**Using computational tools and a corpus lexicography framework in developing an isiZulu LSP Dictionary.**

**Mthuli Buthelezi (buthelezi@bfsu.edu.cn)**

**University of KwaZulu-Natal: Department of Linguistics; Durban, South Africa**

**Cael Marquard (cael.marquard@gmail.com)**

**University of Cape Town: Department of Computer Science; Cape Town, South Africa**

We propose compiling an isiZulu Language for Special Purposes (LSP) dictionary to be deployed and made available through an online software interface. This study develops alongside the much needed studies on open-source code for creating dictionaries[[1]](#footnote-0) and on the work of writing online dictionaries for languages like isiZulu[[2]](#footnote-1).

The herein proposed isiZulu LSP dictionary offers distinct advantages over existing online isiZulu dictionaries. Firstly, it is tailored to specialized linguistic and literature terms. Secondly, its open-source code enables *localisation* to other African languages, fostering inclusivity and accessibility. Thirdly, its online interface is available in isiZulu. This is a novel approach in interactive interface design that has not been seen in other online isiZulu dictionaries. Finally, its development addresses the critical need for wide-ranging dictionaries in African languages. This endeavour aligns with the imperative to advance Human Language Technologies (HLT) in African languages and contributes to pedagogy and terminology development, ultimately transforming isiZulu pedagogy.

Importantly, the algorithm of the online isiZulu LSP dictionary is a *template-based* algorithm (Keet & Khumalo, 2014), and it is *localised* from the isiXhosa.click algorithm.

The dictionary utilizes an LSP corpus, which can be scaled, reused, and redistributed, thus enhancing its linguistic richness. Therefore, this dictionary has the capacity to be improved in the future.

This study utilises the corpus lexicography framework to establish an isiZulu dictionary for the linguistics and literature terms. Corpus lexicography as a theoretical framework and a research methodology provides for the use of term extraction and computational techniques (Khumalo, 2015). The textual datum in the isiZulu LSP dictionary is intrinsically a corpus datum. Therefore, in this study, the corpus lexicography framework guides the process of term extraction in the process whereby the potential lexical entries are identified as headwords for the possible addition to the dictionary.

The selected isiZulu LSP corpus from Kumalo (1995) provides the terms which are incorporated into the dictionary. The extracted isiZulu terms have their English translation equivalents. The proposed dictionary is intended for the teachers and learners of linguistics and literature. The established algorithm for the proposed isiZulu LSP dictionary also serves as an electronic database for term compilation, analysis and storage.

The isiXhosa.click[[3]](#footnote-2) online dictionary software will be adapted to provide a usable front-end to the isiZulu LSP dictionary. Initially developed as an online dictionary for isiXhosa, it allows for live word search and community participation in the dictionary. Community participation is facilitated through tools on the site to suggest additions, modifications, and removals of words in the database. These suggestions are visible in a moderation dashboard, where they are manually edited and approved. This review-by-experts process is intended to prevent vandalism and the capturing of incorrect information.

IsiXhosa.click provides a search function which is the primary way users discover words. Words may be searched either by their isiZulu form, their isiZulu stem, or their English translation. The search is typo-tolerant and allows for even misspelled queries to return results. Therefore, the software is easy to use even for inexperienced users. Additionally, the site is suitable for mobile use, increasing its accessibility.

IsiXhosa.click also supports capturing of grammatical information, addition of example sentences for words, and linking related words together. This software has been selected for the following reasons. Firstly, it currently targets isiXhosa, which is similar to isiZulu, meaning that the software already fulfills many of the dictionary’s requirements. Secondly, it is open-source, meaning that the knowledge acquired from the isiZulu LSP dictionary development project may be directly applicable to other efforts aiming to adapt the software to other languages. Thirdly, the online interface is suitable for mobile use, expanding its potential reach. Lastly, its community editing tools may extend the lifespan of the isiZulu LSP dictionary by allowing interested third parties to continue its maintenance and improvement. Therefore, there are plans to expand these dictionaries and make them into ‘living dictionaries’ with the continual appraisal and addition of lexical items and their example sentences.

Key tasks involved in the adaptation of the isiXhosa.click software to the isiZulu LSP include importing the isiZulu LSP data into the database, adapting the interface to support localisation for isiZulu, creating translations for interface text and labels, and deployment of the modified software such that it may be accessed publicly. In order to better support searching by isiZulu stems, a new isiZulu stemmer should be bootstrapped from the existing isiXhosa algorithm. Because this is only used as a search heuristic, accuracy is not important, hence a simple greedy stemming approach will suffice. Importantly, the software will be extended to support *localisation* of its interface to other languages. Thus, the language development efforts intended in the proposed study underscore the importance of integrating the lexicographic and the corpus linguistic insights with computational methodologies to address the challenges and opportunities presented by African languages in the digital age.

**References**

Keet, M. & Khumalo, L. 2014. Toward verbalizing ontologies in isiZulu. *Controlled Natural Language* 8625: 78-89.

Khumalo, L. 2015. Semi-automatic term extraction for an isiZulu linguistic terms dictionary. *Lexikos* 25: 495-506.

Kumalo, M. 1995. *English/IsiZulu linguistic and literary terms: Amatemu esiNgisi/esiZulu elingwistiksi nelithrisha.* Cape Town: Vivlia Publishers.

1. [https://github.com/topics/dictionary-software[open-source dictionaries code] [↑](#footnote-ref-0)
2. [https://zuludict.acdh.oeaw.ac.at/zuludict/[example of initiatives on creating online isiZulu dictionaries] [↑](#footnote-ref-1)
3. [https://ched.uct.ac.za/dot4d/implementation-student-led-initiatives/isixhosaclick] [↑](#footnote-ref-2)