



Contribution ID: 283

Type: **Talk**

## Computational Model for Identifying Sepitori in Multilingual Texts

*Wednesday, 3 December 2025 15:50 (15 minutes)*

The Sepitori language (also known as Pitori or Pretoria Sotho) is a dynamic and evolving creole language predominantly spoken in urban townships of Pretoria, South Africa. It blends Setswana, Sesotho, Afrikaans, and English, with frequent instances of code-switching and slang. Despite its widespread usage, Sepitori remains underrepresented in natural language processing (NLP) tasks, particularly in language identification and text processing.

This paper proposes the development of a Sepitori Language Identification (ID) Model, designed to classify and distinguish Sepitori text from other South African languages. The model addresses the unique challenges of multi-language mixing, informal vocabulary, and varying dialects within the Sepitori speech community. By leveraging machine learning techniques and deep learning models, including convolutional neural networks (CNN) and transformer-based models (e.g., BERT), the model utilizes a large-scale corpus of annotated Sepitori, Setswana, Sesotho, Afrikaans, and English samples. The model incorporates multiple linguistic features, such as n-grams, word embeddings, and syntactic patterns, to accurately identify Sepitori text, even when it involves heavy code-switching or slang.

This work contributes to the linguistic field by providing a novel computational tool for processing Sepitori, enabling the automatic detection of Sepitori in a variety of contexts, including social media, web scraping, and corpus development. It also lays the foundation for improving language resources for underserved African languages, with potential applications in speech recognition, machine translation, and sentiment analysis. The model is expected to improve the accessibility and representation of Sepitori in digital and computational platforms, fostering greater inclusivity for African language speakers in the digital age.

### Presenting Author

Dan Masethe

### Email

[masethehd@tut.ac.za](mailto:masethehd@tut.ac.za)

### Student or Postdoc?

PhD or DTech

### Institute

### Registered for the conference?

**CHPC User**

## **CHPC Research Programme**

**Primary authors:** Dr MASETHE, Dan (Tshwane University of Technology); Ms MASETHE, Mosima (Sefako Makgatho University of Technology)

**Presenter:** Dr MASETHE, Dan (Tshwane University of Technology)

**Session Classification:** DIRISA

**Track Classification:** DIRISA