Siswati Part of Speech Tagger: A Quantitative Evaluation

Muzi Matfunjwa

South African Centre for Digital Language Resources, North-West University

Muzi.Matfunjwa@nwu.ac.za

Abstract

This article evaluates the performance of the Siswati Text Annotation Tool part of speech (STAT POS) tagger using Recall, Precision and F1 score metrics. A quantitative research design was adopted for analysis, and data was collected through purposive sampling. Python 3 was utilised to calculate the Recall and Precision of the STAT POS tagger outputs. The results show that the Recall for nouns was 0.761, Precision 0.417, with an F1 score of 0.54; for verbs, the Recall was 0.756, Precision 0.798 and F1 score 0.54; for adverbs, the Recall was 0.571, Precision 0.8, and F1 score 0.67; for possessives, the Recall was 0.963, Precision 0.813 and F1 score 0.88. For relatives (REL), the Recall was 0.706, Precision 0.523, and the F1 score 0.60; for class-indicating demonstratives, the Recall was 0.333, Precision 0.25, and the F1 score 0.29; and for copulatives (COP), the Recall was 0.75, Precision 0.75, and the F1 score 0.75. For conjunctions, the Recall was 0.85, the Precision was 0.68, and the F1 score was 0.76; for pronouns, the Recall was 0.563, the Precision was 1.0, and the F1 score was 0.72; for adjectives, the Recall was 0.75, the Precision was 0.75, and the F1 score was 0.75. However, question words, interjections and ideophones received 0.0, highlighting the need for refinement of the STAT POS tagger.

Keywords: Siswati, Part of speech tagger, Recall, Precision, F1 score

1 Introduction

Siswati is one of the under-resourced languages in South Africa, especially in Human Language Technology (HLT) tools (Mlambo and Matfunjwa 2025). Although HLT has existed in South Africa for the past two decades, Siswati still does not receive sufficient attention in this field compared to other Nguni languages such as isiZulu and isiXhosa. In recent years, there has been notable progress in the development of HLT for Siswati, including machine translation, morphological analysers and decomposers, lemmatisers, and part of speech taggers created by various entities such as the Centre for Text Technology (CTexT) at North-West University. Typically, these tools' accuracy and efficacy are lower than those of well-resourced languages because they are developed using inadequate data, mainly from government documents (Bosch, Pretorius & Fleisch 2008; Grover et al. 2011; Heeringa, De Wet and Van Huyssteen 2015; Koehn & Knowles 2017; Mlambo and Matfunjwa 2024). This makes it imperative to evaluate the HLT tools to determine the quality of their output using a different corpus. Therefore, this article evaluates the performance of the Siswati Text Annotation Part of Speech (POS) tagger using Recall, Precision and F1 score.

2 Related Works

Scholars have developed and assessed a variety of POS taggers for the official languages of South Africa using various metrics. Eiselen and Puttkammer (2014) developed and evaluated POS taggers for ten South African languages, excluding English, as part of the National Centre for Human Language Technology project. The taggers were evaluated, revealing superior performance for Afrikaans, Xitsonga, Tshivenda, Sesotho sa Leboa, Setswana, and Sesotho, which use disjunctive writing systems, in contrast to the Nguni languages that employ an agglutinative writing system. This was attributed to the morphological complexity of the Nguni languages and the need for more data to mitigate this. In contrast, using annotated corpora, Du Toit and Puttkammer (2021) developed four text annotation language tools for the Nguni languages: isiNdebele, Siswati, isiXhosa, and isiZulu. These tools included POS taggers, lemmatisers, morphological analysers, as well as morphological decomposers. Previously created tools as part of the NCHLT project were improved upon by this.

Heid, Prinsloo, Faaß, and Taljard (2009) created a noun POS tagger for Northern Sotho, also referred to as Sesotho sa Leboa. The tagger underwent testing, yielding results that indicated a 90% accuracy rate in recognising nouns along with their corresponding noun-class numbers. Mathe and Eiselen (2021) then assessed the performance of two Sesotho sa Leboa POS taggers created by NCHLT and CTexT. The authors determined that the NCHLT tagger exhibited difficulties in accurately tagging noun classes, specifically failing to differentiate between nouns in Class 9 and those in Class 10, resulting in an accuracy of 88.40%. The CTexT tagger achieved an accuracy of 94.18%, indicating minimal errors in the tagging of noun classes. The errors identified in the CTexT tagger were linked to foreign words and proper nouns, which were categorised as either Class 9 nouns or foreign words.

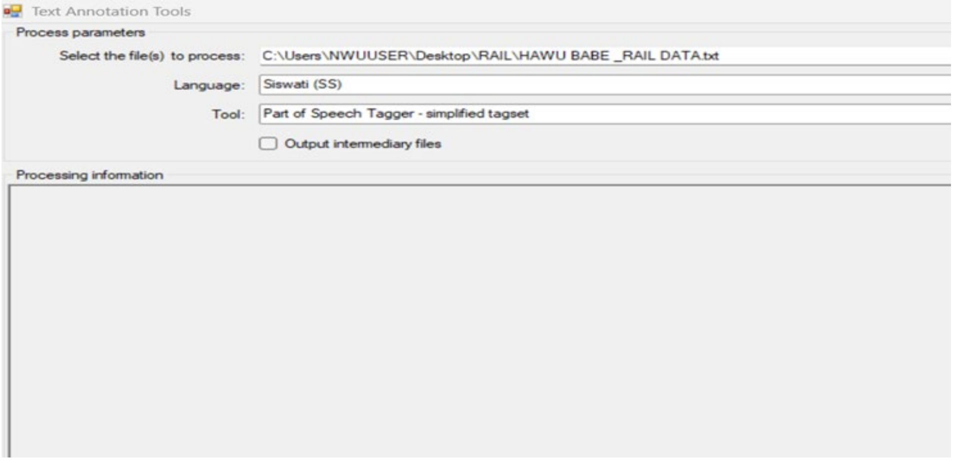
Malema, Okgetheng and Motlhanka (2017) created and assessed a Setswana POS tagger for the identification of relatives. A set of ten pages of Setswana text, comprising 77 relatives, was analysed. The tagger successfully identified 65 of these relatives, resulting in an overall performance rate of 84%. It was determined that the language's disjunctive orthography made it difficult to carry out precise automatic POS tagging and tokenisation of the relatives. Dibitso, Owolawi, and Ojo (2019) employed annotated corpora comprising 65,784 tokens derived from governmental documents to create and train a Setswana POS tagger. The corpus was annotated with tagsets developed following EAGLES guidelines and the tagsets proposed by Taljard, Faaß, Heid, and Prinsloo (2008). The utilised tagsets comprised 128 tags formulated to satisfy the morphosyntactic needs of Setswana text. The developed POS tagger underwent an accuracy assessment, revealing that its accuracy improved with an increase in the number of training words.

Matfunjwa (2025) evaluated and compared the CTexT NCHLT Web Service POS tagger with the Text Annotation Tools POS tagger for Siswati based on accuracy using data sourced from a novel titled *Tinyembeti*. The author found that the overall accuracy of the Text Annotation Tools' POS tagger was higher than that of the CTexT NCHLT Web Service POS tagger in all word categories tagged. It was also determined that these taggers struggled to identify interjections, ideophones, and Class 1a nouns.

From the consulted work, it is evident that no study has evaluated the performance of the Siswati Text Annotation Tools POS tagger based on the metrics: Recall, Precision and F1 score.

3 Methodology

This study is quantitative and uses data obtained from the Siswati drama book titled *Hawu Babe* written by Malindzisa (2005). A sample of 389 tokens, including words, punctuation, and numbers, was purposively collected in a PDF format from the book and converted into a text document. These words were selected because they represent the POS found in Siswati. This data was then manually cleaned to remove numerals, English words, and punctuation, resulting in a clean corpus of 345 Siswati words. This was then uploaded to the Siswati Text Annotation Tool (henceforth, STAT POS) tagger for automatic tagging, as presented in Figure 1. This POS tagger in Figure 1 was obtained from the South African Centre for Digital Resources website at https://repo.sadilar.org/handle/20.500.12185/548. The results received from the STAT POS tagger were then exported to an Excel spreadsheet and stored as a CSV file. A gold standard was created, assigning accurate tags to each word against the predicted POS tags by the tagger. The tagsets used in the POS tagger are presented in Table 1. This resulted in three columns: pos, which contained the words, prediction, which consisted of the tagging by the tagger, and the truth, which was the gold standard being created. Following that, the CSV file was loaded into Jupyter Notebook and then imported into Python 3 using the Pandas package. The Recall and Precision were calculated using Python 3. The formulas for $Recall = \frac{TP}{TP+FN}$ and for $Precision = \frac{TP}{TP+FP}$ were utilised. Thereafter, F1 scores for the POS were calculated using the formula $\text{F1 score} = 2 \text{ x} \frac{\text{Precision x Recall}}{\text{Precision+ Recall}}$.



Text Annotation Tools

Process parameters

Select the file(s) to process: C:\Users\NWUUSER\Desktop\RAIL\HAWU BABE _RAIL DATA.txt

Language: Siswati (SS)

Tool: Part of Speech Tagger - simplified tagset

□ Output intermediary files

Processing information

Figure 1: Siswati Text Annotation POS tagger

Part of Speech	Tagsets
Noun	N
Verb	V
Adjective	ADJ
Adverb	ADV
Class-indicating demonstrative	CDEM
Conjunction	CONJ
Copulative	COP
Interjection	INT
Question word	INTER
Place and brand name	NPP
Possessive	POSS
Pronoun	PRO
Quantitative pronoun	PROQUANT
Relative	REL
Abbreviation (incl. acronyms)	ABBR

Table 1: Tagsets used in the STAT POS tagger

4 Results

4.1 Recall

Table 2 shows the results for the calculated Recall for each POS using Python 3.

POS	Recall
N	0.761
V	0.756
ADV	0.571
POSS	0.963
REL	0.706
CDEM	0.333
COP	0.75
CONJ	0.85
PRO	0.563
ADJ	0.75
INTER	0.0
INT	0.0
IDEO	0.0

Table 2: Recall for all POS in the data

In Table 2**,** the Recall for N (nouns) is 0.761, V (verbs) 0.571, ADV (adverbs) 0.571, POSS (possessives) 0.963, REL (relatives) 0.706, CDEM (Class-indicating demonstrative) 0.5, COP (copulatives) 0.75, CONJ (conjunctions) 0.85, PRO (pronouns) 0.563, ADJ (adjectives) 0.75, INTER (question words) 0.0, INT (interjections) 0.0 and IDEO (ideophone) 0.0.

4.2 Precision

Table 3 shows the results for the calculated Precision for each POS using Python 3.

POS	Precision
N	0.417
V	0.798
ADV	0.8
POSS	0.813
REL	0.523
CDEM	0.25
COP	0.75
CONJ	0.68
PRO	1.0
ADJ	0.75
INTER	0.0
INT	0.0
IDEO	0.0

Table 3: Precision for all POS in the data

In Table 3, the Precision for N is 0.417, V 0.798, ADV 0.80, POSS 0.813, REL 0.523, CDEM 0.25, COP 0.75, NPP 0.0, CONJ 0.68, PRO 1.0, ADJ 0.75, INTER 0.0, INT 0.0 and IDEO 0.0.

4.2 F1 Score

Table 4 shows the results for the calculated F1 scores for each POS.

Part of Speech	F1 score
N	0.54
V	0.78
ADV	0.67
POSS	0.88
REL	0.60
CDEM	0.29
COP	0.75
CONJ	0.76
PRO	0.72
ADJ	0.75
INTER	0.0
INT	0.0
IDEO	0.0

Table 4: F1 score for POS

In Table 4, the calculated F1score for N is 0.54, V 0.78, ADV 0.67, POSS 0.88, REL 0.60, CDEM 0.29, COP 0.75, CONJ 0.76, PRO 0.72, ADJ 0.75, INTER 0.0, INT 0.0 and IDEO 0.0.

5 Discussion

5.1 Nouns

The Recall of 0.761 for nouns means that out of all the true nouns in the data, the STAT POS tagger accurately tagged 76.1% of them, showing that it identified most nouns in the data.

The Precision of 0.417 reveals that out of all the words the tagger labelled as nouns, only 41.7% were true nouns, demonstrating that the tagger is also tagging a lot of words that are not nouns as nouns. The F1 score of 0.54 shows that the tagger achieved a moderate overall performance, even though there is a disparity between its Recall and Precision.

5.2 Verbs

The Recall of 0.756 for verbs reveals that out of all the actual verbs in the data, the tagger correctly identified 75.6% of them, with the remaining 38.5% being false negatives, which are the real verbs missed or misclassified by the tagger. The Precision of 0.798 means that out of all the words the tagger labelled as verbs, 79.8% were indeed verbs, with the remaining 20.2% being false positives, that is, words the tagger incorrectly tagged as verbs when they were other POS. The F1 score of 0.78 indicates that the tagger attained a strong and balanced performance, assigning verb tags with high accuracy.

5.3 Adverbs

The Recall of 0.571 for adverbs means that out of all the adverbs that appear in the dataset, the POS tagger correctly identified 57.1% of them as adverbs, and the remaining 42.9% are false negatives, which are the real adverbs that the tagger missed or classified as other POS. The Precision of 0.80 shows that of all the words the tagger labelled as adverbs, 80% were indeed adverbs, with the remaining 20% being false positives, that is, words the tagger incorrectly tagged as adverbs when they were something else. The F1 score of 0.67 shows a moderate overall performance of the tagger, with a fairly accurate prediction of adverbs, but missing many of them. This F1 score reveals an imbalance between the low recall and high precision.

5.4 Possessives

The Recall of 0.963 means that out of all the actual possessives in the dataset, 96.3% were correctly identified by the POS tagger, with the remaining 3.7% being false negatives or real possessives missed by the tagger. The Precision of 0.813 reveals that out of all the words the tagger labelled as possessive, 81.3% were real possessives, and the remaining 18.7% were words it labelled as possessive but were not. The F1 score of 0.88 indicates a strong and very good overall performance of the tagger, with a high, well-balanced accuracy.

5.5 Relatives

The Recall of 0.706 for relatives means that out of all the relatives in the data, the POS tagger correctly tagged 70.6% of them. The Precision of 0.523 means that out of all the words the tagger labelled as relatives, only 52.3% were correct, demonstrating that the tagger is producing a lot of false positives, that is, many words it marks as relatives are not relatives. The F1 score of 0.60 indicates that the POS tagger has moderate performance, successfully identifying many correct relatives with frequent tagging errors. This reveals an imbalance as the tagger has a strong recall at the expense of precision.

5.6 Class indicating demonstratives

The Recall of 0.333 for demonstratives means that out of all the actual demonstratives in the data, the POS tagger correctly identified 33.3% of them and missed 66.7%, meaning that many of the demonstratives were tagged as something else. The Precision of 0.25 shows that out of all the words the tagger labelled as demonstratives, only 25% were real demonstratives. The 75% were false positives; words incorrectly tagged as demonstratives. The F1 score of 0.29 indicates a poor performance of the tagger, showing failure in consistently and accurately tagging demonstratives.

5.7 Copulatives

The Recall of 0.75 for copulatives means that out of all the actual copulatives in the dataset, the POS tagger correctly identified 75% of them. This implies that it missed about 25% of the copulatives, and these were false negatives. The Precision of 0.75 shows that out of all the words the tagger predicted as copulatives, 75% were demonstratives, implying that 25% of its copulative predictions are false positives (words incorrectly tagged as copulatives). The F1 score of 0.75 indicates a balanced and strong overall performance of the tagger.

5.8 Conjunctions

The Recall of 0.85 for conjunctions means that out of all the actual conjunctions in the dataset, the tagger found 85% of them, missing about 15% of conjunctions (false negatives). The Precision of 0.68 means that out of all the words the tagger labelled as conjunctions, only 68% were real conjunctions, suggesting that 32% of the tagger's 'conjunctions' predictions are wrong (false positives). The F1 score of 0.76 indicates that the tagger has good performance overall, accurately identifying most conjunctions, despite a considerable number of tagging errors.

5.9 Pronouns

The Recall of 0.563 on pronouns means that out of all actual pronouns in the data, the tagger correctly identified only 56.3% of them. The Precision of 1.0 means that out of all the words the tagger predicted as pronouns, every single one was correct. This shows that there were no false positives, and it never tagged a non-pronoun as a pronoun. The F1 score of 0.72 indicates that the tagger has good performance but lacks comprehensiveness, resulting in perfect precision but average recall.

5.9.1 Adjectives

The Recall of 0.75 for adjectives means that out of all actual adjectives in the dataset, the POS tagger correctly identified 75% of them. The Precision of 0.75 shows that out of all the words the tagger predicted as adjectives, 75% were true adjectives, implying that 25% of its adjective predictions were false positives. The F1 score of 0.75 indicates a balanced and strong overall performance of the tagger.

5.9.2 Question words, interjections, and ideophones

The Recall of 0.0 for question words (INTER), interjections (INT) and ideophones (IDEO) means that the POS tagger failed to identify any of these POS in the dataset correctly. The Precision of 0.0 also means that none of the words the tagger labelled as INTER, INT and IDEO were in these word categories. The F1 score of 0.0 therefore, shows that the STAT POS tagger underperformed in the tagging INTER, INT and IDEO.

6 Conclusion

The results show that the Recall for nouns was 0.761, Precision 0.417, with an F1 score of 0.54; for verbs, the Recall was 0.756, Precision 0.798 and F1 score 0.54; for adverbs, the Recall was 0.571, Precision 0.8, and F1 score 0.67; for possessives, the Recall was 0.963, Precision 0.813 and F1 score 0.88. For relatives, the Recall was 0.706, Precision 0.523, and the F1 score 0.60; for class-indicating demonstratives, the Recall was 0.333, Precision 0.25, and the F1 score 0.29; and for copulatives, the Recall was 0.75, Precision 0.75, and the F1 score 0.75. For conjunctions, the Recall was 0.85, the Precision was 0.68, and the F1 score was 0.76; for pronouns, the Recall was 0.563, the Precision was 1.0, and the F1 score was 0.72; for adjectives, the Recall was 0.75, the Precision was 0.75, and the F1 score was 0.75. Meanwhile, the Recall, Precision, and F1 score for question words, interjections and ideophones were 0.0. These results show that the STAT POS performed better in tagging possessives, verbs, conjunctions, copulatives, adjectives, relatives, and nouns. However, it performed poorly on demonstratives, question words, interjections and ideophones, highlighting the need for refinement of the STAT POS tagger to enhance its tagging, especially when data from a non-governmental domain is used.

In future, the researcher will use data from Wikipedia and other sources that demonstrate multifaceted use of Siswati to evaluate the performance of the STAT POS tagger using the metrics accuracy, Recall, Precision and F1 score. For improving the performance of the STAT POS tagger, this study recommends that collaboration between Siswati POS tagger developers, language speakers, and linguists is essential for enhancing the efficiency of the tagger. Speakers of the language are required to provide extensive corpora that can be derived from Siswati literature and Wikipedia. The corpora would include all POS, including ideophones and interjections, utilised in various contexts within this language for training, refining and testing of the POS tagger.

7 Limitations of the study

This study utilised a limited corpus from a single genre to evaluate the performance of the STAT POS tagger.

References

Sonja Bosch, Laurette Pretorius and Axel Fleisch. 2008. Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies*, 17(2), 66–88.

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

Roald Eiselen and Martin J. Puttkammer. 2014. *Developing text resources for ten South African languages*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3698-3703, Reykjavik, Iceland. European Language Resources Association.

Aditi Sharma Grover, Gerhard B. Van Huyssteen, and Marthinus W. Pretorius. 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.

Wilbert Heeringa, Febe De Wet, and Gerhard B. Van Huyssteen.2015. Afrikaans and Dutch as closely related languages: A comparison to West Germanic languages and Dutch dialects. *Stellenbosch Papers in Linguistics Plus* 47: 1–18. https://doi.org/10.5842/47-0-649.

Ulrich Heid, Danie J. Prinsloo, Gertrud Faaß, and Elsabé Taljard. 2009. Designing a noun guesser for part of speech tagging in Northern Sotho. *South African Journal of African Languages*, 29(1): 1–19.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872.

Gabofetswe Malema, Boago Okgetheng, and Moffat Motlhanka. 2017. Setswana part of speech tagging. *International Journal of Natural Language Computing*, 6(6): 15–20.

Gubudla Aaron Malindzisa. 2005. *Hawu babe!* Shuter & Shooter, Pietermaritzburg.

Dimakatso S. Mathe and Roald Eiselen. 2021. Quantitative analysis of Sesotho sa Leboa part of speech tagger. South African. *Journal of African Languages, 41*(3): 259-269 https://doi.org/10.1080/02572117.2021.2010921.

Muzi Matfunjwa (2025). The efficacy of a Siswati part-of-speech tagger. In H. Ekkehard Wolff and Justus C. Roux, editors, Contextualising African Language Dynamics of Change, pages 251-268. African Sun Media, Stellenbosch, https://doi:10.52779/9781991260581.

Respect Mlambo and Muzi Matfunjwa. 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.

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

Siswati Text Annotation Tools Part of Speech Tagger. Available:
https://repo.sadilar.org/handle/20.500.12185/548. Accessed on 5 August 2025.