Using Content Analysis to Explore AmaXhosa Language Identities in Social Media Texts

Nkazimlo Ngcungca
University of Western Cape
Robert Sobukwe Rd, Bellville, Cape Town
6509848@myuwc.ac.za

Abstract

In the age of increasing digital participation, the role of social media in shaping and revealing identity has become an important area of scholarly inquiry. This article explores the viability of content analysis as a method for inferring identity markers of amaXhosa in multilingual online discourse, focusing on isiXhosa- and English-dominant YouTube comments. Drawing on the theoretical framework of performativity, the study examines how amaXhosa construct and express cultural and linguistic identities through language use in digital spaces shaped by English dominance. A curated selection of 80 YouTube videos related to isiXhosa culture, interviews, and pranks was analysed. Comments were mined using the YouTube API, and the South African Language Identification Tool (SA-LID) was applied for language categorisation. By grounding its approach in local language use and identity, the study contributes to African sociolinguistics and the broader field of African Digital Humanities, demonstrating how social media data can inform context-identity research in multilingual societies.

1 Introduction

The rapid growth of digital platforms such as X , Facebook , and YouTube has transformed how people communicate, share cultural knowledge, and express aspects of their identities. Thus, informal, user-generated texts become increasingly central to everyday life. Even so, there is a growing debate around the usability of social media texts in identity research. At the same time, studies outside South Africa have contributed extensively to digital identity research (Knight and Weedon, 2014; Androutsopoulos, 2015; Darvin, 2016; Baldauf et al., 2017; Jakaza, 2022). Questions remain, especially in African contexts, about whether social media texts offer reliable insights into identity expression. In my broader Master of Arts study (see Ngcungca (2025)), I engaged with this question and demonstrated that such discourses can indeed be used reliably for language identity research.

I base the research in this article on the theory of performativity, which is typically used in gender studies (Butler, 1999, 2025). Through this lens, I view identity not as a fixed trait but as something that is shaped through language use. For me, identity is enacted through repeated linguistic and discursive choices rather than something inherent or static (Butler, 2007; Pennycook, 2007; Benzehaf, 2023).

1.1 Background

To better understand how language relates to identity formation, I first outline my conceptualisation of the self. As Meng et al. (2024) argues, identity is not derived from a single source but emerges from a constellation of affiliations, ranging from sporting and religious communities to national identity, family, and linguistic repertoire. Each of these dimensions contributes meaningfully to an individual's evolving sense of self.

In this sense, linguistic and cultural identities are closely intertwined. Language use often reflects this interplay, serving as both a marker of belonging and a resource through which individuals signal other dimensions of identity(Bucholtz and Hall, 2005; Benzehaf, 2023).

Despite this recognition, much identity research remains bound to curated, edited texts and formal interviews (Bucholtz and Hall, 2005; Norton, 2010; Praeg, 2014; Baxter, 2016; Léglise and Migge, 2021). These studies, while valuable, may fall short in capturing the fluidity of multilingual and informal expressions found in everyday digital discourse. This is particularly significant in South Africa, where English dominates public life. Especially in African multilingual settings where digital identity performances are underexplored.

This article joins the conversation by examining

how linguistic identity is signalled and negotiated through YouTube comments in South Africa. I demonstrate that content analysis can be a viable method for inferring identity markers, grounding this in findings from my MA research on identity expression through language use by amaXhosa see (Ngcungca, 2025).

1.2 Research Questions

In this article, I aim to demonstrate how content analysis can be used to infer identity markers by examining how identity is constructed and made visible in multilingual online spaces, particularly within the English-dominant South African linguistic landscape. This article contributes to emerging scholarship on digital discourse, African sociolinguistics, and the study of language and identity in contemporary society.

The central question guiding this study is: *Can content analysis be used to infer identity markers of Xhosa identity in YouTube comments ?* In other words, this study asks: How do amaXhosa construct, negotiate, and signal their cultural and linguistic identities through language use in multilingual YouTube comments?

According to White and Marsh (2006) and Devi (2009), content analysis enables researchers to analyse the presence, meanings, and relationships of words and concepts and to infer broader messages, cultural elements, and temporal influences embedded within the data. I thus find it useful in this article.

This article is structured as follows: The introduction (this section) provides an introduction by presenting the theory, the background, and outlining the guiding research question. This is followed by a review of related work in Section 2. Section 3 outlines the methodology. Section 4 presents the findings, Section 5 engages with their implications, and Section 6 concludes the article.

2 Related work

2.1 Methodologically-related Literature

Content analysis is a well-established method for analysing data, particularly in its evolving application to textual material (Kleinheksel et al., 2020). According to Devi (2009), content analysis involves the systematic examination of text documents, while Gheyle and Jacobs (2017) expand this understanding to encompass the analysis of unstructured content, including text, images, symbols, or audio data.

Before conducting content analysis, a researcher must address six fundamental questions as outlined by Krippendorff (2004) and Mayring (2021): (i) Which data are analysed? (ii) How are the data defined? (iii) What is the population from which the data are drawn? (iv) What is the relevant context for analysis? (v) What are the boundaries of the analysis? and (vi) What is the target of the inference?

Accordingly, I present my analysis below,

- The data analysed comprise YouTube comments extracted from selected publicly available videos relevant to the research topic.
- These comments are defined as user-generated texts expressing various reactions, opinions, or sentiments.
- The population includes commenters who interact with the chosen content, assumed to be part of or familiar with the broader amaXhosa community.
- The relevant context is the multilingual and digital nature of YouTube, where users express themselves in multiple languages, often mixing isiXhosa and English.
- The boundaries of the analysis are limited to written comments posted under the selected videos.
- The target of the inference is to identify and interpret linguistic markers that may point to expressions of cultural and linguistic identity among amaXhosa in this online space.

Content analysis is a widely applied research method across numerous disciplines, including mass media (Ahmed and Matthes, 2017; Reifegerste and Wiedicke, 2023), small-group research (Bonito, 2004; O'Hagan et al., 2021) , instructional communication (Goodboy, 2011; Alhamami, 2023), and public relations (Taylor and Kent, 2010). Early studies by Osgood and Walker (1959) and more recent work by Synnott et al. (2018) illustrate how content analysis can illuminate emotional and psychological dimensions of language use, such as through the examination of suicide notes. Similarly, De Laat's (2023) analysis of online communities for seniors reveals how digital communication reflects social identities and collective experiences.

Filename	Line	Words	Emojis	Punctuation
af.txt	475	2144	858	1437
en.txt	23412	129871	38254	99776
nr.txt	251	869	31	994
nso.txt	377	905	372	627
ss.txt	375	1799	686	1177
st.txt	211	1010	407	837
tn.txt	105	436	396	374
ts.txt	337	457	322	501
ve.txt	2433	2347	9617	11228
xh.txt	2247	14648	276	433
zu.txt	1810	3760	5666	11439
unsure.txt	2820	2789	10279	12224
Total	34853	161035	67164	141047

Table 1: Text analysis generated after language identification.

Together, these studies demonstrate how content analysis can be used to explore emotional, psychological, and social dimensions of language use. Building on this foundation, the current study applies content analysis to social media texts to infer identity markers, specifically examining how amaXhosa speakers express aspects of cultural identity through their linguistic practices in YouTube comments.

In this study, quantitative content analysis is employed to identify the most frequently occurring words, under the premise that high frequency may signal meaningful linguistic patterns. I hope to infer how recurring linguistic patterns, such as frequent use of particular words, phrases, or expressions, reflect shared cultural references, practices of self-identification, and linguistic repertoires associated with Xhosa identity.

3 Methodology

3.1 Comment mining

A total of 80 videos were selected from YouTube using a set of predetermined search terms, including: (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 selection process was guided by the relevance of each video to the broader focus of the study, with additional attention to identifiable linguistic elements in the content or comments that marked contributors as isiXhosa speakers.

The procedure entailed cataloguing the unique video IDs of selected material and subsequently extracting the text-based comments using YouTube API. This extraction included both comments and emojis, while explicitly excluding user metadata, timestamps, and other ancillary data.

3.2 Data-Cleaning

In the data-cleaning process, the presence of unexpected characters was systematically addressed by replacing them with appropriate punctuation. For instance, the sequence (39;) was converted to an apostrophe (’), resulting in corrections across 7,729 instances. Similarly, occurrences of (quot;) were replaced with opening and closing quotation marks (“ ”), with a total of 3,542 errors identified and rectified. Additionally, instances of (lt;3) were amended to (lt;br>).

3.3 Identifying comments for further analysis

The South African Language Identifier (SA-LID) was employed in this study to categorize the languages present in YouTube comments . The analysis focuses on the 2 languages, but the lang identification does not focus on any specific language. Prior to its application, a pilot study was conducted to assess the tool’s feasibility (Ngcungca et al., 2024).

Table 1 presents the results of language identification, detailing the identified languages, the number of comments per language, total word counts, and counts of emojis and punctuations within each language file.

Despite the file categorisations, code-switching remained prevalent across comments. Notably, the SA-LID classified 2,820 lines as uncertain. Subsequent analysis attributed this uncertainty primarily to the presence of emojis, informal slang terms such as *wow* and *yeah*, and acronyms including *lol* and *omg*. However, this ambiguity was not problematic for the study’s objectives, which focuses on isiXhosa-dominant comments (see `xh.txt` in Table 1) and English-dominant comments (see `en.txt` in Table 1). The characterisation of ’dominant’ acknowledges the inherently multilingual nature of the comments, wherein some isiXhosa-labelled comments incorporated English elements and other languages like Sesotho and isiZulu.

I preprocessed the text through the normalisation of case and the removal of punctuation to prevent extraneous tokens, such as contractions like *I’m* to *im*,— from skewing frequency counts. I also retained, stop words due to the lack of a validated stop word lexicon for isiXhosa social media texts. Note that in multilingual texts such as those explored in this article, certain stop words in English

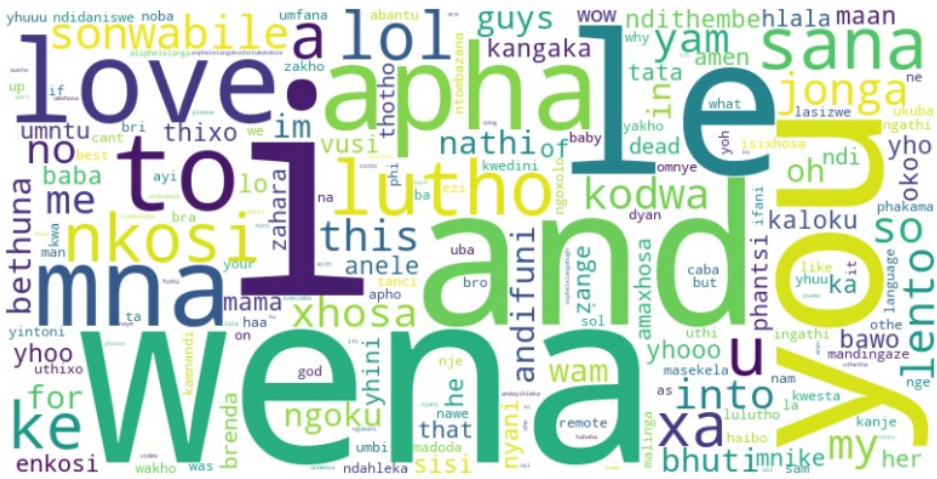

Figure 1: Frequently used words in the isiXhosa comments.

may possess semantic value in isiXhosa or contribute structurally to word formation and syntactic meaning; for instance, the English stop word *i* can appear embedded within isiXhosa particles essential for grammatical coherence. Accordingly, all English lexical items were preserved in the analysis.

3.4 Data analysis procedure

The analytical process in the main study, see (Ngcungca, 2025), encompassed three stages: content analysis, thematic analysis, and discourse analysis. In this article, I present results from content analysis. In the larger study, I relied on content analysis to ascertain the presence of elements pertinent to identity expression within the datasets. This initial procedural step, akin to many distant reading methodologies, functioned to evaluate dataset suitability, identify linguistic markers indicative of identity, and develop an intimate familiarity with the textual material before traditional qualitative interpretive analysis.

4 Findings

4.1 isiXhosa

In this section, I explore isiXhosa comments to analyse how commenters expressed their identities as amaXhosa. The section begins with the investigation of the most frequently used words, followed by the discussion of the most frequently used collocations, concluding with an investigation of word co-occurrences.

4.1.1 Most frequently used words

First, I investigate the top 50 most frequently used words in the isiXhosa comment dataset, illustrated in Figure 1 above, and Figure 5 in the Appendix.

Figures 2 and 5 present a word cloud and a bar graph of the top 50 most frequent words, with larger words like *wena* (you) and *love* indicating higher occurrences. The data shows multilingualism, mixing English (e.g., *people, man, beautiful, love*) and isiXhosa words (e.g., *wena, bhuti, enkosi*), reflecting code-switching and diverse language use.

The word *im* (57 occurrences) may indicate identity positioning, aligning with principles of social identity theory. Similarly, the possessive pronouns *yam* (mine) and *wam* (mine), each appearing 37 times, point to notions of ownership. This assumption of ownership was supported in Ngcungca (2025), where I demonstrated through qualitative analysis how these terms are used to express identity and ownership within the comments.

4.1.2 Frequently collocated words

In this section, I investigate collocations of the most frequently appearing words in the data. The top 50 collocated words are presented in Figure 6 in the Appendix.

The content analysis is aimed at inferring the ways in which amaXhosa express their identities through language use in the dataset. Thus, while words like *nkosi ndithembe* [Lord, I trust you] ap-

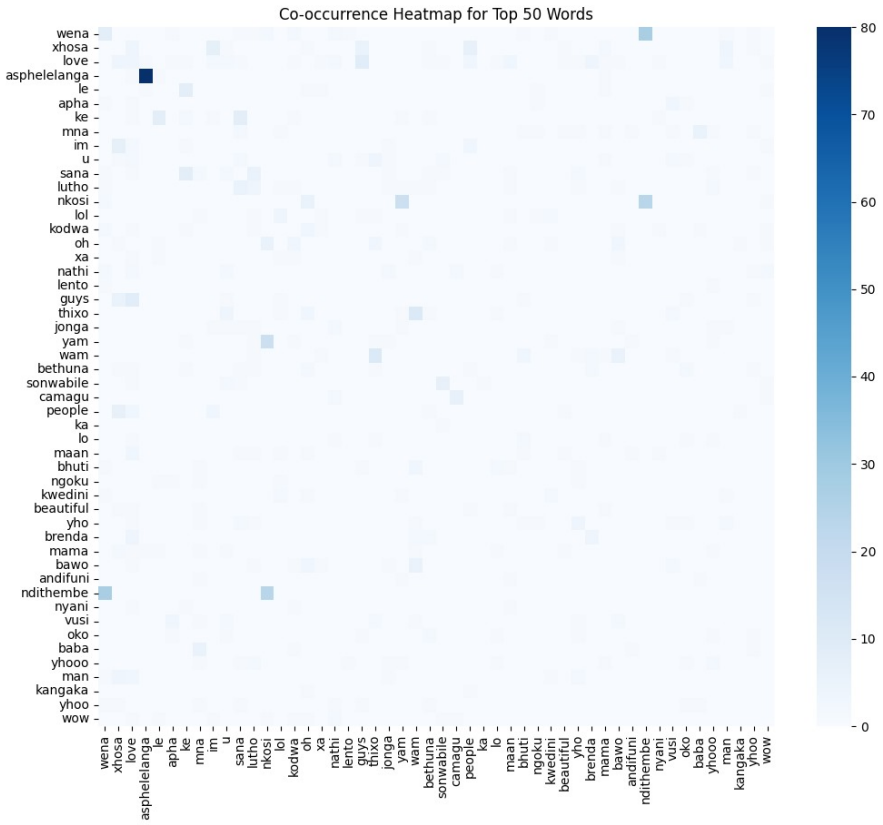

Figure 2: Co-occurrences of the 50 most frequently used words in the isiXhosa comments.

pear 23 times and others appear prominently in Figure 6, they remain irrelevant as they do not provide sufficient cues for identity expression. Instead, it seems that these phrases are primarily lyrics from a song - as such one cannot assume that they are expressions of identity other than references to a song or artist.

To infer identity markers of amaXhosa in data set, I paid special attention to the use of the word *us* and the words with which it frequently collocates. In short, us appears frequently in collocation with words like *us Xhosas* (6), which reveals solidarity and a more direct identity expression strategy, which emphasises a collective identity while communicating the commenter's identity.

Additionally, strings such as *isiko lethu* (our tradition) (6) highlight ownership and cultural pride. These words suggest strategies, which amaXhosa might be using to express their social identities, reinforcing a sense of belonging and ownership to an isiXhosa in-group.

Figure 6 reveals the bigram *Camagu maXhosa* (6). The cultural term *camagu* suggests that the commenter may be expressing their identity through it. *Camagu* is an isiXhosa interjection, used to express appeasement, calmness, forgiveness or soothing (Tshabe and Shoba, 2006).

It serves a purpose similar to amen in the Hebrew language. By addressing amaXhosa directly with camagu, the commenter demonstrates familiarity with the proper way to refer to amaXhosa collectively, implying that they identify as amaXhosa.

4.1.3 Word co-occurrences

To further understand the relationships between the frequently used words, I used a heatmap to study co-occurrences between the most frequently used words in the data. As can be observed in Figure 2, the top 50 words are presented on the left (y-axis) and the bottom (x-axis). Note that darker shades indicate higher frequencies of these co-occurrences while lighter shades represent less frequent relationships.

In Figure 2, several observations can be made. For instance, the strength of the co-occurrences between *man* and *Xhosa* may suggest that commenters are either referring to themselves as Xhosa

Figure 3: Frequently used words in the English comments.

men or discussing Xhosa men in general.

Another noteworthy pairing is *Xhosa* and *love*. This combination could indicate that commenters are expressing affection for either the content itself or the Xhosa language and culture. It is important to explore whether the word *love* here is used as an expression of pride, emotional attachment, admiration, or sarcasm.

In this section, I evaluated the suitability of the isiXhosa data subset for investigating identity expression. The findings indicate that certain words and specific terms are indicators of an individual's identity, based on their language use.

4.2 English

In this section, English comments are analysed to explore how commenters might express their identities. This examination provides insights into how language reflects individual and collective identities in the data.

4.2.1 Most frequently used words

First, I examine the top 50 most frequently used words in the English comments illustrated in the wordcloud in Figure 3. Like in Section 4.1, I also present the list of frequent words in a bar graph in Figure 7 to illustrate the different frequency rates.

The graph in Figure 7 reveals *love* as the most frequently used word with (196) appearances, indicating a high level of expressing affection, assumedly towards a particular topic, video or person. Adjectives such as *best* (68), *great* (62), and *beautiful* (59) have also been used often, presumably to describe something (perhaps the language or the people who speak it as many of the videos have a cultural focus).

Based on these observations, the analysis of frequently occurring words suggests the potential for employing the English sub-dataset in identity-related analyse.

4.2.2 Frequently collocated words

Furthermore, I also examine collocates, which are the immediate neighbours of these words in the dataset.

The results of the collocates presented in Figure 8 reveals the presence of *South African* (6) demonstrating a strong connection to South Africa, suggesting that much of the conversation may have occurred among South Africans or at least about South Africa. There is no evidence that the focus on South Africa is linked to amaXhosa individuals.

In comparison, phrases like *Xhosa people* (10) and *Xhosa guys* (5) are more specific as they refer to a particular ethnic group. This specificity could indicate references to a collective identity in the group. Thus, the presence of collocates such as *Xhosa people* and *Xhosa guys* suggests the dataset's potential for analysing third-person identity constructions. Figure 8 also reveals minimal occurrences, such as *rest peace* (4) and *come back* (5), which do not contribute meaningfully to identifying the speaker's identity.

4.2.3 Word co-occurrences

Finally, I examined the relationships between frequently occurring words. The results are presented in Figure 4.

The word *Xhosa* co-occurs with positive adjectives, such as *style*, *perfect*, *beautiful*, *tremendous* and *pure*. These words seem to place Xhosa in a positive light, potentially favourably defining the culture and language. It would be valuable to explore whether these adjectives are used to describe the language or the people who identify as Xhosa. A more qualitative, manual analysis is needed to determine how these collocates are used and whether they serve as expressions of identity for the audience.

5 Discussion

The isiXhosa and English data subsets reflect dominant language use but are not mutually exclusive. To allow for fair comparison, I followed Miller (2021) guide to normalise frequencies per 1,000 words. The English dataset contains 129,871 words, while the isiXhosa dataset comprises 14,648 words. For instance, *love* appears 196 times in English ($196 \div 129{,}871 \times 1{,}000 = 1.51$) and 88 times in isiXhosa ($88 \div 14{,}648 \times 1{,}000 = 6.01$), showing that it is proportionally more frequent in isiXhosa.

The discussion begins with a comparison of the most frequent isiXhosa and English words, followed by an analysis of the common word strings in both datasets.

5.1 IsiXhosa against English Frequent Words

The Word clouds in Figures 5 and 3 illustrate the prominence of identity-related terms in both datasets. In the isiXhosa comments, code-switching is common, with English words like *people* (2.04) and *beautiful* (1.91) appearing alongside isiXhosa terms such as *mna* (me), *wena* (you)(6.92), and camagu (2.31). This reflects linguistic flexibility and the multilingual strategies used to express social identities, suggesting that online discourse could be a site where cultural and linguistic boundaries are negotiated and identity is actively performed

5.2 Frequent IsiXhosa against English collocates

The bigram analysis reveals differences in identity expression across the two data sets. In the English data Figures 8, collocations like *best episode* (6.54), *feel like* (8.32), and love song (12.67) are common but do not strongly reflect identity. In contrast, the isiXhosa data Figures 6 shows expressions of solidarity and cultural identity, such as *Xhosa guys* (5), *us Xhosas* (6), and *mna ndinguXhosa* (I am Xhosa) (10.09). While English commenters often use I'm with adjectives like *proud, mesmeriSed*, and *rude* to convey emotions, isiXhosa speakers use similar phrases to express both individual stance and group affiliation.

6 Conclusion

This article examined whether language identities can be observed through the use of a content analysis of social media texts. Specifically, I looked at how identity is constructed and signalled through isiXhosa- and English-dominant YouTube comments, with a specific focus on speakers affiliated with the amaXhosa community. By applying content analysis to user-generated digital texts, the study demonstrated that online discourse can offer meaningful insights into how individuals perform, negotiate, and attribute identity in contemporary South Africa.

A key contribution of this study lies in its methodological orientation. That is, while previous research has often relied on curated, edited texts or structured interviews, this article shows that informal digital spaces—particularly multilingual social media comment sections—can serve as viable sites for identity research.

Notably, Krippendorff (2004) conceptualises content analysis as a research technique designed to make replicable and valid inferences from texts within their contexts of use. This study affirms the potential by applying content analysis to a multilingual dataset drawn from YouTube, thereby extending the technique into underexplored, culturally specific digital terrains.

Although to a limited extent, the findings also highlight how performativity theory can be productively applied to digital discourse, where language choices—whether in isiXhosa, English, or a mix of both—are acts through which identity is enacted. By doing so, the study contributes to ongoing conversations about the intersection of language, identity, and digital culture, particularly within the South African context.

This approach not only adds to the growing field of digital sociolinguistics but also offers a scalable method for future research on identity in multi-

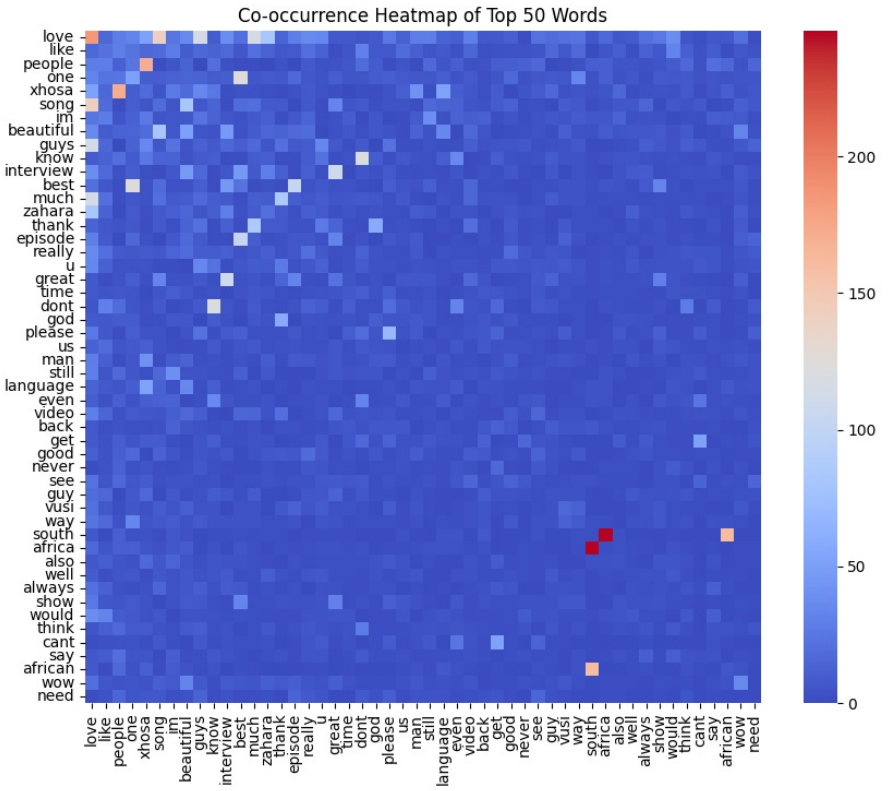

Figure 4: Co-occurrences of the 50 most frequently used words in the English comments.

lingual and low-resourced language communities. Scholars interested in language and identity, digital culture, or African sociolinguistics may benefit from the insights offered here.

This work opens up new avenues for future research. It would be valuable for upcoming studies to explore similar approaches in other South African languages. Additionally, future research could consider developing tailored stop word lists for isiXhosa YouTube texts, which would further enhance the accuracy and depth of computational linguistic analysis in this low-resourced language.

Limitations

Despite the insights provided by these analyses, it is important to acknowledge the limitations of the methodology applied in this article. First, the findings from the content analysis approach are confined to isolated words, and the interpretations rely heavily on the researcher's attributions and assumptions.While content analysis was used to infer identity markers, it does not allow for definitive claims about identity expression without deeper contextual understanding.

Furthermore, the lack of a tailored stop word list for isiXhosa limits the precision of text processing and may influence the reliability of the finding. Even so, this article has revealed possible strategies that amaXhosa may use to express their identities on YouTube. A more thorough, contextualised analysis of the texts would be necessary to confidently determine which strategies were used to express identity in the comments.

Acknowledgments

I would like to express my sincere gratitude to Dr Johannes Sibeko for reviewing this work and offering constructive feedback. My appreciation also goes to the Digital Humanities Association of Southern Africa for introducing me to Digital Humanities and other ways of doing research, and to The Escalator Project at the South African Centre for Digital Language Resources for funding my training. I am deeply thankful to Dr Sharon Rudman and Dr Johannes Sibeko for their invaluable guidance throughout the broader study that informed this article, my MA at Nelson Mandela University.

References

Saifuddin Ahmed and Jörg Matthes. 2017. Media representation of muslims and islam from 2000 to 2015: A meta-analysis. *International Communication Gazette*, 79(3):219–244.

Munassir Alhamami. 2023. Instructional communication and medium of instruction: Content instructors' perspectives. *SAGE Open*, 13(2):21582440231172713.

Jannis Androutsopoulos. 2015. Networked multilingualism: Some language practices on facebook and their implications. *International Journal of Bilingualism*, 19(2):185–205.

Heike Baldauf, Christine Develotte, and Magali Ollagnier-Beldame. 2017. The effects of social media on the dynamics of identity: Discourse, interaction and digital traces. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 20(1).

Judith Baxter. 2016. Positioning language and identity: Poststructuralist perspectives. In Siân Preece, editor, *The Routledge Handbook of Language and Identity*, pages 34–49. Routledge, London and New York.

Bouchaib Benzehaf. 2023. Multilingualism and its role in identity construction: A study of english students' perceptions. *International Journal of Multilingualism*, 20(3):1145–1163.

Joseph A. Bonito. 2004. Shared cognition and participation in small groups: Similarity of member prototypes. *Communication Research*, 31(6):704–730.

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

Judith Butler. 1999. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York, USA.

Judith Butler. 2007. *Gender Trouble: Feminism and the Subversion of Identity*, 2nd edition. Routledge, New York. Originally published in 1990; this is the second edition widely cited.

Judith Butler. 2025. Performative acts and gender constitution: An essay in phenomenology and feminist theory. In Henry Bial, editor, *The Performance Studies Reader*, pages 186–196. Routledge, New York, USA.

Ron Darvin. 2016. Language and identity in the digital age. In *The Routledge Handbook of Language and Identity*, pages 523–540. Routledge.

Maarten De Laat. 2023. Network and content analysis in an online community discourse. In *Computer Support for Collaborative Learning*, pages 625–626. Routledge.

Naorem Binita Devi. 2009. Understanding the qualitative and quantitative methods in the context of content analysis. In *Proceedings of the International Conference on Qualitative and Quantitative Methods in Libraries (QQML)*, Chania, Crete, Greece.

Niels Gheyle and Thomas Jacobs. 2017. Content analysis: A short overview. Internal research note, Ghent University.

Alan K. Goodboy. 2011. Instructional dissent in the college classroom. *Communication Education*, 60(3):296–313.

Ernest Jakaza. 2022. Identity construction or obfuscation on social media: A case of facebook and whatsapp. *African Identities*, 20(1):3–25.

A. J. Kleinheksel, N. Rockich-Winston, H. Tawfik, and T. R. Wyatt. 2020. Demystifying content analysis. *American Journal of Pharmaceutical Education*, 84(1):127–137.

Julia Knight and Alexis Weedon. 2014. Identity and social media. *Convergence*, 20(3):257–258.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2 edition. SAGE Publications, Thousand Oaks, CA.

Isabelle Léglise and Bettina Migge. 2021. Language and identity construction on the french guiana-suriname border. *International Journal of Multilingualism*, 18(1):90–104.

Philipp Mayring. 2021. *Qualitative Content Analysis: A Step-by-Step Guide*. SAGE Publications Ltd, London.

Liang Meng, Xiao-Fei Zhang, Jia-Min Li, and Zhao-Yu Sun. 2024. Occupational stigma perception and emotional labor: The role of ambivalent occupational identification and leaders' emotional intelligence. *Current Psychology*, 43(19):17225–17238.

Don Miller. 2021. Analysing frequency lists. In Tony McEnery, Robbie Love, and Vaclav Brezina, editors, *A Practical Handbook of Corpus Linguistics*, pages 77–97. Springer, Cham.

Nkazimlo Ngcungca. 2025. Identity expression in language use by amaxhosa on youtube. Master's thesis, Nelson Mandela University.

Nkazimlo N. Ngcungca, Johannes Sibeko, and Sharon Rudman. 2024. A qualitative inquiry into the south african language identifier's performance on youtube comments. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages (RAIL) @ LREC-COLING 2024*, pages 45–54.

Bonny Norton. 2010. Language and identity. In Nancy H. Hornberger and Sandra Lee McKay, editors, *Sociolinguistics and Language Education*, pages 349–369. Multilingual Matters, Bristol and Blue Ridge Summit.

Edel T. O'Hagan, Adrian C. Traeger, Samantha Bunzli, Hayley B. Leake, Siobhan M. Schabrun, Benedict M. Wand, Sean O'Neill, Ian A. Harris, and James H. McAuley. 2021. What do people post on social media relative to low back pain? a content analysis of australian data. *Musculoskeletal Science and Practice*, 54:102402.

Charles E. Osgood and Evelyn G. Walker. 1959. Motivation and language behavior: A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology*, 59(1):58–63.

Alastair Pennycook. 2007. Language, localization, and the real: Hip-hop and the global spread of authenticity. *Journal of Language, Identity, and Education*, 6(2):101–115.

Leonhard Praeg. 2014. *A Report on Ubuntu*. University of KwaZulu-Natal Press, Scottsville, Pietermaritzburg, South Africa.

Doreen Reifegerste and Annemarie Wiedicke. 2023. Content analysis in the research field of health coverage. *Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft–Standardized Content Analysis in Communication Research*, page 179.

John Synnott, Maria Ioannou, Angela Coyne, and Siobhan Hemingway. 2018. A content analysis of online suicide notes: Attempted suicide versus attempt resulting in suicide. *Suicide and Life-Threatening Behavior*, 48(6):767–778.

Maureen Taylor and Michael L. Kent. 2010. Anticipatory socialization in the use of social media in public relations: A content analysis of prsa's public relations tactics. *Public Relations Review*, 36(3):207–214.

Sonwabo Lungile Tshabe and F. M. Shoba, editors. 2006. *The Greater Dictionary of isiXhosa: Volume 1, A–J (The official isiXhosa–English–Afrikaans trilingual dictionary of the Republic of South Africa)*. IsiXhosa National Lexicography Unit, University of Fort Hare, Alice, South Africa.

Marilyn Domas White and Emily E. Marsh. 2006. Content analysis: A flexible methodology. *Library Trends*, 55(1):22–45.

Appendices

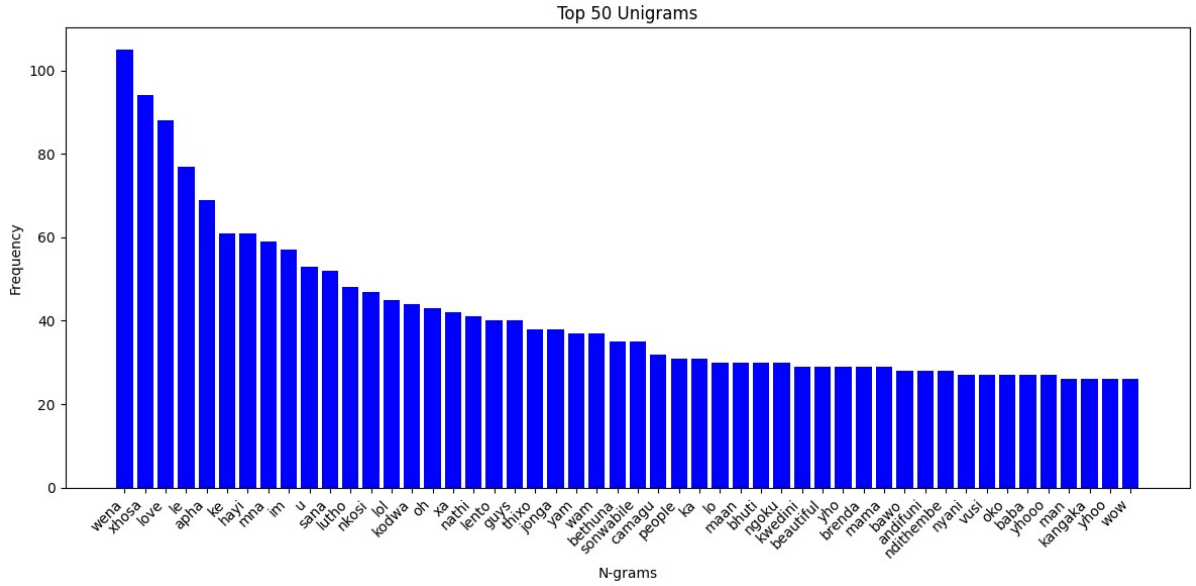

Figure 5: 50 Frequently used words in the isiXhosa comments data set of the selected videos.

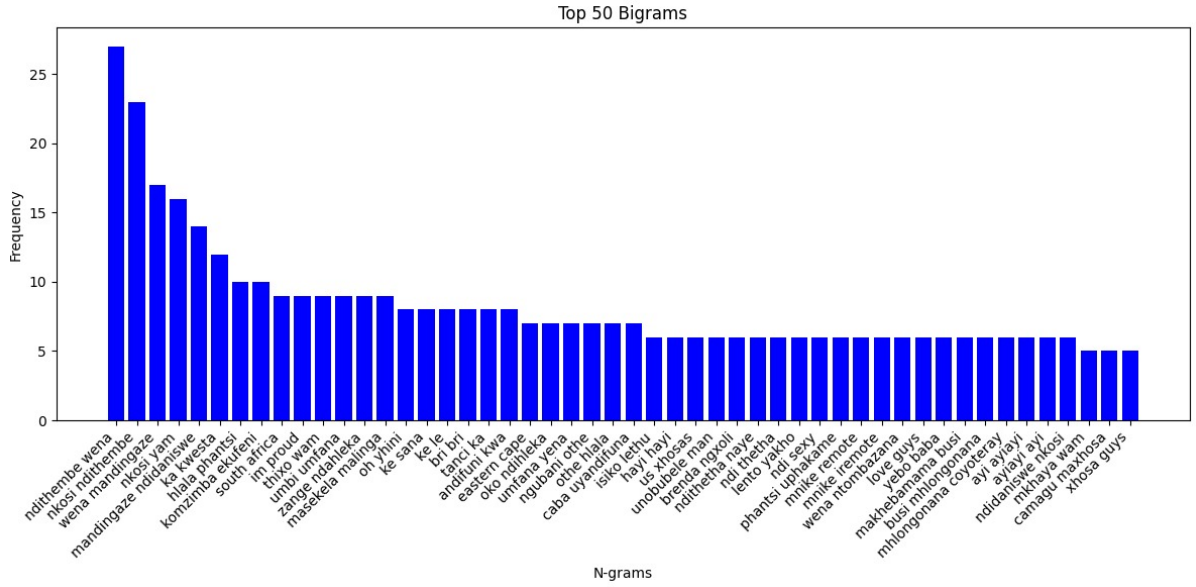

Figure 6: Most frequent bigrams in the isiXhosa comments.

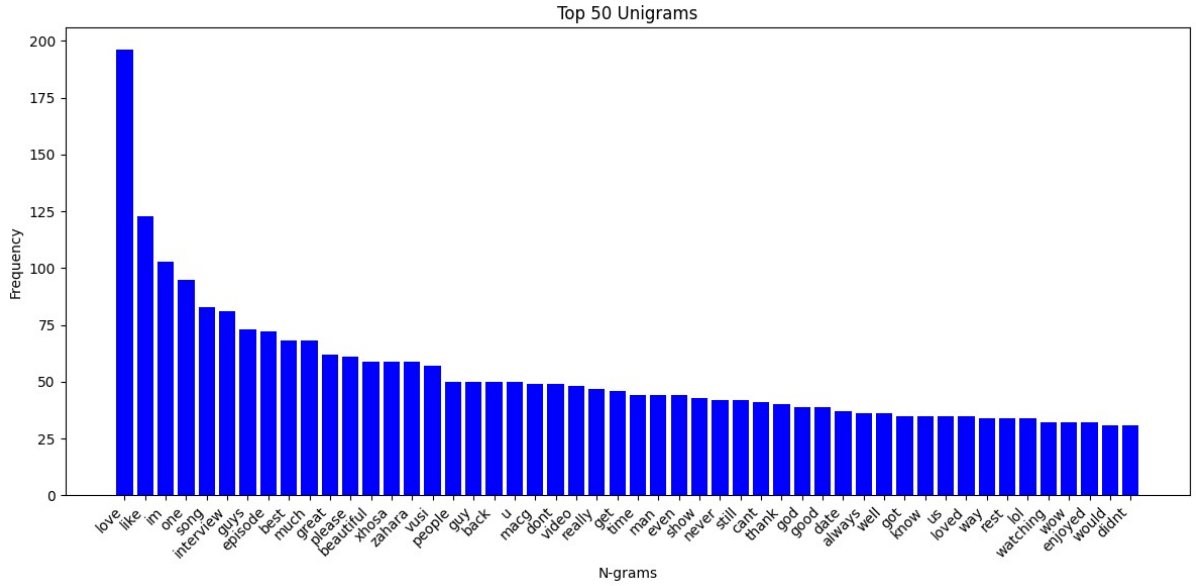

Figure 7: 50 frequently used words in the English comments.

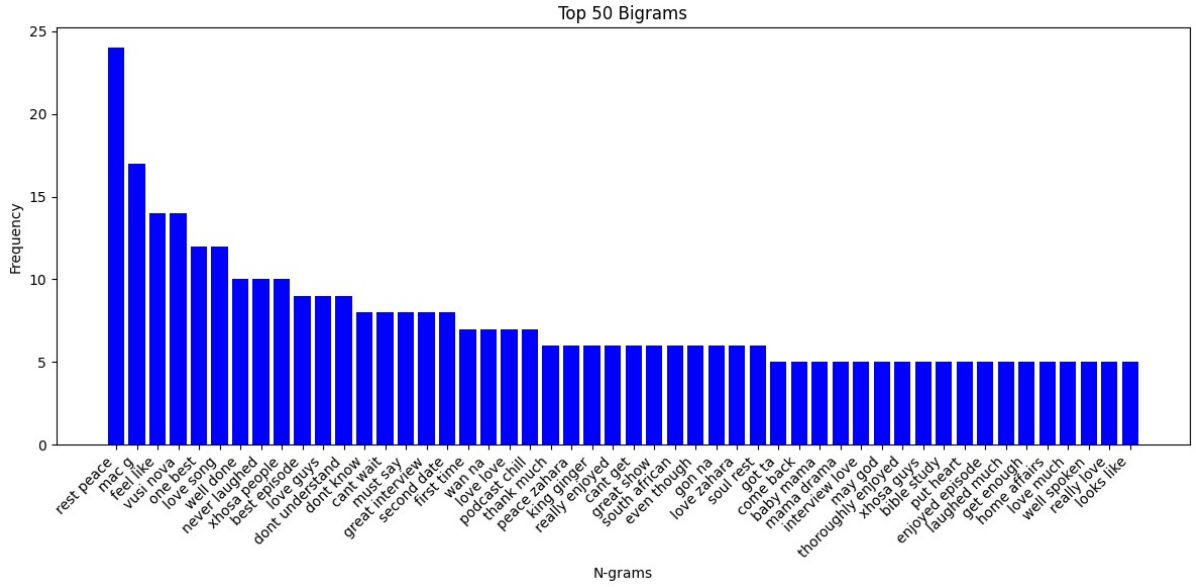

Figure 8: The most frequent collocations in the English comments.