**Semi-automatic linguistic annotation for lexicography with machine learning**

Benjamin GUBB (gbbben001@myuct.ac.za)

Affiliation: School of Languages and Literatures: African Languages, University of Cape Town, Cape Town, South Africa

Cael MARQUARD (cael.marquard@gmail.com)

Affiliation: Department of Computer Science, University of Cape Town, Cape Town, South Africa

Many terminology lists for South African languages provide only the translations of each term without grammatical information such as part-of-speech and noun class. This makes it difficult to know how to use these words correctly in context. Annotating these terminology lists manually requires linguistic expertise and can be costly, time-intensive, and error prone. Instead, we propose annotating these terminology lists using a machine-learning classifier. An expert will then review the generated output to ensure accuracy. We will apply this approach to isiXhosa and integrate the results into the IsiXhosa.click online dictionary (Marquard 2024). This progresses the annotation of lexicographic works by making it easier to input linguistic information in dictionaries.

We build on IsiXhosa.click’s crowd-sourced lexicography and focus on streamlining linguistic annotation. Previously, terms were manually annotated, which is a labour-intensive, time-consuming process. With an automated machine-learning classifier, we hypothesise this process will be faster and less tedious. We propose applying machine-learning classifiers to annotate UCT’s mechanical engineering (Mechanical Engineering IsiXhosa Glossary 2023) and medical school glossaries.

A variety of data-driven Human Language Technologies (HLTs) have been developed for isiXhosa (Agbeyangi & Jere 2024), including automatic classifiers for linguistic information such as part-of-speech and noun class. However, their performance is not yet good enough to rely on them solely. Additionally, most are developed for words in a sentence-level context, and not standalone as appearing in a dictionary.

We adopt a process similar to the one described by Gaustad & Puttkammer (2022) in developing a linguistically annotated corpus. The pipeline begins with words being automatically annotated by machine-learning classifiers with their part-of-speech and noun class, as well as the classifiers’ confidence. It is then reviewed by a human annotator and checked for quality control.

Key tasks include developing the noun-class tagger, applying the taggers to the terminology lists, hiring an expert, native isiXhosa speaker to review the classifier’s output, and comparing the classifiers to each other and the manual corrections. The corrected terms will be added to the IsiXhosa.click database.

Two types of machine-learning classifiers are applied: one for part-of-speech annotation and one for noun-class annotation. Two options are considered for the part-of-speech classifier: the classifier developed in the NLAPOST21 Shared Task (Pannach et. al 2021), and the classifier developed by du Toit & Puttkammer (2021). Output from both will be analysed for accuracy at the end of the project. The classifier for noun-class annotation will be developed from scratch, we hypothesise that a purpose-built noun-class classifier will outperform general taggers like morphological analysers.

The noun-class classifier is a simple bi-LSTM model trained on linguistically annotated data (Gaustad & Puttkammer 2022). A bi-LSTM model combines two Long Short-Term Memory Models (LSTMs) (Hochreiter & Schmidhuber 1997), one reading the input forward and one backward. An LSTM is a type of Recurrent Neural Network, a class of models storing past events and are thus well-suited for sequence processing tasks. Bi-LSTMs have successfully been applied to similar tasks for Nguni languages, such as morphological (MorphParse 2024) and part-of-speech tagging (Pannach et. al 2021).

We experimentally evaluate the approach through two metrics: classifier accuracy, and the time a human takes to review outputs. These allow us to conclude whether the process is useful, and whether the hypothesis will be accepted.

By using automated machine learning for lexicographic annotation, we plan to evaluate the accuracy of machine-learning classifiers for linguistic annotation in isiXhosa. This approach streamlines the annotation process, saves time manually reviewing suggestions, maintains accuracy, and preserves the isiXhosa language, all while improving the efficiency of adding grammatical information to lexicographic works, making them more informative and accessible.

**References**

Agbeyangi, A., & Jere, N. 2024. IsiXhosa in the Digital Age: Navigating Language Preservation and Innovation: A Systematic Scoping Review. *IEEE Access* (12): 125835.

Gaustad, T., & Puttkammer M.J. 2022. Linguistically annotated dataset for four official South African languages with a conjunctive orthography: IsiNdebele, isiXhosa, isiZulu, and Siswati. *Data in Brief* (41): 107994*.*

Hochreiter, S., & Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* (9): 1735-1780.

Marquard, C. 2024. IsiXhosa.click: online, open, user-friendly, and searchable isiXhosa-English dictionary software. In: *African Association for Lexicography. 28th International Conference.* 1 July, 2024. Pretoria, South Africa: 35-36. Available: <https://www.afrilex.co.za/_files/ugd/bd0c6e_369dc939341044a6b4671d36d459717c.pdf>. Accessed on 24/03/2025.

Mechanical Engineering IsiXhosa Glossary. 2023. Available: https://ched.uct.ac.za/sites/default/files/media/documents/ched\_uct\_ac\_za/1667/Mechanical%20Engineering%20IsiXhosa%20Glossaries%20Project%20verified%20%28XHO%20NLB%20VERIFIED%2030Mar2023%29.pdf. Accessed on 30/03/2025.

MorphParse. 2024. Available: <https://projects.cs.uct.ac.za/honsproj/cgi-bin/view/2024/marquard_mawere.zip/index.html>. Accessed on 24/03/2025.

Pannach, F., Meyer, F., Jembere, E., & Dlamini S.Z. NLAPOST2021 1st Shared Task on Part-of-Speech Tagging for Nguni Languages. (2022). *Journal of the Digital Humanities Association of Southern Africa* (3): 1.

du Toit, J.S., & Puttkammer, M.J. 2021. Developing Core Technologies for Resource-Scarce Nguni Languages. *Information* (12): 520.