A Qualitative Inquiry into the South African Language Identifier's Performance on YouTube Comments

Nkazimlo Ngcungca, Johannes Sibeko, Sharon Rudman
Nelson Mandela University
University Way, Summerstrand, 6019, South Africa
zizingcungca@gmail.com, {johanness, srudman}@mandela.ac.za

Abstract

The South African Language Identifier (SA-LID) has proven to be a valuable tool for data analysis in the multilingual context of South Africa, particularly in governmental texts. However, its suitability for broader projects has yet to be determined. This paper aims to assess the performance of the SA-LID in identifying isiXhosa in YouTube comments as part of the methodology for research on the expression of cultural identity through linguistic strategies. We curated a selection of 10 videos which focused on the isiXhosa culture in terms of theatre, poetry, language learning, culture, or music. The videos were predominantly in English as were most of the comments but the latter were interspersed with elements of isiXhosa, identifying the commentators as speakers of isiXhosa. The SA-LID was used to identify all instances of the use of isiXhosa to facilitate the analysis of the relevant items. Following the application of the SA-LID to this data, a manual evaluation was conducted to gauge the effectiveness of this tool in selecting all isiXhosa items. Our findings reveal significant limitations in the use of the SA-LID, encompassing the oversight of unconventional spellings in indigenous languages and misclassification of closely related languages within the Nguni group. Although proficient in identifying the use of Nguni languages, differentiating within this language group proved challenging for the SA-LID. These results underscore the necessity for manual checks to complement the use of the SA-LID when other Nguni languages may be present in the comment texts.

Keywords: Language Identity, IsiXhosa, Language Identification, SA-LID

1. Introduction

The global linguistic landscape, comprising approximately 7168 languages, is dynamic and demands continuous exploration (Aitchison, 2005; Trask, 2003). This is particularly true in light of the core role played by language in terms of the social, cultural, intellectual and political vitality in any society (Lo Bianco, 2010). As such, there is a need for continuing research in order to understand the characteristics of each language as well as the cultures and identities that are linked to the concerned linguistic communities.

Given the global linguistic diversity, an ability to distinguish between the languages being used in a particular context is understandably significant (Jaspers, 2015). Such an ability facilitates the decoding of the content of the message and thus fosters effective communication and comprehension (Hardan, 2013). This is particularly significant in a linguistically diverse country like South Africa (Fishman et al., 2008), where most citizens are multilingual (Evans, 2015; Sithole, 2015; Adelani et al., 2021).

Through language, individuals not only communicate but also articulate their origins, making it a fundamental dimension of cultural identity. In this, and many other ways, language and identity are intricately linked (Bucholtz and Hall, 2004). This aspect of language use extends beyond oral expression to also encompass written interactions and specifically so in colloquial contexts which allow for a more spontaneous and free use of language - for example, social media platforms. A tool which has the ability to accurately identify the languages used in a multilingual text could carry numerous benefits and play an important role in a linguistically diverse society. This would be especially true in terms of research focused on the actual use of language by those fluent in more than one language and the manner in which their language use expresses their cultural identity.

Identity, in its simplest form, is an expression of individuality and reflects the uniqueness of every human being (Buckingham, 2008). However, identity is also influenced to a large extent by the social groups to which an individual belongs (Baxter, 2016). This is particularly so in terms of the cultural and linguistic background into which one is born as this is the context in which one first learns about – and learns how to express – aspects of the world (Praeg, 2014). Linguistic and cultural identity are generally conflated and language use is often reflective of these aspects – along with other 'hints' about a speaker's identity (Bucholtz and Hall, 2005). For this reason, the relationship between language usage and cultural/linguistic identity is rife with possibilities.

Research on the link between language use and cultural or linguistic identity in a multilingual con-

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 45–54
25 May, 2024.

text assumes the ability to discern between the languages employed by the language users in that community. This is indeed the case with the research project of which this paper forms a component. The broader study aims to investigate linguistic strategies employed by isiXhosa language users to express their language identities on YouTube through the use of comments.

The scope of this study, therefore, underscores the necessity of a reliable language identifier to accurately detect the language(s) used within a particular text. The use of such a tool becomes indispensable when navigating through the substantial pool of comments in order to extract comments written in isiXhosa or code-switched between isiXhosa and other languages. The identification of instances of isiXhosa usage from our corpus of YouTube comments is thus necessary in order to delineate the data on which our study will focus.

This paper aims to evaluate the reliability of the South African Language Identifier (Puttkammer et al., 2016) when it is applied to a corpus of YouTube comments to ascertain the languages used. The following sections of this paper provide a brief literature review in terms of the core concepts in Section 2, an overview of our methodology for data collection in Section 3 and analysis as well as a summary of our findings in Section 4. The paper concludes with a discussion of our conclusions and recommendations in Section 5.

2. Background

2.1. The isiXhosa Language

While there are between 24 and 30 spoken languages in South Africa (Finlayson and Madiba, 2002), the constitution of the Republic of South Africa recognises 12 of these as official languages (Republic of South Africa, 1996, 2023). These official languages are typically grouped into six language groups, including: (i) South African Sign Language, (ii) Sotho-Tswana, which includes Sesotho, Setswana, and Sepedi, (iii) Sotho-Makua-Venda, which includes Tshivenḓa, (iv) West Germanic, which includes Afrikaans and English, (v) Nguni-Tsonga, which includes Xitsonga, and (vi) Nguni, which includes isiZulu, Siswati, siNdebele and isiXhosa.

The Nguni language group occupies a significant position as the largest language group in South Africa. IsiXhosa, the second-most prominent Nguni language within South Africa, (Wheeler, 2018), is predominantly spoken in the Eastern Cape and the Western Cape Provinces. Notably, it is also officially recognised in Zimbabwe (Republic of Zimbabwe, 2021). According to Wheeler (2018), isiXhosa has much in common with isiZulu in terms of their linguistic roots. In fact, as discussed later in Section 4.3, isiXhosa demonstrates close linguistic ties and mutual intelligibility with other languages in the Nguni group as well (Dyers, 2000). Additionally, isiXhosa stands out for its use of clicks, a feature present in only about 0.5% of the world's languages (Brenzinger and Shah, 2023), including a few Bantu languages. These clicks are represented by the use of three consonants: /c, q, and x/ (Nogwina et al., 2013; Gxowa-Dlayedwa, 2015, 2018; Wheeler, 2018).

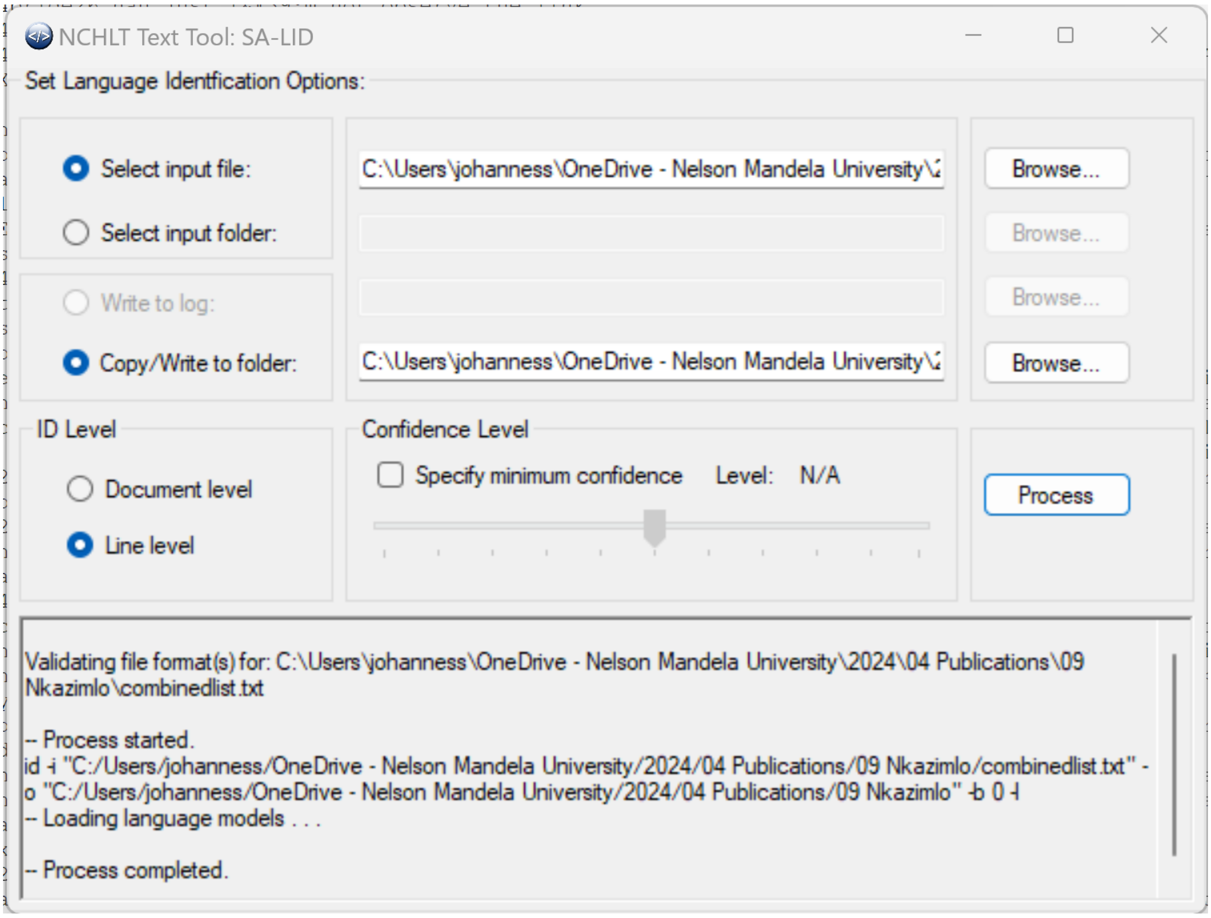
2.2. Identifying Languages

The initial step in comprehending written text is to ascertain the language in which it is written (Babhulgaonkar and Sonavane, 2020). Various language identification tools are developed for this purpose, with the goal of discerning the language(s) present in the text (Jauhiainen et al., 2024). Note that these language identifiers are designed to encompass both speech and written texts. However, due to the inherent differences between written text, composed of discrete characters, and speech, which involves a continuous signal relying on acoustic features, different natural language processing methods are traditionally employed for text and speech, resulting in limited methodological overlap between the two (Murthy and Kumar, 2006; Ambikairajah et al., 2011).

For the purpose of this discussion, our focus is specifically on language identifiers for written texts. Text language identification involves analyzing written linguistic features, including character n-grams and word frequency patterns. This analytical process often makes use of statistical models and machine learning algorithms (Nezhadi et al., 2017).

Traditionally, human beings are regarded as the most accurate language identifiers (Deshwal et al., 2020). Unfortunately, their ability to detect languages is limited by their language repertoires. As such, the limits of relying on humans for language identification become obvious when considering the estimated 7168 languages worldwide (Al-Jarf et al., 2022), or the twelve official languages in South Africa. Simply put, humans are unable to detect languages that are outside their current linguistic repertoires.

As a result, more non-human dependent approaches are needed in the task of identifying languages. Over time, computational approaches employing tailored algorithms and indexing structures have developed to discern language usage without human intervention (Calvo et al., 2017). This evolution includes the use of advanced techniques such as neural networks (Talpur and O'Sullivan, 2020) and Natural Language Processing (NLP) approaches (Saji et al., 2022), which are integrated into language identification tools.



Set Language Identfication Options:

● Select input file: C:\Users\johanness\OneDrive - Nelson Mandela University\2

○ Select input folder:

○ Write to log:

● Copy/Write to folder: C:\Users\johanness\OneDrive - Nelson Mandela University\2

ID Level

○ Document level

● Line level

Confidence Level

□ Specify minimum confidence Level: N/A

```
Validating file format(s) for: C:\Users\johanness\OneDrive - Nelson Mandela University\2024\04 Publications\09
Nkazimlo\combinedlist.txt

-- Process started.
id -i "C:/Users/johanness/OneDrive - Nelson Mandela University/2024/04 Publications/09 Nkazimlo/combinedlist.txt" -
o "C:/Users/johanness/OneDrive - Nelson Mandela University/2024/04 Publications/09 Nkazimlo" -b 0 -l
-- Loading language models . . .

-- Process completed.
```

Figure 1: A screenshot of the SA-LID.

2.3. Language Identification Tools

The introduction of automatic language identifiers serves as a valuable advancement in language detection. The primary task of a language identification tool is to analyse a given spoken or written text and generate a prediction of the language in which the text is spoken or written (Navrátil, 2006; Bergsma et al., 2012; Solorio et al., 2014). This process includes assessing the probability of each word in the provided text as belonging to one or more of the languages in the tool's library (Lui and Baldwin, 2012). The identified language is determined by the highest probability, initiating a competition among language models to determine the most likely match for the entire sample. Like humans, automatic language identifiers also rely on libraries (Agarwal et al., 2023). In this way, the automatic tools need to be trained using different languages and will not be able to detect new languages that are not in their existing libraries.

In this paper, as indicated in Section 1, we conduct a qualitative evaluation of the South African Language Identifier (henceforth SA-LID). The SA-LID was designed to classify text into one of the 11 official written languages of South Africa, either at the document or line level (Puttkammer et al., 2018). The SA-LID has been trained using government text corpora obtained during the National Centre for Human Language Technology (NCHLT) Text project and collected through collaboration between the South African Department of Arts and Culture and the Centre for Text Technology (Puttkammer et al., 2018).

The SA-LID uses feature extraction, identifying language-specific patterns through the analysis of character n-grams, ranging from bigrams to 6-grams. The model was trained using the Multinomial Naive Bayes classifier, incorporating labelled training samples and the selected feature extractor. These components collectively enable the model to discern languages effectively based on the extracted features. In the subsequent step of text classification, the trained classifier is applied to text inputs, resulting in a list of probable languages arranged by their respective probabilities. The final language determination is achieved by selecting the language with the highest probability, based on the model's consideration of learned patterns and characteristics during the training process, ensuring accurate identification of the language in the given text.

Video Identifier	Release Year	Comments	Views-to-Date	Likes-to-Date	Type
iZcx_akfXe4	2010	202	306,211	870k	Documentary
zEoYl4Ok6Ks	2012	107	394,025	2k	Music and Dance
baEiWB2aM9Y	2013	894	1,797,004	17k	Interview
ZnnlJzINWs8	2018	179	426,724	4.9k	Praise Poetry
ZcRykTbiva4	2015	236	373,909	4.7k	Lesson
zOUvWM6Yx3Q	2015	115	521,804	1.7k	Drama
RfcnDHYFETs	2020	266	14,458	345k	Lesson
rjo8h5qLpU0	2020	2549	2,230,543	92k	Music
v4iOTPFz0-c	2021	1538	222,000	3.5k	Documentary
zPM8Qid9VSY	2021	831	121,577	4.6k	Lesson
Total	-	**6885**	**6,127,486**	**126.7k**	

Table 1: Video Statistics (Ordered by Release Year) with Totals

2.4. Related Research

Previous research has explored the application of language identification specifically to isiXhosa texts (Kyeyune, 2015). In their study, Kyeyune (2015) utilised corpora from the Language Resource Management Agency and employed the Java Text Categorising Library to extract *n*-grams for identifying isiXhosa using an *n*-gram language model. The study conducted by Duvenhage et al. (2017) investigated the use of a naive Bayes classifier for accurate language group identification. Additionally, they incorporated a lexicon-based classifier to differentiate the specific South African language in which the text is composed. Furthermore, in their work, Duvenhage (2019) introduces a hierarchical classifier that combines naive Bayesian and lexicon-based approaches for short-text Language Identification (LID). This approach proves particularly beneficial for under-resourced languages.

In this paper, we investigate the reliability of the SA-LID to assess its usability in detecting isiXhosa from YouTube comments, employing a qualitative approach for our discussion.

3. Methodology

A total of ten videos were selected from YouTube using a variety of pre-determined search terms such as: (i) amaXhosa ase South Africa, (ii) Introduction to the Xhosa culture (iii) The History of isiXhosa language, (iv) The history of isiXhosa culture, (v) Clicks used in isiXhosa music, and (vi) isiXhosa language-use in South Africa. The video selection process was based on the relevance to the title of the broader study, as well as evidence of the use of linguistic elements which identified commentators as isiXhosa.

We employed the YouTube API to mine comments from the 10 selected videos for our study. This process involved specifying the video IDs of the chosen content and extracting the associated comments. The API facilitated the extraction of text-based comments and emojis. We excluded information such as user details, timestamps, and other information.

The data collection was conducted on 19 January 2024. We identified videos that were uploaded more than a year before our investigation. As a result, we were not expecting any surge in the new comments on the videos.

3.1. Data Cleaning

During the data cleaning process, we addressed the presence of unexpected characters by replacing them with relevant punctuation. For example, we transformed $(\&\#39;)$ into the apostrophe $(')$, resulting in modifications for a total of 2371 + 83 instances. Additionally, occurrences of $(\&quot;)$ were replaced with (“) and (”), with a total of 730 errors identified and rectified. Furthermore, instances of $(\&lt;3)$ were amended to <3, with twelve occurrences addressed. Finally, we removed line breaks that were indicated by $(< br >)$ as we needed the comments to be counted as one and not to be separated.

This study employed the YouTube Data API to systematically extract comments and replies from specified YouTube videos using associated video IDs. The video IDs are provided in Table 1. An API key was configured for authentication, and a systematic approach was adopted to retrieve comments and replies for each video. The script iterated through the list of video IDs, employing the YouTube Data API to retrieve comments in batches of 100, with pagination support for handling larger datasets. The retrieved comments and replies were processed and organised into a pandas DataFrame for each video, facilitating subsequent analysis.

The dataset initially comprised 6885 lines. However, this figure was reduced after we eliminated duplicate lines. punctuation-only lines (such as question marks), and closing quotation marks (indi-

Language	**Confidence levels**						
	40%	**50%**	60%	70%	80%	90%	99%
Afrikaans	13	**75**	4	1	1	1	-
English	4225	**5215**	2153	896	177	24	-
SiNdebele	14	**37**	6	3	1	1	-
Sepedi	2	**13**	1	1	1	1	-
SiSwati	12	**44**	3	1	1	1	-
Sesotho	3	**20**	2	-	-	-	-
Setswana	4	**15**	-	-	-	-	-
Xitsonga	2	**20**	-	-	-	-	-
TshivVenda	2	**19**	1	-	1	-	-
IsiXhosa	340	**458**	148	75	27	5	-
IsiZulu	194	**318**	85	48	13	3	-
Unsure	1762	**339**	4170	5548	6352	6539	6573

Table 2: Results considering different confidence levels.

cating the end of quotes from preceding lines) and instances involving full stops, numbers and further duplicates. Ultimately, our analysis is grounded in a dataset encompassing 6573 lines, inclusive of both text and emoji comments. The overview of videos and comment counts is provided in Table 1.

The study prioritised user privacy and adhered to the terms of service of the YouTube platform. No personally identifiable information was collected, and the data were used exclusively for research purposes.

3.2. Data Analysis

We considered seven confidence levels available through the Language Identifier (namely 40%, 50%, 60%, 70%, 80%, 90%, and 99%) in order to evaluate the consistency of the findings. In the end, our discussion is based on the results of the default confidence setting for language identification, that is, 50% confidence. The results of the analyses at the different confidence levels are presented in Table 2.

The SA-LID utilises an input file or folder, and the accepted file types are .txt files. It then outputs to either a log file or into a folder. The identity levels include document level and line level. It is important to note that when the line level option is selected, the output setting defaults to the “Copy/Write to folder” option. A screenshot of the interface is illustrated in Figure 1. The output files classify the sentences, so each output file includes only sentences identified as the specified language. The output file names append the language code as a prefix to the original file name. For example, when using the original file name for a bilingual dataset for English and isiZulu, dataset.txt, the SA-LID will output zu.dataset.txt and en.dataset.txt.

4. Qualitative Error Analysis

In evaluating the performance of the SA-LID on our YouTube corpus (identified for our larger project on the language identities of amaXhosa), we employed a qualitative error analysis. Our analysis was based on the default confidence setting, specifically the 50% confidence level (refer to Table 2 for the results of the SA-LID which reflects the different confidence levels from 40% to 99%).

4.1. The Unsure Caterogy

The SA-LID encountered 339 comments which it found uncertain. Upon examination, we identified a few reasons for the uncertainty. The uncertainty arose primarily from the identification of emojis and unfamiliar slang, such as *‘wow’* and *‘yeah’* as well as acronyms such as *‘lol’* and *‘omg’*. We assume that such words were not included in the development data for the SA-LID.

Secondly, a notable challenge emerged as we realised that the SA-LID encountered difficulties in accurately categorising language when spelling mistakes were present. Consequently, a significant number of comments ended up in the unsure category. For instance, one comment under the unsure category featured the misspelling *‘qween*,’ which is appropriately spelt as *‘queen*.’ These instances illustrate that when words are misspelt, it becomes more challenging for the SA-LID to accurately identify the languages. This underscores the critical importance of accurate spelling for the SA-LID to perform effectively in language categorisation.

Thirdly, notable instances of unexpected scripts were observed. For instance, the data contained comments in Japanese, Russian, and Arabic. Such scripts are not official in South Africa and, as such, they are not expected to be identified using the SA-LID.

We also noted that some comments in English were also categorised as unsure. For example, consider the comment:

```
(a) Nice, Lucky you, I am so Jealous.
```

We are unable to account for these results. However, such occurrences prompt questions about whether the majority of words in such comments were absent from the language library used by the identifier.

4.2. Multilingual Comments

Ideally, the SA-LID as outlined earlier, will assign the language based on the higher probability. As an example, consider the sentence below:

```
(b) Awume kancane wena. uShaka
uhlanganaphi nokubaleka kwamaXhosa.
Ehamba nabelungu. Babuya Kuphi?
Asibafuni iningi lethu Kwazulu
KwaZulu. Loyalt to nothing
asibafuni.
```

The SA-LID has categorised this sentence under isiZulu since it contains 15 isiZulu words although it also contains three English words, one of which is spelled incorrectly.

Furthermore, we observed that the SA-LID shows a preference for indigenous South African languages when an equal number of words from multiple languages are present in the same comment. To illustrate, consider the following comment classified under the isiXhosa comments:

```
(c) This woman is talking sh\%t...
lo othi xa ungenamgidi awuyndoda...
mxfm.... The only part I like is
lena athi yimisebenzi yakho
ebonakalisa ubudoda.
```

In this example, the term *'sh%t'* is not recognised as English due to the inclusion of punctuation. Upon tokenisation, the word is divided into three tokens, making it less likely to be identified as a valid English word. Additionally, the term *'mxfm'* is a misspelt word. Consequently, there are only ten English words in the comment. Similarly, the isiXhosa word *'awuyndoda*' is spelt incorrectly, as such there are also ten isiXhosa words. Thus, the comment contains an equal number of English and isiXhosa words but even so, the SA-LID identified the comment as isiXhosa, thereby illustrating a preference for isiXhosa.

In more extreme circumstances, the SA-LID identified code-switched comments that are predominantly English under indigenous languages such as isiXhosa. To illustrate, consider the example below:

```
(d) Singabantu abanye.
Xhosas from Zim moved to Zim
from the Eastern Cape with Cecil
John Rhodes, Ndicelumenywa.
```

This code-switched comment exhibits a prevailing use of English, interspersed with three isiXhosa words. The classification of this sentence as isiXhosa reinforces our inference that the SA-LID tends to favour indigenous languages when categorising code-switched texts.

Note that the SA-LID has no category for sentences that are multilingual. This is particularly problematic in the context of South African multilingual social media. That is, there is a need for an additional category in terms of those sentences considered 'multilingual' (rather than assigning them to one of the two language groups present in the sentence). The ability to identify the use of more than one language within a single text effectively ensures an alignment with real-life language use. However, the SA-LID currently identifies at least one language from the comment and then assigns a language label rather than noting the sentence as 'multilingual. Nonetheless, this ability to classify code-switched texts is a significant asset for our study which focuses on how amaXhosa articulate their linguistic identity.

Our aim in this paper was to investigate the ability of the SA-LID to identify comments in isiXhosa. Overall, the SA-LID was able to identify instances of the use of isiXhosa including sentences that are purely in isiXhosa and those that are code-switched.

In the larger study on language identities, we also hope to identify and analyse strategies used by multilingual commentators in their interactions on YouTube as a social media platform. Consequently, the accurate identification of isiXhosa through the SA-LID holds particular significance for our research objectives, facilitating the exclusion of comments lacking isiXhosa content.

4.3. Mutual Intelligibility

In our analysis, we observed challenges for the SA-LID in distinguishing between similar languages from the same language group. For instance,isiZulu and isiXhosa share some characteristics, enabling speakers of one language to understand the other due to their akin dialects.

While there are some similarities in vocabulary stemming from their common Bantu origin, specific words differ between isiXhosa and isiZulu. The table below provides an illustrative example:

Despite these distinctions, the SA-LID encountered difficulty and misidentified some texts written in isiZulu as isiXhosa. For instance, consider the following example:

English	isiXhosa	isiZulu
I want (it)	Ndiyayifuna	Ngiyayifuna
I noticed that/it	Ndiyibonile	Ngiyibonile
I am happy	Ndiyavuya	Ngiyajabula
We appreciate	Siyakuvuyela	Siyakujabulela

```
(e) Ngiyalithanda isiko lamaXhosa,
thanks for this content bhudi''
```

In this example, the term *'ngiyalithanda'* is of isiZulu origin, while the isiXhosa equivalent would be *'ndiyalithanda'*. We suspect that the confusion might have arisen due to the inclusion of the term *'lamaXhosa'* in the sentence. Nevertheless, the term *'isiko'* can be identified in either of the two languages. Furthermore, examples such as:

```
(f) ``nazoke ezakuthi madoda''
(g) ``gaaa ! hlala phansi.''
```

The first example was identified as isiNdebele, while the second example was identified as Siswati. While these may be correct, the same sentences could be identified as other languages in the Nguni group too. For instance, while the use of the word *'ezakuthi'* in the first example rules out isiZulu, which would be *'ezakithi'*, it can be identified as isiXhosa. However, the second example could be isiZulu because of the word *'phansi'*, whose equivalent in isiXhosa is spelled *'phantsi'*. Note that local dialects may actually identify these languages as either one in the group based on language contact influences. Nonetheless, these examples demonstrate the mutual intelligibility of the languages. Furthermore, this illustrates that a thorough manual check is necessary to distinguish between the Nguni languages before commencing an official analysis, as the SA-LID may be confounded by the linguistic similarities.

4.4. Assumed linguistic and Cultural identities

As we analyzed the comments, we observed a diverse array of languages employed by commentators, including code-switching between indigenous South African languages as well as the unexpected occurrences of Japanese, Russian, and Arabic. We inferred that individuals who use both monolingual and multilingual sentences were concurrently expressing both their thoughts and cultural identities. Drawing on the insights of scholars like Bucholtz and Hall (2004) and others who explore the intricate relationship between language and identity, our findings suggested that commentators were strategically situating their linguistic and cultural identities through their language use.

Nevertheless, we acknowledge the inherently multilingual nature of the world, where individuals can learn languages beyond those spoken at home. According to Kinginger (2004), when individuals speak and learn a new language, they simultaneously adopt a new identity or engage in the reconstruction of their existing one. This concept is illustrated by Johanson Botha (2009)'s example of an English man learning isiXhosa. When he speaks isiXhosa, he becomes loud, which causes embarrassment to his wife. This loudness, not commonly associated with Western culture, is stereotypically linked to isiXhosa culture, portraying the construction of amaXhosa as assertive or loud. Therefore, we understand that to definitively ascertain whether someone identifies as isiXhosa or any other language, further investigation (for example conducting interviews) would be imperative.

Given this context, the primary investigation of our broader study will focus on conducting interviews to confirm linguistic identities. This was not, however, necessary for this paper as the objective was solely to assess the accuracy of the SA-LID when applied to a corpus of YouTube comments.

5. Conclusion

In this paper, we explore the use of automatic language identification using the South African Language Identifier (SA-LID) to discern languages within a YouTube comments corpus. This study forms part of a broader project aiming to uncover linguistic strategies employed by isiXhosa speakers in expressing their language and cultural identities in YouTube comments. As digital platforms continue to shape communication patterns, understanding language identities becomes crucial for fostering inclusive and accurate representation. To facilitate this, there is a need for accurate language identification in multilingual texts. As such, the context of our broader study led us to evaluate the reliability of the SA-LID in identifying any use of isiXhosa language elements in the relevant comments which we mined from YouTube. The question which underpinned our research, as reflected in Section 1, related to whether we could rely on the language identification results generated through the use of the SA-LID to accurately identify all instances of the use of isiXhosa.

Our analysis of the SA-LID reveals both strengths and challenges. The tool demonstrates proficiency in identifying languages used in multilingual comments, showcasing its versatility in capturing dynamic language use within the amaXhosa community on YouTube. This aspect prompts us to conclude that the SA-LID can indeed be effectively employed in situations where two languages coexist or code-switching occurs, showcasing its robust capabilities in language categorisation.

However, challenges arise, particularly in cases of mutual intelligibility between closely related languages like isiXhosa and other Nguni languages. This highlights the complicated nature of language identification and, therefore, urges further exploration.

Other challenges encountered with this tool include uncertainties related to emojis, slang, unconventional spelling and spelling errors. This emphasises the need for continuous refinement in language identification tools to accommodate diverse linguistic expressions. In the context of our corpus, the non-Latin scripts in the dataset further complicated language identification as they are unexpected in the South African context.

Our findings contribute to the ongoing discourse on language use and identity in digital spaces, offering insights into methodologies which can be employed in further research. The misidentification of languages, as noted in this study, opens up opportunities for future studies to explore how the choice of words or phrase structure in a text can potentially confuse a language identifier. In this study, we did not delve into grammatical complexities or sentence structures; our primary focus was to ascertain the ability of the SA-LID to identify the use of isiXhosa from written YouTube comments accurately. We acknowledge that potential issues may have arisen from variations in pre-processing steps. Specifically, different processes may have been employed for tokenisation, text normalisation, and handling special characters compared to those used in the training of the SA-LID.

6. Bibliographical References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

DOSUL ÁFRICA. 2020. Constitution of the republic of south africa, 1996. *As adopted*, 704:705.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. Limit: Language identification, misidentification, and translation using hierarchical models in 350+ languages. *arXiv preprint arXiv:2305.14263*.

Jean Aitchison. 2005. Language change. In *The Routledge Companion to Semiotics and Linguistics*, pages 111–120. Routledge.

Reima Al-Jarf et al. 2022. Text-to-speech software for promoting efl freshman students' decoding skills and pronunciation accuracy. *Journal of Computer Science and Technology Studies*, 4(2):19–30.

Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu. 2011. Language identification: A tutorial. *IEEE Circuits and Systems Magazine*, 11(2):82–108.

Arun Babhulgaonkar and Shefali Sonavane. 2020. Language identification for multilingual machine translation. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 401–405. IEEE.

Judith Baxter. 2016. Positioning language and identity: Poststructuralist perspectives. *The Routledge handbook of language and identity*, pages 34–49.

Kenneth R Beesley. 1988. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th annual conference of the American Translators Association*, volume 47, page 54.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the second workshop on language in social media*, pages 65–74.

Matthias Brenzinger and Sheena Shah. 2023. A typology of the use of clicks. *Stellenbosch Papers in Linguistics Plus*, 67(1):59–77.

Mary Bucholtz and Kira Hall. 2004. Language and identity. *A companion to linguistic anthropology*, 1:369–394.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

David Buckingham. 2008. *Introducing identity*. MacArthur Foundation Digital Media and Learning Initiative.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Deepti Deshwal, Pardeep Sangwan, and Divya Kumar. 2020. A language identification system using hybrid features and back-propagation neural network. *Applied Acoustics*, 164:107289.

Bernardt Duvenhage. 2019. Short text language identification for under resourced languages. *arXiv preprint arXiv:1911.07555*.

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. 2017. Improved text language identification for the south african languages. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, pages 214–218.

Charlyn Dyers. 2000. *Language, identity and nationhood: Language use and attitudes among Xhosa students at the University of the Western Cape, South Africa*. Ph.D. thesis, University of the Western Cape.

John Edwards. 2009. *Language and identity: An introduction*. Cambridge University Press, New York.

Moyra Sweetnam Evans. 2015. Language use and language attitudes in multilingual and multicultural south africa. *Heritage and exchanges: Multilingual and intercultural approaches in training context*, pages 43–62.

Rosalie Finlayson and Mbulungeni Madiba. 2002. The intellectualisation of the indigenous languages of South Africa: Challenges and prospects. *Current issues in language planning*, 3(1):40–61.

Joshua A Fishman, Monica Barni, and Guus Extra. 2008. *Mapping linguistic diversity in multicultural contexts*. Mouton de Gruyter.

Ntombizodwa Gxowa-Dlayedwa. 2018. Investigating click clusters in isixhosa syllables. *South African Journal of African Languages*, 38(3):317–325.

Ntombizodwa Cynthia Gxowa-Dlayedwa. 2015. Ukufundisa izicuku zeziqhakancu emagameni. *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 31(3):32–48.

Abdalmaujod A Hardan. 2013. Language learning strategies: A general overview. *Procedia-Social and Behavioral Sciences*, 106:1712–1726.

Jürgen Jaspers. 2015. Modelling linguistic diversity at school: the excluding impact of inclusive multilingualism. *Language Policy*, 14:109–129.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Tommi Jauhiainen, Marcos Zampieri, Timothy Baldwin, and Krister LindÚn. 2024. Introduction to language identification. In *Automatic Language Identification in Texts*, pages 1–17. Springer.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Liz Johanson Botha. 2009. ‘them and us’: Constructions of identity in the life history of a trilingual white south african. *African Identities*, 7(4):463–476.

Celeste Kinginger. 2004. Alice doesn’t live here anymore: Foreign language learning and identity reconstruction. *Negotiation of identities in multilingual contexts*, 21(2):219–242.

Michael J Kyeyune. 2015. Isixhosa search engine development report. Technical report, University of Cape Town.

Joseph Lo Bianco. 2010. The importance of language policies and multilingualism for cultural diversity. *International Social Science Journal*, 61(199):37–67.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proceedings of the 5th workshop on language analysis for social media (LASM)*, pages 17–25.

Thembinkosi Mtonjeni. 2013. *An investigation of discriminatory language used in communicating with South Africans born in Tanzania and Zambia*. Ph.D. thesis, Stellenbosch: Stellenbosch University.

Kavi Narayana Murthy and G Bharadwaja Kumar. 2006. Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(01):57–80.

Jirı Navrátil. 2006. Automatic language identification. *Multilingual speech processing*, pages 233–272.

Mohammad M Alyan Nezhadi, Majid Forghani, and Hamid Hassanpour. 2017. Text language identification using signal processing techniques. In *2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*, pages 147–151. IEEE.

Mnoneleli Nogwina, Zelalem Shibeshi, and Zoliswa Mali. 2013. Towards developing a stemmer for the isixhosa. In *WIP. SATNAC Conference*.

Bonny Norton. 2010. Language and identity. *Sociolinguistics and language education*, 23(3):349–369.

Jacobus Christiaan Oosthuysen. 2016. *The grammar of isiXhosa*. African Sun Media.

Leonhard Praeg. 2014. *A report on Ubuntu*. University of KwaZulu-Natal Press Pietermaritzburg.

Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. 2018. Nlp web services for resource-scarce languages. In *Proceedings of ACL 2018, System Demonstrations*, pages 43–49.

Republic of South Africa. 1996. *Constitution of the Republic of South Africa*. Department of Justice, Pretoria.

Republic of South Africa. 2023. *Constitution Eighteenth Amendment Bill*. Department of Justice and Correctional Services, Pretoria.

Republic of Zimbabwe. 2021. *The Constitution of Zimbabwe*. Veritas, Harare.

Lourdes C Rovira. 2008. The relationship between language and identity. the use of the home language as a human right of the immigrant. *REMHU-Revista Interdisciplinar da Mobilidade Humana*, 16(31):63–81.

Ami Katherine Saji, Laura Morales, and Meredith Winn. 2022. *Feasibility report on setting up a collection on questionnaires realting to Ethnic and Migrant Minorities in the European Question Bank (Version 1.0)*. Ph.D. thesis, Sciences Po, CEE.

NE Sithole. 2015. *The functional viability of Indigenous African Languages in South Africa: challenges and prospects of their survival*. Ph.D. thesis, University of Zululand.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Bandeh Ali Talpur and Declan O'Sullivan. 2020. Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter. *Informatics*, 7(4):52–74.

Robert Lawrence Trask. 2003. *Language: the basics*. Routledge.

Mihoko Wheeler. 2018. Phonetic analysis of clicks, plosives and implosives of isixhosa: A preliminary report. *Florida Linguistics Papers*, 5(2).

Marcos Zampieri. 2016. *Pluricentric languages: automatic identification and linguistic variation*. Ph.D. thesis, Universität des Saarlandes, Sao Paulo.

7. Language Resource References

Martin Puttkammer and Justin Hocking and Roald Eiselen. 2016. *NCHLT South African Language Identifier*. South African Centre for Digital Language Resources. PID https://repo.sadilar.org/handle/20.500.12185/350.