



Communication and Marketing Department
Isebe loThungelwano neNtengiso
Kommunikasie en Bemarkingsdepartement

Private Bag X3, Rondebosch 7701, South Africa
Welgelegen House, Chapel Road Extension, Rosebank, Cape Town
Tel: +27 (0) 21 650 5427/5428/5674 Fax: +27 (0) 21 650 5628

www.uct.ac.za

5 May 2026

UCT researchers build multilingual AI system for SA's official languages



The UCT researchers behind MzansiLM. From left: Simbarashe Mawere, Anri Lombard, Dr Jan Buys and Dr Francois Meyer.

Photo: supplied.

A team of researchers at the University of Cape Town (UCT) has developed a new artificial intelligence (AI) language model trained specifically on South Africa's 11 official written languages. This marks a significant step toward closing a gap that has left millions underserved by mainstream AI tools.

The research, which will be presented at the Language Resources and Evaluation Conference (LREC) in Mallorca, Spain in May 2026, introduces a landmark dual contribution to African language AI development. The first is MzansiText, a curated multilingual dataset covering the 11 official written languages, and MzansiLM, a language model trained on that dataset from scratch. The work was led by Anri Lombard and Dr Jan Buys from UCT's [Department of Computer Science](#), together with Dr Francois Meyer and a broader team of collaborators.

The paper arrives at a time when AI-powered language tools are rapidly shaping how people access information, work and communicate globally. But for speakers of most South African languages, that reality looks quite different. Ask a popular AI assistant a question in isiNdebele or Sepedi, and the response is likely to be poor, inconsistent or simply wrong. The reason, the researchers explain, comes down to data.

“In language modelling, languages are considered low resource, primarily because there are much fewer and smaller textual datasets available in these languages for training language models,” said Dr Buys, a senior lecturer in the Department of Computer Science. “Our dataset, MzansiText, is still small compared to data available for high-resource languages such as English and major European and Asian languages, but larger than previous datasets for South African languages.”

Nine of South Africa’s 11 official written languages fall into this low-resource category. Languages like isiZulu and isiXhosa have received some attention from the global research community, but others, including isiNdebele and Sepedi, have been largely overlooked. MzansiLM is believed to be the first publicly available decoder-only language model designed to support all 11 official written languages in a single system.

“There has been real progress in language modelling for African languages, including some South African ones like isiXhosa and isiZulu,” said Dr Meyer, a lecturer in the Department of Computer Science. “But most existing models only cover a subset of languages. With MzansiLM, we wanted to build a single model focused specifically on South Africa that covers all 11 official written languages, including those that are often left out.”

From master’s research to a baseline for the field

For Lombard, a master’s student in computer science, the project began with a recurring question in his research.

“I came into this work through my master’s research, which looks at how different language-model architectures perform for low-resource languages, since that is still a relatively underexplored area,” he explained. “One thing that stood out to me is that publicly available models tended to cover only a subset of the South African languages we care about. MzansiLM was meant to provide a small decoder-only baseline that future work can compare against and build on.”

The model itself, with 125 million parameters, is modest by today’s commercial AI system standards. Despite its size, the model demonstrated strong performance in targeted tasks. It outperformed much larger open-source models on benchmarks in several South African languages. On isiXhosa text generation, for instance, it produced results that competed with encoder-decoder models more than 10 times its size.

Not a chatbot, but a foundation

It is worth being clear about what MzansiLM is and what it is not. Unlike tools such as ChatGPT or Claude, it is not designed for open-ended conversation. It is a base model, a foundation that developers and researchers can adapt for specific purposes through a process known as fine-tuning.

“In practice, that means developers could build tools for specific use cases, for example, summarising information or annotating raw data, in South African languages,” Meyer said. “Adapting MzansiLM for a limited use case might be more effective and affordable than

relying on proprietary large language models, if you want users to be able to interact with a system in their home language.”

The more immediate benefits for everyday users will come from future, larger versions of the model and from systems built on top of this foundation. The research also offers important insight into a broader global challenge, why even advanced AI systems still struggle beyond dominant languages like English.

“Our findings show that the model can work well when fine-tuned for specific tasks but is not yet able to work well for general-purpose user interaction or instruction following, due to the limited training data,” Buys explained. “This helps to explain why even larger language models don’t yet work as well when used in languages other than English.”

An open research community is essential

The team is clear that MzansiLM is a step, not a destination. Closing the gap between South African languages and the capabilities now available in English will require sustained, collective effort.

“A lot of the progress we were able to make depends on earlier open research from the African Natural Language Processing research community, so continuing that openness is essential,” Lombard said. “We still need better and broader data sources, stronger benchmarks and the kind of shared datasets, models, code and results that make it possible for others to reproduce and extend the work.”

Meyer echoed that view. “The research community plays an important role here by working openly, sharing datasets, models and findings so others can build on them. That kind of openness is often what leads to progress, especially compared to proprietary systems where much of the data and methodology isn’t accessible.”

In line with this approach, the UCT team has made both MzansiText and MzansiLM publicly available to support further research and innovation. The paper, “MzansiText and MzansiLM: An Open Corpus and Decoder-Only Language Model for South African Languages”, is available on arXiv.

- [Read the full paper.](#)
- [Access the MzansiText dataset.](#)
- [Download the MzansiLM model.](#)

Story by Mari van der Merwe, UCT News.

Issued by: UCT Communication and Marketing Department

Velisile Bukula

Head: Media Liaison
Communication and Marketing Department

University of Cape Town
Rondebosch
Tel: 021 650 2149
Cell: 071 642 3495
Email: velisile.bukula@uct.ac.za
Website: www.uct.ac.za