover, 13 out of 26 relatives were also correctly tagged, which is 50%, 44 out of 49 possessives, which is 90%, and 12 out of 17 class-indicating demonstratives, which is 71%. Overall, out of the 522 words, 319 were tagged correctly, which shows an accuracy performance of 61%, while 203 words were tagged incorrectly, which gives an inaccuracy of 39%.

a

```
File  Edit  View
.       PUNC
He      N
.       PUNC
.       PUNC
nawubiza        ADV
abantu  N
kangaka         ADV
nje     ADV
,       PUNC
iinkomo         N
nezimvezo       ADV
uzozithathaphi      N
?       PUNC

Ungathi         N
isilarha        N
sakho POSS
sikghona        N
ukuceda         N
iveke N
tomu    N
singananyama        V
uthi    V
Ln 1, Col 1   100%   Windows (CRLF)   UTF-8
```

b

```
File  Edit  View

Usize N
ubatjele    V
kuhle ADJ
bona    PRO
bangazitshwenyi     POSS
bahle V
behla POSS
benyuka     POSS
baya    V
kezinye     POSS
iinlarha    N
.       PUNC
Ubatjele    N
bona    PRO
ikomo N
nemvu ADV
zizokuthengeka      N
eLAL'UDLILE N
.       PUNC

Kanti-ke    N
ungalibali  N
Ln 1, Col 1   100%   Windows (CRLF)   UTF-8
```

Figure 3. (a)Inaccurately tagged ideophones and interjections (b)Some inaccurately tagged verbs

From Figure 3, it emerges that the interjection *He* and the ideophone *tomu* were inaccurately tagged as nouns. Some verbs that begin with class 1 and 1a subject concord /u-/, such as *usize* (help) and *ubatjele* (tell them), were tagged as nouns rather than verbs. Unexpectedly, the same word *ubatjele* used in another context was accurately tagged as a verb. These figures show the limitations of the NDE POS tagger.

Recall

Recall is one of the prevalent metrics used for POS tagging performance. Chiche and Yitagesu (2022, p. 10) describe recall as "the ratio of all samples correctly tagged as tagged to all the samples that are tagged by an expert (aka a Detection Rate)". Therefore, recall determines how well the POS tagger recognises all specified parts of speech in the sentences. It is calculated using the formula: $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$, where True Positive (TP) denotes the accurately tagged words as annotated by language experts, and False Negative (FN) refers to instances where the POS tagger failed to tag a specific POS correctly (Chiche & Yitagesu, 2022). The results of the calculated recall are shown in Figure 4.

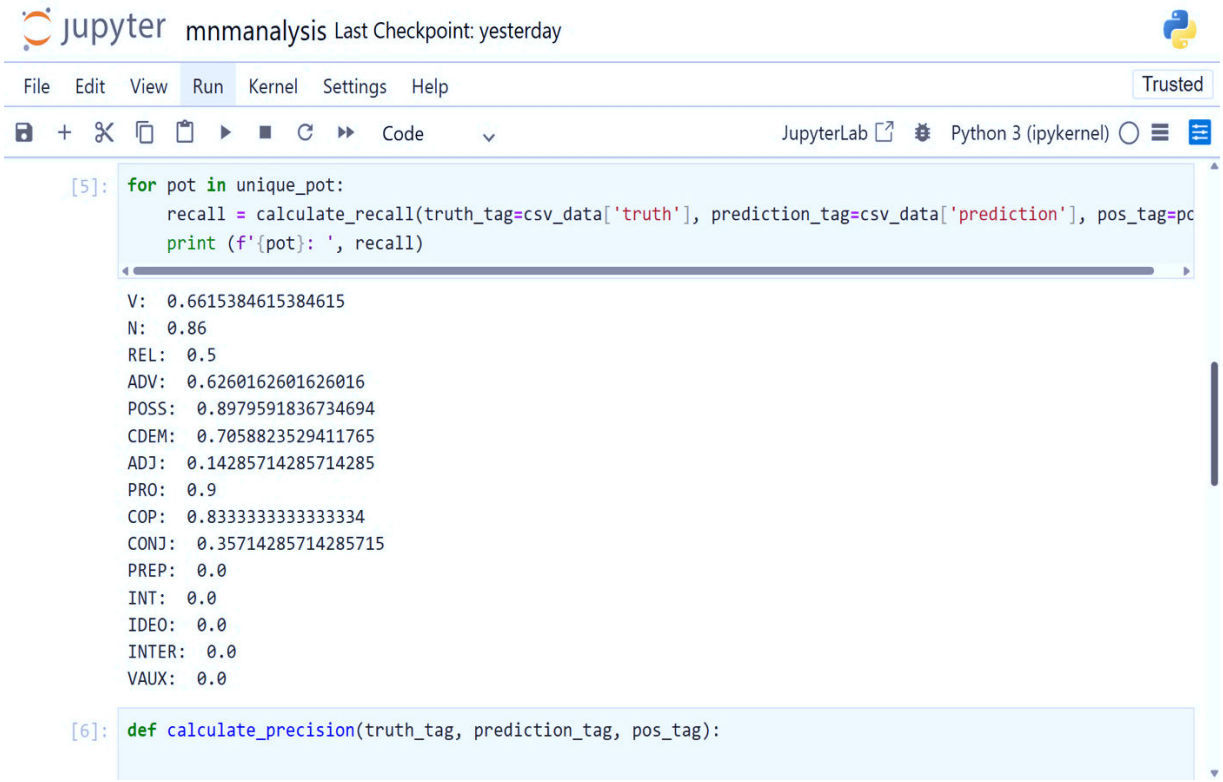

Figure 4. Recall results

In Figure 4, the result shows that the calculated recall for verbs tagged is 0.661, nouns is 0.86, relatives is 0.5, adverbs is 0.626, possessives is 0.897, demonstratives is 0.705, adjectives is 0.142, pronouns is 0.9, copulatives is 0.833, and conjunctions is 0.375. While for prepositions, interjections, ideophones, question words and auxiliary verbs, the recall is 0.0.

Precision

Precision is another metric mostly used to validate the performance of Machine Learning and Deep Learning tools, such as POS taggers, and it is the ratio of accurately identified POS within all tagged words (Chiche & Yitagesu, 2022). This metric is calculated using the formula: $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$, in which False Positive (FP) refers to words tagged inaccurately by the POS tagger (Chiche & Yitagesu, 2022). The results for the calculated recall are presented in Figure 5.

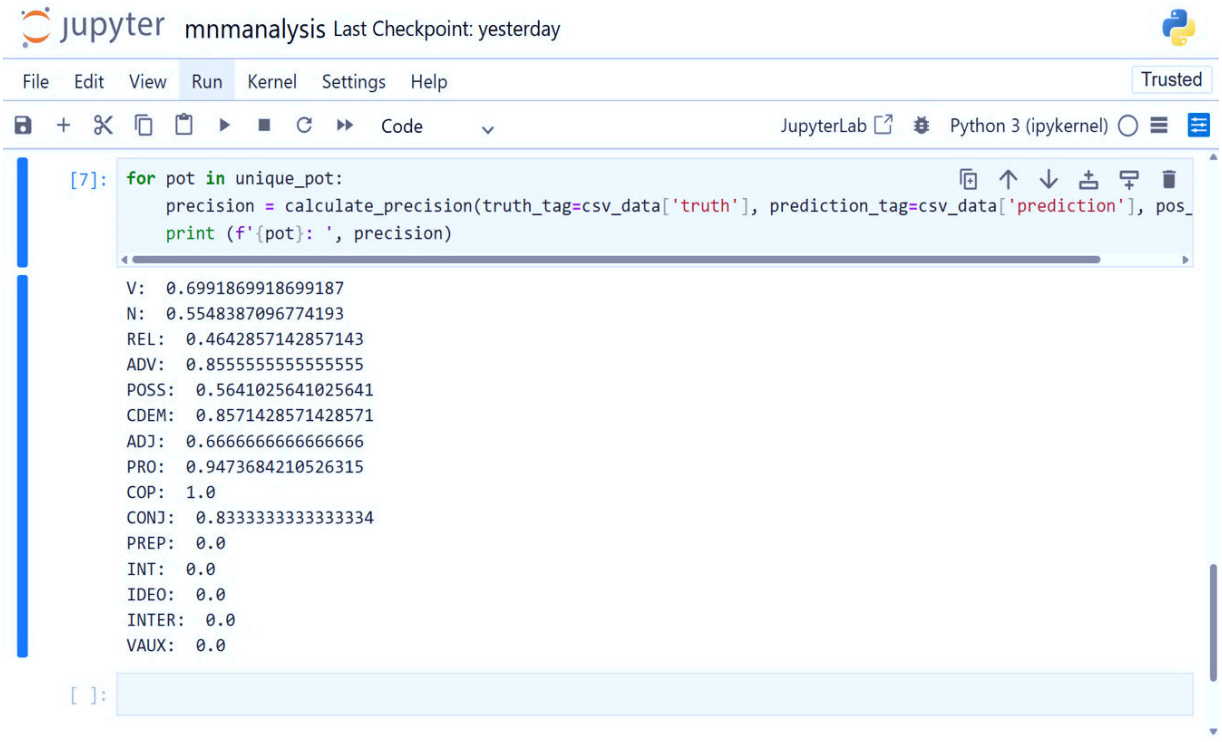

Figure 5. Precision results

In Figure 5, the result shows that the calculated precision for verbs tagged is 0.699, nouns is 0.554, relatives is 0.464, adverbs is 0.855, possessives is 0.564, demonstratives is 0.857, adjectives is 0.666, pronouns is 0.947, copulatives is 1.0, conjunctions is 0.833, while for prepositions, interjections, ideophones, question words and auxiliary verbs, it is 0.0.

In Table 2, the overall results of the calculated precision against the recall show that the weighted precision of 0.674 (67.4%) is slightly higher than the weighted recall of 0.667 (66.7%). This overall slightly high precision indicates that the NDE POS tagger is more conservative, which means that it tags a lesser number of words with more accuracy while omitting some parts of speech, such as the ideophones and interjections, thus producing a few mistakes. However, the weighted results for recall and precision do not determine the entire performance. To have a conclusive performance of the NDE POS tagger, the recall and precision results must be compared per POS tagging. For example, the recall of 0.66 for verbs indicates that the NDE POS tagger correctly tagged 66% of all true verbs and missed 34%. The precision of 0.70 means that of all the words the tagger predicted as verbs, 70% were correct, while 30% were false positives. The F1 score of 0.68 shows a good balance between precision and recall, indicating a strong overall performance of the NDE POS tagger.

Table 2. Recall, Precision and F1 Score for Each Part of Speech

Part of Speech	**Recall**	**Precision**	**F1 Score**
Verbs	0.66	0.70	0.68
Nouns	0.86	0.55	0.67
Relatives	0.5	0.46	0.48
Adverbs	0.63	0.86	0.73
Possessives	0.90	0.56	0.69
Demonstratives	0.71	0.86	0.78
Adjectives	0.14	0.67	0.23
Pronouns	0.9	0.95	0.92
Copulatives	0.83	1.0	0.91
Conjunctions	0.36	0.83	0.5

(*Continued*)

Table 2. (Continued)

Part of Speech	Recall	Precision	F1 Score
Prepositions	0.0	0.0	0.0
Interjections	0.0	0.0	0.0
Ideophones	0.0	0.0	0.0
Question words	0.0	0.0	0.0
Auxiliary Verbs	0.0	0.0	0.0
Weighted	0.667 (66.7%)	0.674 (67.4%)	

The recall of 0.86 for nouns indicates that the NDE POS tagger correctly identified 86% of all true nouns and missed only 14%. The precision of 0.55 indicates that of all the words tagged as nouns, only 55% were true nouns, with 45% being false positives. The F1 score of 0.67 shows a moderately good overall performance of the tagger despite the low precision. While the recall for relatives is 0.5, suggesting that the NDE POS tagger correctly identified 50% of all true relatives in the data, missing the other 50% of the relatives. The precision of 0.46 indicates that of all the words the tagger predicted as relatives, only 46% were true relatives. This precision shows a high false-positive rate. The F1 score of 0.48, which is 48%, indicates an overall poor performance of the tagger in tagging relatives.

The recall for adverbs of 0.63 conveys that the NDE POS tagger correctly identified 63% of all adverbs in the texts, while 37% are the adverbs that it could not tag. The precision for adverbs of 0.86 signifies that 86% of instances where the tagger labelled words as adverbs were correct, but 14% refers to cases where it mistakenly labelled non-adverbs as adverbs. The F1 score of 0.73, which is 73%, shows good overall performance of the NDE POS tagger for adverb tagging. On possessives, the POS tagger attained a recall of 0.9 and a precision of 0.56. The recall of 0.9 means that the tagger successfully tagged 90% of all true possessives while missing only 10%. The precision for possessives is 0.56, which means that of all the words the tagger predicted as possessives, only 56% were correct, while 44% were false positives. The high recall of 0.9 reveals that the tagger is effective at identifying most possessives in the text, missing only a small fraction. Meanwhile, the low precision of 0.56 indicates that the tagger often mislabels non-possessive words as possessives. The F1 score of 0.69 reflects a good, but imperfect overall performance of the tagger.

The recall of 0.71 on demonstratives means that the NDE POS tagger identified 71% of the demonstratives in the dataset and missed 29% demonstratives. A precision of 0.86 means 86% of the words the tagger labelled as demonstratives were true positives, while 14% were incorrectly labelled as false positives. An F1 score of 0.78 indicates a strong overall performance, combining good recall with impressive precision. This shows that the POS tagger is efficient in tagging demonstratives. Meanwhile, for adjectives, the recall of 0.14 means that the tagger correctly tagged only 14% of true adjectives in the data, while the other 86% were either untagged or incorrectly tagged as adjectives, indicating a poor performance of the tagger. The precision of 0.67 shows that of the words labelled as adjectives by the POS tagger, 67% were actually adjectives in the gold-standard data, indicating few false positives. An F1 score of 0.23 indicates a poor overall performance for adjective tagging.

For pronouns, the POS tagger obtained a recall of 0.9 (90%), which means that the tagger detected almost all pronouns in the dataset. There were a few false positives and false negatives in the tagging of the pronouns. The precision of 0.95 (95%) means that almost all the words the tagger tagged as pronouns are correctly identified as a pronoun. The F1 score of 0.92 shows that the tagger achieved an outstanding accuracy for pronouns, with fewer errors in either metric. This score indicates a perfect balance between precision and recall, reflecting a very good performance.

In copulatives, the NDE POS tagger achieved a recall of 0.83. This means the tagger identified 83% of true copulatives, which is decent but not excellent. It missed 27% of actual copulatives. For precision, it achieved 1.0, indicating that the tagger is extremely reliable when it labels a word as a copulative. There are no false positives, and every predicted copulative is correct. This suggests that the tagger is highly conservative and confident when assigning the copulative tag. The F1 score is 0.91, which reflects a good balance between precision and recall, demonstrating a strong performance of the tagger in copulative tagging. On the other hand, the recall for conjunctions is 0.36, which means the NDE POS tagger correctly identified only 36% of all true conjunctions, missing 64% of them. This low recall score indicates that the tagger failed to identify most conjunctions in the data. The precision is 0.83, indicating that of all the words tagged as conjunctions, 83% were true conjunctions.

This reflects a high prediction rate in conjunction. The F1 score of 0.5 shows a moderately overall performance, despite the low recall bringing down this F1 score, in which the tagger favours correctness over coverage.

For prepositions, interjections, ideophones, question words and auxiliary verbs, the recall is 0.0, the precision is 0.0, and the F1 score is 0.0, indicating a failure by the NDE POS tagger to identify or predict any of these parts of speech correctly. These results show that every true preposition, interjection, ideophone, question word and auxiliary verb was either missed or tagged as another word category. The tagger's F1 score of 0.0 indicates that it fails to strike any balance between recall and accuracy, as both are 0.0, producing no usable output for these parts of speech. These statistical results indicate that the NDE POS tagger encounters difficulties in accurately tagging prepositions, interjections, ideophones, question words, and auxiliary verbs. This notable underperformance underscores the need for additional training data containing these parts of speech to enhance the tagger's overall performance.

Conclusion

The NDE POS tagger achieved an accuracy of 86% for nouns, 66% for verbs, 59% for adverbs, 90% for pronouns, 14% for adjectives, 33% for conjunctions, 83% for copulatives, 50% for relatives, 90% for possessives, and 71% for demonstratives. It scored 0% for ideophones, interjections, prepositions, question words and auxiliary verbs. Nouns also had a recall of 0.86, precision of 0.55, and F1 score 0.67, verbs 0.66, 0.7 and 0.68, relatives 0.5, 0.46 and 0.48, adverbs 0.63, 0.86 and 0.73, possessives 0.9, 0.56 and 0.69, demonstratives 0.71, 0.86 and 0.78, adjectives 0.14, 0.67 and 0.23, pronouns 0.9, 0.95 and 0.92, copulatives 0.83, 1.0 and 0.91 and conjunctions 0.36, 0.83 and 0.5 respectively. These results indicate that the NDE POS tagger was effective at tagging pronouns, copulatives, demonstratives, adverbs, verbs, and nouns. While it performed poorly on tagging relatives and adjectives. The F1 score of 0.0 for ideophones, interjections, prepositions, question words and auxiliary verbs indicates that the tagger could not tag correctly any of these word categories. These findings highlight the need to improve the NDE POS tagger's performance, especially in these word categories where it underperformed. This can be achieved by further training the NDE POS tagger with data not only sourced from government documents but from other sources such as literary works, social media and Wikipedia. Therefore, the study recommends using training and testing data from diverse sources to ensure better accuracy of the NDE POS tagger. This study highlights that while this tagger is efficient in tagging some parts of speech, it still faces challenges in achieving a high level of recall and precision on others.

Declarations

Interdisciplinary Scope: This article demonstrates an interdisciplinary scope by integrating linguistics with computational analysis to evaluate the performance of isiNdebele part of speech tagger using the metrics accuracy, recall, precision and F1 score.

Author Contributions: Conceptualisation (Matfunjwa and Skosana); literature review (Skosana); methodology (Matfunjwa); analysis (Matfunjwa and Skosana); investigation (Matfunjwa); drafting and preparation (Matfunjwa and Skosana); review and editing (Matfunjwa and Skosana). All authors have read and approved the published version of this article.

Conflict of Interest: No potential conflict of interest was reported by the authors.

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

Availability of Data: All relevant data sources are included in the article. However, more information is available upon reasonable request from the corresponding author(s).

References

Almeida, F. (2018). Strategies to perform a mixed methods study. *European Journal of Education Studies*, *5*(1), 37–151.

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*(1), 10–25. https://doi.org/10.1186/s40537-022-00561-y

Creswell, J., & Plano Clark, V. (2011). *Designing and conducting mixed methods research*. (2nd ed.). Thousand Oaks.

Dibitso, M. A., Owolawi, P. A., & Ojo, S. O. (2019). *Part of speech tagging for Setswana African language*. 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Vanderbijlpark, South Africa. https://doi.org/10.1109/IMITEC45504.2019.9015871

Du Toit, J. S., & Eiselen, R. (2017, November 29). *A comparative evaluation of open-source part-of-speech taggers for South African languages*. Pattern Recognition Association of South Africa and Robotics and Mechatronics Conference (PRASA-RobMech), Bloemfontein, South Africa.

Du Toit, J. S., & Puttkammer, M. J. (2021). Developing core technologies for resource-scarce Nguni languages. *Information*, *12*(12), 520. https://doi.org/10.3390/info12120520

Eiselen, R., & Puttkammer, M. J. (2014, May 26–31). *Developing text resources for ten South African languages*. Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland.

Grover, A. S., van Huyssteen, G. B., & Pretorius, M. W. (2011). The South African human language technology audit. *Language Resources and Evaluation*, *45*(3), 271–288. https://doi.org/10.1007/s10579-011-9151-2

IsiNdebele Part of Speech Tagger. (n.d.). *Core technologies for conjunctively written South African languages*. Retrieved February 10 2024, from https://repo.sadilar.org/handle/20.500.12185/548

Jahara, F., Barua, A., Iqbal, M. A., Das, A., Sharif, O., Hoque, M. M., & Sarker, I. H. (2020, December). *Towards POS tagging methods for Bengali language: A comparative analysis*. P. Vasant, I. Zelinka, & G. W. Weber (Eds.), International Conference on Intelligent Computing & Optimization (pp. 1111–1123). Springer International Publishing. https://doi.org/10.1007/978-3-030-68154-8_93

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *arXiv Preprint arXiv:1706.03872*.

Kumar, D., & Josan, G. S. (2010). Part of speech taggers for morphologically rich Indian languages: A survey. *International Journal of Computer Applications*, *6*(5), 1–9. https://doi.org/10.5120/1078-1409

Mabena, C. S. (2020). *Terminology development in IsiNdebele: Challenges and solutions* [Master's thesis]. University of Pretoria.

Malema, G., Okgetheng, B., & Motlhanka, M. (2017). Setswana part of speech tagging. *International Journal on Natural Language Computing*, *6*(6), 15–20. https://doi.org/10.5121/ijnlc.2017.6602

Mathe, D. S., & Eiselen, R. (2021). Quantitative analysis of Sesotho sa Leboa part-of-speech taggers. *South African Journal of African Languages*, *41*(3), 259–269. https://doi.org/10.1080/02572117.2021.2010921

Mlambo, R., & Matfunjwa, M. (2024). The use of technology to preserve indigenous languages of South Africa. *Literator*, *45*(1), a2007. https://doi.org/10.4102/lit.v45i1.2007

Mlambo, R., & Matfunjwa, M. (2025). Human language technology tools for indigenous South African languages and their potential use. *Literator*, *46*(1), a2049. https://doi.org/10.4102/lit.v46i1.2049

Mnguni, A. (2004). *The use of the isiNdebele language in the South African public service* [Doctoral dissertation, Tshwane University of Technology].

Mngunı, A. (2025). Translation of selected Zakes Mda's plays into the IsiNdebele language: Perspectives on accuracy and naturalness. *IJASOS-International E-Journal of Advances in Social Sciences*, *10*(30), 415–419.

Nyamunda, G. (2019). *Ukukhamba Kubona*. Nyamunda.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106. https://doi.org/10.1162/0891201053630264

Pannach, F., Meyer, F., Jembere, E., & Dlamini, S. Z. (2021). NLAPOST2021 1st shared task on part-of-speech tagging for Nguni languages. *Journal of the Digital Humanities Association of Southern Africa (DHASA*), *3*(1). https://doi.org/10.55492/dhasa.v3i01.3865

Polit, D. F., & Hungler, B. P. (1995). *Nursing research: Principles and methods*. (6th ed.). Lippincott.

Puttkammer, M., & Gaustand, T. (2021). *Protocol: Part-of-speech tagging (isiNdebele)*. Retrieved February 3 2025, from https://hdl.handle.net/20.500.12185/546

Skhosana, P. B. (1994). *Mbala ngubaba*. Aktua Press.

Skhosana, P. B. (2003). The literary history of isiNdebele. *South African Journal of African Languages*, *23*(2), 111–119. https://doi.org/10.1080/02572117.2003.10587210

Skosana, N. J., & Mlambo, R. (2021). A brief study of the Autshumato machine translation web service for South African languages. *Literator*, *42*(1), a1766. https://doi.org/10.4102/lit.v42i1.1766

The Presidency Republic of South Africa. (n.d.). *Ikulumo yobujamo belizwe (i-SoNA) kamengameli Cyril Ramaphosa*. Retrieved July 22 2025, from https://www.gov.za/nr/speeches/2025StateOfTheNation

Thurmond, V. A. (2001). The point of triangulation. *Journal of Nursing Scholarship*, *33*(3), 253–258. https://doi.org/10.1111/j.1547-5069.2001.00253.x

Van Zaanen, M., Trollip, B., Ramukhadi, P. M., & Mlambo, R. (2020, July 20–25). *Identifying relations between characters in Afrikaans*. Tshivenda, and Xitsonga books [Conference session]. Annual Conference of the Alliance of Digital Humanities Organizations (ADHO), Ottawa, Canada. https://works.hcommons.org/records/dahtp-nx381