Contribution ID: **14** Type: **Talk**

Developing a Word Sense Disambiguation Dataset for Sesotho sa Leboa: Manual Annotation and Authentic Text Sources”

Word Sense Disambiguation (WSD) plays a critical role in Natural Language Processing (NLP), particularly for low-resourced languages like Sesotho sa Leboa (Northern Sotho), which lack comprehensive linguistic resources. This study presents the development of a manually annotated WSD dataset tailored for Sesotho sa Leboa, addressing the scarcity of such corpora. The dataset was constructed using text collected from a variety of authentic and standardized sources, including academic dissertations, research papers, dictionaries, and curated web pages featuring ambiguous lexical items. These sources were deliberately chosen for their formal use of language, ensuring reliability in the representation of word meanings. Data collection was conducted through web scraping, utilizing the Beautiful Soup library in Python to parse HTML documents and systematically extract relevant content while removing extraneous HTML tags. Manual annotation was then performed by native language experts to label word senses based on contextual usage. This dataset provides a foundational resource for training and evaluating WSD models in Sesotho sa Leboa, promoting the development of more accurate and context-aware NLP tools for underrepresented languages.

Primary author: Ms MASETHE, Mosima Anna

Co-author: Mr MASETHE, Hlaudi Daniel (Tshwane University of Technology)

Presenter: Mr MASETHE, Hlaudi Daniel (Tshwane University of Technology)