Creating Bilingual Corpus for isiZulu: A Case Study from the University of KwaZulu-Natal

Thandeka Mbali Gumede[1], Rooweither Mabuya[2], Njabulo Hadebe[3]
[1]University Language Planning and Development Office, UKZN
[2]South African Centre for Digital Language Resources, NWU
[3]University Language Planning and Development Office, UKZN
Gumedet1@ukzn.ac.za , roo.mabuya@nwu.ac.za , hadeben3@ukzn.ac.za

Abstract

Although several bilingual resources exist, there is a lack of domain-specific, institutionally verified parallel corpus focusing on academic and administrative texts. Existing datasets such as Autshumato English–isiZulu corpus, UNISA English/Zulu Parallel Corpus, and the WebCrawl African Corpus hosted on GitHub provide valuable material but differ in accessibility, domain coverage, and documentation. To complement these initiatives, the University Language Planning and Development Office (ULPDO) at the University of KwaZulu-Natal has developed a curated isiZulu–English Parallel Corpus comprising 10,000 carefully aligned sentence pairs drawn from institutional and academic texts. This paper outlines the corpus compilation process, including data sourcing, cleaning, alignment, and validation, and discusses key structural and linguistic challenges encountered. The resource contributes to translation studies, terminology development, and multilingual natural language processing, while supporting ongoing efforts to advance the digital presence and intellectualisation of isiZulu.

1. Introduction

Language resources play an important role in the development of digital technologies such as translation tools, chatbots, and language learning platforms. However, African languages like isiZulu still do not have enough digital resources to support these technologies. This makes it difficult to build accurate and useful tools that include isiZulu, even though it is one of the most widely spoken languages in South Africa. Most available isiZulu data is either small, unbalanced, or not well aligned for research and natural language processing (NLP) work. Some efforts have been made to create bilingual corpora, such as the Autshumato project, the UNISA English–isiZulu corpus, and the Masakhane translation initiatives. These are valuable, but there is still a need for clean, verified, and domain-specific isiZulu–English data that can be used for both research and real-world applications. To help address this problem, the ULPDO at the University of KwaZulu-Natal has developed a new isiZulu–English parallel corpus. The corpus includes 10,000 aligned sentence pairs that come from official university documents, academic texts, and policy materials. All content was carefully translated and verified by language experts to ensure accuracy and consistency. This paper describes how the corpus was created, how the data was collected and aligned, and how quality control was done. It also explains how this resource can support language development, translation improvement, and NLP projects involving isiZulu. The goal is to contribute to the digital growth of isiZulu and make it more visible in modern technology and research.

2. Related work

From an global perspective, efforts such as the IIT Bombay English–Hindi Parallel Corpus (Kunchukuttan, Mehta et al. 2017) and the multilingual Indian corpora collection (Siripragada 2020) demonstrate scalable models for compiling large, multi-domain corpora in low-resource languages. Similarly, the Italian–Chinese corpus by (Tse, Mirri et al. 2020) showcased innovative web-scraping methods for automatic sentence alignment and techniques that could be adapted to African corpus development, where online content is increasing.

Within the African context, African scholars have made significant strides in compiling parallel corpora for low-resource indigenous languages. For instance, a conducted study described the creation of an Emakhuwa–Portuguese corpus using legal texts, religious content, and children's stories to fill the resource gap for Mozambican languages (Ali, Caines et al. 2021). Similarly, there was a manually created English–Igala corpus of 50,000 aligned sentences, due to the absence of digital Igala content online (Ayegba, Onoja et al. 2017). The above reflect the collaborative, hybrid approaches, often combining manual translation with semi-automated tools that are necessary in African contexts due to infrastructural and linguistic challenges (Ayegba, Onoja et al. 2017).

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit
The copyright remains with the authors.

Furthermore, the Twi-English corpus by (Afram, Weyori et al. 2022) and the Nigerian Pidgin-English discourse-annotated corpus (Scholman, Marchal et al. 2025) further emphasise the importance of local, context-specific corpus development. These corpora not only support machine translation but also provide insight into linguistic structures, discourse features, and the lexical characteristics of African languages, opening new avenues for applied research and natural language processing (NLP).

In recent years, there has been a growing emphasis on the need to develop African languages into languages of science, technology, and education. One of the critical enablers of this transformation is the availability of multilingual parallel corpora, structured, aligned datasets containing sentence pairs in two or more languages. These corpora are foundational to building terminological dictionaries, training machine translation systems, and supporting research in digital humanities, translation studies, and computational linguistics (Ndhlovu 2016, Shoba 2018, Khumalo 2020). However, despite this recognition, the development of such resources for African languages has lagged, with isiZulu and other indigenous South African languages remaining significantly under-resourced (Kotzé 2016). Scholars and institutions alike have acknowledged that building representative, high-quality corpora is essential for the intellectualisation of African languages and the development of tools that meet academic and societal multilingual needs (Ndhlovu 2016, Shoba 2018, Khumalo 2020).
Moreover, several South African scholars have recognised the importance of parallel corpora and explored innovative methodologies to extract bilingual terminology from such resources. (Ndhlovu 2016), for instance, used ParaConc, a bilingual concordance, to extract English–Ndebele terminology for the creation of specialised dictionaries. This approach provided accurate results efficiently and demonstrated the value of corpus-based lexicography in promoting indigenous language development (Ndhlovu 2016). Shoba (2018) echoed this sentiment in her study on English–isiXhosa terminology, highlighting the role of parallel corpora in addressing the terminology gap in science, law, and commerce. Shoba (2018) underscored the effectiveness of corpus interrogation tools such as frequency lists and concordances (KWIC) for extracting headwords, synonyms, and usage examples in context elements critical for producing functional and user-centred bilingual dictionaries (Shoba 2018).

Khumalo (2020) shifted the focus towards pedagogical innovation, exploring how digital corpora such as the IsiZulu National Corpus (INC), the English–isiZulu Parallel Corpus (EiPC), and the IsiZulu Oral Corpus (IOC) were used to support online isiZulu teaching during the COVID-19 lockdown. Khumalo's work demonstrated that corpus-based tools like AntConc could be repurposed not only for research but also for multilingual digital teaching, thereby extending the utility of parallel corpora into the realm of e-learning and digital scholarship. From a computational linguistics perspective, Kotzé (2016) examined the preprocessing techniques needed to improve the quality and alignment of the English–Zulu parallel corpus for statistical machine translation. Kotze's findings reaffirmed that corpus quality, including sentence splitting, alignment, and manual verification, has a direct impact on translation accuracy, highlighting the ongoing need for rigorous corpus development methodologies (Kotzé 2016).

Additionally, corpus creation has progressed through national and institutional initiatives that align with the country's Human Language Technologies (HLT) strategy (Moors, Wilken et al. 2018). A landmark effort in this regard is the Autshumato Project, initiated in 2007 by the South African Department of Sports, Arts and Culture (DSAC), which represents a significant national effort to advance language technology and multilingual communication. The primary aim of Autshumato is to develop, release, and support open-source translation technologies that facilitate the translation process and enhance access to information for all South Africans (https://mt.nwu.ac.za/#). As a result, the South African Centre for Digital Language Resources (SADiLaR) is a national corpus portal and infrastructure that hosts parallel and monolingual corpora for South African languages (Autshumato datasets, NCHLT derivatives, and new parallel corpora for 11 South African official languages (https://corpus.sadilar.org/corpusportal/explore/corpus).

This paper, therefore, responds to the urgent need to fill that gap by documenting the creation, structure, and potential applications of the ULPDO English–isiZulu Parallel Corpus. With 10,000 aligned sentence pairs, the corpus is positioned to contribute meaningfully to terminology development, machine translation research, multilingual education, and the broader intellectualisation of isiZulu. By situating this contribution within the growing body of African and international corpus-based research, the study illustrates how locally developed linguistic resources can serve both scholarly and societal needs, ultimately

advancing multilingualism, digital inclusivity, and language equity.

3. Methods of building English-isiZulu Parallel Corpus

The compilation of the ULPDO English–isiZulu Parallel Corpus followed a structured and carefully considered methodology to ensure the quality, accuracy, and usability of the resource. Given the linguistic complexity of isiZulu and its structural differences from English, a hybrid approach was adopted combining automated tools with human linguistic expertise. This methodology is informed by established practices in corpus linguistics and bilingual corpus development (Khumalo 2020), with adaptations made to address local language realities and institutional contexts. The process involved several key stages: sourcing and selecting appropriate bilingual materials, preparing and cleaning the data, aligning English and isiZulu sentences, applying basic metadata tagging, and validating the final corpus. Each step was guided by the need to preserve meaning, maintain translation fidelity, and support potential reuse of the corpus in machine translation, terminology development, and multilingual research. The subsections below outline these steps in detail.

3.1 Data Collection and Preprocessing

The ULPDO English–isiZulu Parallel Corpus was developed using bilingual texts officially translated by ULPDO at UKZN. ULPDO serves as the central institutional unit responsible for translation and language policy implementation at the university. It plays a key role in promoting the use of isiZulu as an academic language, in line with the university's language policy and broader national language planning goals.

The primary sources for the corpus included formally translated university documents such as institutional policies, academic regulations, student handbooks, faculty circulars, and public information. These texts were selected because they are consistently produced in both English and isiZulu by professional language practitioners within the office. Each document undergoes a rigorous internal review process before publication, ensuring linguistic fidelity and terminological consistency. As such, they represent high-quality bilingual data suitable for inclusion in a parallel corpus (Egbert, Biber et al. 2022).

In addition to institutional documents, the corpus includes a substantial collection of PhD abstracts translated under UKZN's Doctoral Rule 09 (DR9), which mandates that all doctoral candidates submit their final abstracts in both English and isiZulu prior to graduation (Zungu 2021). This rule is unique to the institution and provides a rich body of formally edited discipline-specific bilingual academic texts. These abstracts were sourced across all four academic colleges of the university, covering a range of disciplines including Humanities, Health Sciences, Law and Management Studies, and Agriculture, Engineering and Science. Since the translations are conducted or vetted by ULPDO, these texts offer terminologically rich and structurally aligned bilingual content, making them particularly suitable for sentence-level parallel alignment and linguistic analysis.

For the construction of the parallel corpus, only documents that were fully available in both English and isiZulu were considered. Translations were required to have been produced or thoroughly vetted by qualified language practitioners within ULPDO, ensuring linguistic accuracy and terminological consistency. Crucially, the source and target texts are needed to exhibit structural and semantic equivalence to facilitate reliable sentence-level alignment. The corpus prioritized documents reflecting the university's core domains of academic scholarship, administrative governance, and public communication. Materials that were incomplete, loosely translated, or exhibited significant paraphrasing were systematically excluded to preserve the corpus's integrity and support its intended applications in linguistic research and language technology development.

All documents were sourced from ULPDO's internal digital repository. Files in non-editable formats (e.g., scanned PDFs) were converted to machine-readable text through a combination of manual transcription and the use of mobile OCR technology (iOS Live Text), selected for its robustness with the variable quality of the source documents. All extracted text underwent manual post-editing to ensure a final, high-quality version for corpus alignment.

Because all materials used in the corpus are institutionally owned and publicly communicable

documents, no personal data was processed, and formal ethics clearance was not required. However, the entire process adhered to the data governance principles outlined in the Protection of Personal Information Act (POPIA), Act No. 4 of 2013.

3.2 Data Preparation and Cleaning

After collecting the bilingual documents, the data was prepared and cleaned to ensure its suitability for sentence-level alignment. All English texts and their isiZulu translations were stored in a shared Microsoft OneDrive repository for centralized access and version control. Each document pair was carefully named and organized by document type, academic domain, and year. All bilingual files were exported into UTF-8 encoded plain-text format to preserve character integrity across languages. Non-linguistic elements such as tables of content, page numbers, and signatures were removed.

The documents that were exceptionally long or contained highly complex structures such as embedded tables, or non-textual elements that hindered automated processing of structured Excel or CSV files were manually created. In this process, English sentences were carefully copied into the first column, and their corresponding isiZulu translations were pasted into the second column, sentence by sentence. This manual approach ensured accurate pairing for challenging texts, forming a foundational dataset where each translation pair was clearly matched and preserved for semantic fidelity.

For the majority of the cleaned bilingual texts, a custom Python script was developed to automate the sentence alignment process between English source texts and their isiZulu translations. Existing alignment tools were not adopted because most are designed for widely studied language pairs and do not account for the linguistic characteristics of isiZulu, such as its complex morphology, word order, and sentence segmentation conventions. The custom script leveraged heuristics based on sentence length and lexical similarity, while also integrating manually prepared CSVs for complex documents. The output of this automated alignment was structured in a tabular format, typically CSV, ready for further linguistic analysis. This approach enabled scalable, high-quality alignment while maintaining full control over the methodology and ensuring reproducibility.

3.3 Quality Assurance and Validation

To ensure the reliability and accuracy of the ULPDO English–isiZulu Parallel Corpus, a rigorous quality assurance and validation process was implemented following the alignment phase. Both manual and automated checks were conducted to identify and rectify alignment errors, inconsistencies, and translation discrepancies. Initially, a subset of aligned sentence pairs was randomly sampled and reviewed by experienced language practitioners and translators from ULPDO to verify semantic equivalence, grammatical correctness, and terminological consistency. This human validation helped identify common issues such as misaligned sentences, omitted content, or translation deviations.

In parallel, automated scripts were employed to detect anomalies such as empty sentence pairs, excessive length mismatches, and duplicated entries. Statistical measures, including length ratio thresholds and lexical similarity scores, were applied to flag potential misalignments for further manual inspection. Following the sentence alignment described in Section 3.2, length ratio thresholds similar to those used during alignment were applied post-alignment to flag sentence pairs that were unusually long or short relative to each other. Flagged pairs were manually reviewed to ensure correct alignment. In addition, established QA standards were followed, including terminological consistency, structural equivalence (avoiding improper splitting or merging of sentences), and adherence to isiZulu orthographic and formatting conventions. This workflow ensured that the final corpus maintained both high-quality alignment and linguistic integrity, directly building the data preparation steps outlined in Section 3.2. Special attention was given to domain-specific terminology and culturally sensitive expressions to ensure fidelity to source meanings.

Despite these efforts, it is important to note that there are currently limited to no widely available tools specifically designed to fully automate the quality assurance process for isiZulu-English parallel corpus, particularly given the linguistic complexity and resource constraints of African languages (Keet and Khumalo 2017). As a result, much of the validation relies on expert human input and semi-automated methods. Recognizing this gap, the ULPDO is actively developing bespoke tools and workflows aimed at improving automation and

enhancing the quality control process for future corpus development.

Corrections arising from these reviews were incorporated iteratively, with the dataset undergoing multiple rounds of validation until alignment quality met the established standards. The finalized corpus was then formatted consistently and prepared for storage and dissemination. This thorough validation process guarantees that the corpus not only serves as a reliable linguistic resource but also supports downstream applications in machine translation, terminology extraction, and multilingual research.

3.4 Data Preparation and Cleaning

Following successful quality assurance and validation, the ULPDO English–isiZulu Parallel Corpus was formatted to ensure compatibility with various computational tools and ease of access for future research and development. The cleaned and validated sentence pairs were stored in UTF-8 encoded CSV files, a widely supported and versatile format that preserves text integrity across different platforms and software. Each CSV file contains aligned sentences arranged in two columns English in the first and isiZulu in the second accompanied by metadata fields such as document type, academic discipline, and document ID to facilitate filtering and contextual analysis. To ensure secure and centralized data management, the corpus files are maintained on Microsoft OneDrive within the ULPDO's shared digital repository. This setup enables controlled access for authorized personnel, version tracking, and seamless collaboration among corpus developers and language practitioners. Regular backups are scheduled to prevent data loss, and data governance policies compliant with the Protection of Personal Information Act POPIA, 2013 are strictly followed to protect sensitive information (de Waal 2022). The choice of CSV as the primary format supports easy integration with a range of natural language processing (NLP) tools and machine learning frameworks, allowing the corpus to serve diverse applications including machine translation, terminology extraction, and linguistic research (Thanaki 2017).

4. Discussions

Aligning bilingual corpus involving structurally divergent languages such as English and isiZulu presents a complex set of challenges, despite advancements in computational tools and alignment algorithms. These challenges range from initial preprocessing such as tokenization and normalization to sentence-level alignment, which often necessitates a combination of manual verification and rule-based heuristics.

Before delving into deeper linguistic complexities, it is essential to examine the overall corpus statistics for the data contained in our corpus. The bilingual dataset comprises 10,000 aligned sentence pairs, with a total of 165,519 English tokens and 116,710 isiZulu tokens. The average sentence length in English is 16.55 tokens, whereas isiZulu sentences average 11.67 tokens in length. This difference reflects a general tendency for isiZulu translations to be more concise in token count, despite being semantically rich (Schryver and Prinsloo 2000).

Figure 1 compares the average sentence lengths, illustrating that English sentences are typically longer than their isiZulu counterparts. However, this does not imply that isiZulu sentences are less complex. On the contrary, isiZulu is a morphologically rich language, where a single word can encapsulate grammatical features such as tense, agreement, and aspect, which would typically require multiple words in English. This results in alignment difficulty, as one isiZulu word may correspond to several English words, making direct sentence or word-level alignment more complex.

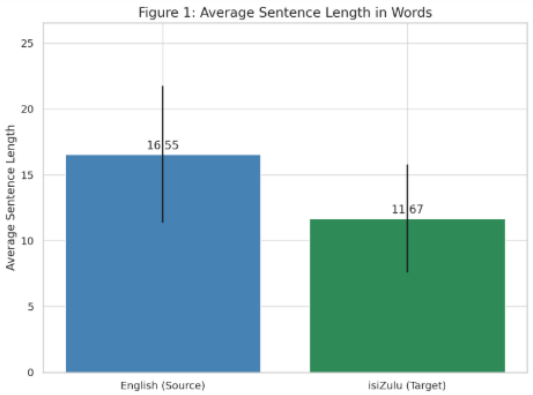

Figure 1: Average sentence length in words

Further linguistic complexity is evident in Figure 2, which presents a comparison of the Type-Token Ratio (TTR) between the two languages. The isiZulu corpus exhibits a significantly higher TTR,

indicating greater lexical diversity and extensive use of inflectional and derivational morphology. This elevated variation in word forms adds another layer of difficulty for automatic alignment systems, which often rely on surface-level token similarity.

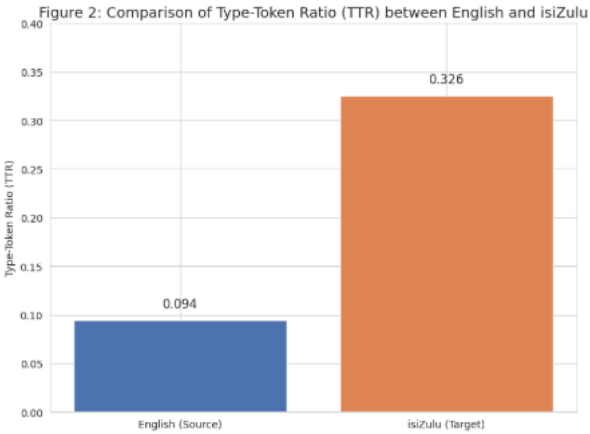

Figure 2: Comparison of TRR between English and isiZulu text

Together, these structural and processing challenges underscore the need for alignment strategies that are linguistically informed, especially when working with morphologically rich languages like isiZulu.

Structural and Technical Alignment Challenges

A significant technical challenge was sentence segmentation. English segmentation was largely automated using Microsoft Word macros, but isiZulu required careful manual segmentation due to its looser punctuation conventions and compound sentence structures. IsiZulu often encodes multiple ideas in a single complex sentence, reducing the likelihood of one-to-one mappings. The lack of localized alignment tools also limited performance. While tools such as Hunalign by Varga, Halácsy et al. (2008) perform well for European language pairs, they underperform on low-resource languages like isiZulu. Consequently, ULPDO used custom Python scripts based on heuristics such as sentence-length ratios and lexical anchor matching. Even so, manual verification was required to address cases involving non-literal translations and paraphrasing. Data quality issues such as typographical errors, skipped lines, or repeated segments further introduced noise into the corpus. As a result, extensive pre-cleaning and filtering were applied before alignment. Despite this, challenges remained, particularly in aligning texts with complex or irregular formatting. **Figure 3** presents a scatter plot comparing sentence lengths (in tokens) across 500 randomly sampled aligned English–isiZulu sentence pairs. Each point represents a single aligned pair, with the x-axis showing the English sentence length and the y-axis the corresponding isiZulu sentence length. The plot shows a general positive correlation, indicating that longer English sentences tend to align with longer isiZulu sentences. However, there are notable outliers where English sentences are much longer or shorter than their isiZulu equivalents. These outliers often reflect complex sentence structures, paraphrasing during translation, or segmentation mismatches, highlighting the limitations of automatic alignment methods that rely mainly on surface-level features like sentence length.

While the scatter plot visualizes these trends and potential misalignments, it was used primarily for illustration rather than as a formal quality control tool. Actual quality assurance relied on numerical length ratio thresholds and lexical similarity scores in automated scripts, with flagged pairs manually reviewed by language experts. This combination ensures both systematic and exceptional alignment errors are detected and corrected, maintaining semantic fidelity and linguistic accuracy, particularly for morphologically rich languages like isiZulu.

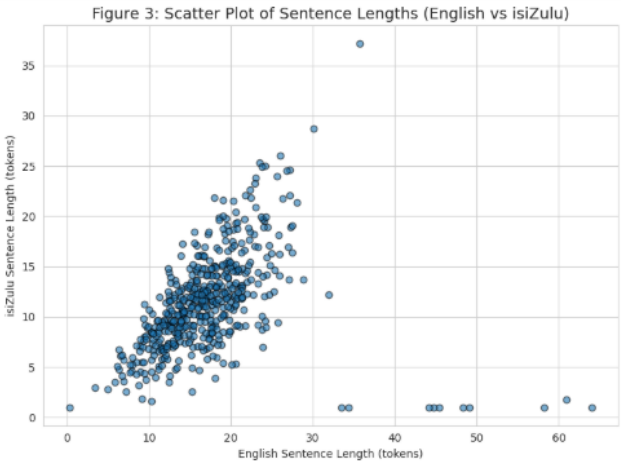

Figure 3: Scatter Plot of Sentence Lengths (isiZulu vs English)

5. Limitations

While the project achieved its goal of producing a well aligned English–isiZulu parallel corpus, a few limitations remain. One key challenge was the limited

availability of tools specifically built for isiZulu, which meant that many steps such as segmentation and alignment had to be done manually or semi automatically. This limited the speed and scale of processing. In addition, the corpus does not yet cover all domains equally, with more academic and formal texts than conversational or technical content. Lastly, some alignment inconsistencies may still exist, particularly in long or complex sentence structures, despite thorough quality checks.

6. Conclusion

This paper explored the structural and processing challenges involved in aligning English–isiZulu corpus, highlighting the implications of linguistic divergence on computational alignment methods. The analysis of 10,000 aligned sentence pairs revealed key asymmetries in average sentence length, type-token ratio, and token distribution, all of which impact the quality and reliability of automated alignment tools. The scatter plot of sentence lengths (Figure 3) particularly underscored the inconsistencies in sentence pairings, with notable outliers reflecting the complex, often non-parallel nature of bilingual translation between English and isiZulu.

Structural differences such as English's analytic syntax versus isiZulu's agglutinative morphology and noun class system introduce alignment complexities that go beyond surface token matching. These linguistic features cause isiZulu sentences to encode dense grammatical information within single words, while English distributes similar meaning across multiple words or phrases. As a result, word sentence alignment becomes difficult without incorporating deeper linguistic knowledge into alignment algorithms.

The findings emphasize the need for more linguistically informed approaches to corpus alignment in African languages. Future work should investigate alignment tools that integrate morphological analysis, syntactic parsing, and machine learning methods tailored to low-resource and morphologically rich languages. By addressing these challenges, researchers and developers can significantly improve the quality of bilingual resources and support the development of robust language technologies for underrepresented languages like isiZulu.

5. References

Afram, G. K., et al. (2022). "TWIENG: A Multi-Domain Twi-English Parallel Corpus for Machine Translation of Twi, a Low-Resource African Language."

Ali, F. D., et al. (2021). "Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique." arXiv preprint arXiv:2104.05753.

Ayegba, S. F., et al. (2017). "English–igala parallel corpora for natural language processing applications." International Journal of Computer Applications **975**: 8887.

de Waal, P. J. (2022). "The protection of personal information act (POPIA) and the promotion of access to information act (PAIA): It is time to take note." Current Allergy & Clinical Immunology **35**(4): 232–236.

Egbert, J., et al. (2022). Designing and evaluating language corpora: A practical framework for corpus representativeness, Cambridge University Press.

Keet, C. M. and L. Khumalo (2017). "Toward a knowledge-to-text controlled natural language of isiZulu." Language Resources and Evaluation **51**(1): 131–157.

Khumalo, L. (2020). "Using corpora in online isiZulu language teaching." N. Mkhize, N. Hlongwa-Ndimande, L. Ramrathan, & J. Smit (Eds.) **3**: 37–53.

Kotzé, G. (2016). Refining semi-automatic parallel corpus creation for Zulu to English statistical machine translation. 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), IEEE.

Kunchukuttan, A., et al. (2017). "The iit bombay english-hindi parallel corpus." arXiv preprint arXiv:1710.02855.

Moors, C., et al. (2018). Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. Proceedings of the annual conference of the South African institute of computer scientists and information technologists.

Ndhlovu, K. (2016). "Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele." Literator (Potchefstroom. Online) **37**(2): 1–12.

Scholman, M. C. J., et al. (2025). "DiscoNaija: a discourse-annotated parallel Nigerian Pidgin-English corpus." Language Resources and Evaluation.

Schryver, G.-M. d. and D. J. Prinsloo (2000). "The compilation of electronic corpora, with special reference to the African Languages." Southern African Linguistics and Applied Language Studies **18**: 87–104.

Shoba, F. M. (2018). Exploring the use of parallel corpora in the complilation of specialised bilingual dictionaries of technical terms: a case study of English and isiXhosa, University of South Africa (South Africa).

Siripragada, S., Philip, J., Namboodiri, V. P., & Jawahar, C. V. (2020). "A multilingual parallel corpora collection effort for Indian languages."

Thanaki, J. (2017). Python natural language processing, Packt Publishing Ltd.

Tse, R., et al. (2020). Building an italian-chinese parallel corpus for machine translation from the web. Proceedings of the 6th EAI international conference on smart objects and technologies for social good.

Varga, D., et al. (2008). Parallel corpora for medium density languages. Recent advances in natural language processing IV: selected papers from RANLP 2005, John Benjamins Publishing Company**:** 247–258.

Zungu, T. G. (2021). "Intellectualization of IsiZulu at the University of KwaZulu-Natal through the Development of IsiZulu Terminology and the Implementation of the Doctoral Rule." Alternation, Special Edition 38b: 637–657.