Multilingual vibes: Visualising linguistic resources and emoji in Southern African online discourse

Zelalem Shibeshi
Rhodes University

Marion Walton
University of Cape Town

Johannes Sibeko
Nelson Mandela University

Vusi Maxwell Mthembu
University of Eswatini

Henri-Count Evans
University of Eswatini

Nkazimlo Ngcungca
University of the Western Cape

Alette Schoon
Rhodes University

Abstract

This article presents Vibes, a prototype interface for visualising multilingual online discourse in Southern Africa. We developed the prototype during a three-day hackathon with a multidisciplinary team. The interface combines computational tools, manual coding and visualisation methods to work with data that standard NLP tools cannot process due to their monolingual design. We tested Vibes on two YouTube datasets: English/isiXhosa comments from the @cmtvsa channel and comments on videos discussing a hair product advertisement controversy. Through this work, we encountered practical challenges, including language identification failures, code-switching within single posts, non-standard orthographies, and multimodal communication through emojis. The challenges led us to propose an interface for collaborative coding that accounts for translanguaging practices. The hackathon development process highlighted the need for context-sensitive tools to study linguistic diversity in the Global South.

1 Introduction

Vibes is an early prototype interface to visualize informal discourse in multilingual Southern African online discourse, where emojis and code-switching (or more accurately, translanguaging) are essential to linguistic interaction. The prototype was designed by a multidisciplinary, multilingual[1] team of researchers from Southern Africa over a three-day hackathon, where visualisations and a prototype interface were developed with the goal of evaluating common Natural Language Processing (NLP) tools, and using visualisation to foreground and understand multilingual practices in these contexts.

Visualisations were developed using free NLP libraries and manual coding where necessary.

We introduce scholars to the theories of multilingualism needed to understand the challenge of NLP in the context of linguistic and multimodal data in Southern African online discourses.

1.1 Background

1.1.1 Multilingualism in Southern Africa

Africa may be home to over 2000 languages (Gerhardt, 2020, p. 126). In pre-colonial Southern African societies, languages came into contact and mutually influenced one another (Eriksen and Oppelt, 2025, p.1767). Other forms of interconnection including contact languages (Sommer, 2020, p.890) came about through the region's history of war, slavery, colonialism, apartheid and migration. Under colonialism and apartheid, local African languages were officially marginalised (Alexander, 1989) with English retaining dominance in the current era (Mesthrie, 2006). Nonetheless, many individuals in this region are multilingual, using Bantu languages as well as Afrikaans, English, and others, moving fluidly between them for communication, creative expression, and identity, stylistic or language play (Williams, 2016; Dowling et al., 2019; Hurst, 2020).

Such patterns of multilingual and fluid language use illustrate why linguists have challenged the separate labelling of distinct languages, arguing that language should be understood as a continuum rather than as discrete boundaries (Makoni, 2003). Translanguaging is one concept that captures this perspective, referring to "the deployment of a speaker's full linguistic repertoire without regard for watchful adherence to the socially and politically defined boundaries of named (and usually national and state) languages" (Otheguy et al., 2015).

Over the past few years, this conceptualisation

---

[1] Authors 1,2,3 and 4 of this paper comprised the hackathon team. Linguistic resources of the team included isiZulu, Sesotho, isiXhosa, Setswana, Siswati, Afrikaans, and English, while disciplinary backgrounds included Computer Science, Media Studies and Linguistics

has been extended to acknowledge that such repertoires are not limited to linguistic features alone but also encompass other semiotic and sensory modes of communication. Thus, translanguaging extends beyond linguistic expression to include the coordinated use of images, sound, gesture, visual cues and material objects in the production of meaning (Canagarajah, 2013; García and Otheguy, 2019). In this article, we further extend this perspective to include digital semiotic resources, such as emoji, which function alongside written and spoken language and sometimes act alone to express emotion.

Researchers have highlighted the importance of multilingualism and translanguaging in understanding online interaction among South African audiences. For instance, Smit and Bosch (2020) observed how Black middle-class viewers of South African television programmes used Twitter (now X) to live-tweet during broadcasts, interacting with one another and with producers of the shows. Such practices demonstrate how multilingualism is deeply embedded in everyday communication, particularly in online environments where people draw from their full linguistic repertoires to participate in public discourse.

Evans and Chetty (2023) explored a multilingual corpus of tweets about power cuts in South Africa, showing subtle differences in emoji use when tweets were categorised according to language (English, Afrikaans, and isiXhosa).

Understanding the dynamics of these multilingual and translanguaging practices is a crucial aspect of decolonising digital methods for communication research in Southern Africa (Bosch, 2022). However, many of the tools and workflows currently available for studying online discourse are built on monolingual assumptions. These assumptions make it difficult to analyse the multilingual practices of online publics in Southern Africa and, indeed, in many other parts of the Global South.

Monolingual biases, understood as the viewpoint that people who speak only one language (monolinguals) are considered the norm, while bilinguals and multilinguals are treated as exceptions, have historically framed multilingualism as problematic (Barratt, 2018; Liyanage and Canagarajah, 2023), preferring to model ideas of 'language' on monolingual written text (Stroud, 2020). These biases stem from historical processes that have privileged certain linguistic practices and standard language varieties as normative (Flores and Rosa, 2015; Makoni and Pennycook, 2005; Mignolo, 2012; Mil-

roy, 2001). Consequently, computational tools for text analysis, largely developed in Northern contexts, inherit these monolingual orientations, while available dictionaries reflect written corpora from elite genres such as news media and standard monolingual variants.

These challenges underscore the urgency of developing locally informed computational tools and resources. Developing NLP technologies such as corpora, lexical analysers, and language detectors for local languages remains a challenge due to the lack of adequate resources for most South African indigenous languages (Obrocka et al., 2019). As a result, existing machine translation systems and related applications continue to perform poorly (Mlambo et al., 2025).

Researchers working with linguistic analysis and computational modelling of these low-resource languages face many challenges since online discourse includes features that are common in translanguaging (code-switching, contractions, non-standard orthography and idiosyncratic or non-standard syntax), and these features reduce the success of existing NLP techniques and require language identification at the token level (Obrocka et al., 2019).

This article contributes to this discourse by presenting Vibes, a prototype interface for analysing and visualising multilingual and code-switched social media data. This project was developed by multidisciplinary experts through collaboration during a hackathon. It brings together researchers from computational linguistics, digital humanities, and social sciences to create tools which explore multilingual online communication in Southern Africa.

1.2 Article Aims

This article aims to address the challenges posed by graphocentric data that standard NLP tools struggle to process due to their monolingual design. In doing so, it introduces Vibes, a prototype interface developed for analysing and visualising nongraphocentric, multilingual and code-switched social media data.

2 Related Work

Ngcungca and Sibeko (2024) report using the Python VADER lexicon and rule-based sentiment analysis tool (Hutto and Gilbert, 2014), with some success on a multilingual corpus. Unfortunately, they do not provide evidence for sentiment-mapping non-English words, which would be in-

teresting for this article. They highlight that a lack of contextual awareness, which is a common shortcoming of such tools, may be exacerbated in multilingual texts, as evidenced by cases of missed sarcasm in their case.

Traditionally, NLP pre-processing 'cleans' emojis from the dataset but these are often crucial to the contextual and affective meaning of text. Emoji, like other multimodal and paralinguistic communicative resources, are frequently interpreted differently across age groups and cultures, while their interpretation and patterns of use may vary (Yurieff, 2021; Docrat and Kaschula, 2024; Evans, 2017), making emojis integral to the study of translanguaging.

Coats (2018) points out that emojis can be utilised to address the question of sentiment in online discourse, although the available tools have not kept up with the expanding repertoire of contemporary emoji characters and sequences.

Thus multilingual text analysis and emoji analysis both need to be addressed in a visualisation tool for researchers interested in translanguaging.

3 Methodology

3.1 Approach and Data

We tested available NLP tools on two different sets of comment data from YouTube videos studied by Ngcungca and Sibeko (2024) and Walton (2024). These datasets can be accessed at (Shibeshi, 2025).

The Ngcungca and Sibeko (2024) dataset is a mixed English/isiXhosa corpus of 95237 comments posted on YouTube videos on the @cmtvsa channel [2]. A second multilingual corpus (Walton, 2024) was compiled after querying YouTube Data Tools (Rieder, 2015) and collecting 3419 comments from a sample of YouTube videos which matched a query about a South African controversy over racism in online adverts. The advertisement promoted Unilever hair products and was published by South African retailer, Clicks. The resulting corpus includes comments on news videos reports and on vlogs by YouTubers covering the controversy.

3.2 Tools

We used regular expressions to detect emojis across multiple Unicode ranges and NLTK for tokenisation and stopword handling for English (there is

---

[2]The YouTube channel can be accessed at `https://www.youtube.com/@cmtvsa`

no standard stopword list for isiXhosa). We experimented with the *langid*, *langdetect* and *polyglot* Python packages for automatic language detection, but they proved unreliable for our data, as discussed below. Google Translate offers strong performance for major languages, but it requires a paid cloud account, so we excluded it to keep our approach open and reproducible.

We used Figma to design the Vibes interface. The New Momentum Analysis Toolkit for R, along with Flourish, 4CAT (Peeters and Hagen, 2022), Matplotlib and WordCloud, were used to visualize emoji–word associations and frequencies. The analysis code and interface mockup are publicly available in a GitHub repository (Shibeshi, 2025).

3.3 Data Analysis

3.3.1 Language Detection Challenges

Natural language tools (*Langdetect*, *Textblob* and *LangId*) (Lui, 202) were not able to identify the languages in the English-isiXhosa dataset. For instance, the LangId tool incorrectly identified isiXhosa as Swahili in some instances.

Polyglot (Al-Rfou et al., 2013) successfully detected isiXhosa in a longer sequence of text from this dataset, but failed to classify individual lines (or shorter sequences of text). As with LangId, some sentences were labelled as isiZulu and others as Kiswahili. In the Clicks dataset, lines were labelled as isiZulu with high confidence levels when they also included English, Afrikaans, and Sesotho, as in the following example:

(1) Anywae ngiyaxolisa siswam vah ... moenie met
    Anyway sorry sister.mine hear ... don't with
    my veg nie ka kopo ungaqumbi uqalekise
    me fight not with please chill don't curse
    'Anyway, I'm sorry my sister, okay... please don't fight with me, don't be angry and curse.'

In other cases, lines from the @cmtvsa dataset were labelled as English, with high confidence levels, even when they included translanguaging, as in the example below:

(2) Ke sfebe dis girl Joh n has no shame
    Is harlot this girl wow and has no shame
    'Wow! This girl is promiscuous and she has no shame.'

The exact details of how and where specific linguistic resources are used are crucial to the study of multilingualism. Unfortunately, confidence levels and single language labels are not analytically precise enough to characterise the deliberate, expressive and political use of more than one language.

As a result, our methodology was challenged in at least three important ways. First, traditional NLP tools are designed for 'clean' monolingual text and code-switching or intra-word mixing breaks, tokenisation, parsing, and tagging. In reality, our data included multiple languages blended within a single post, or even a single word (e.g. "ge-worry"), and thus could not be parsed properly. Language detection tools remained a challenge, highlighting the limitations of existing tools for studying African multilingual social media texts.

Second, local taboo language (swearing, insults and hate speech) were not flagged at all by standard libraries (hatebase.org, 2025).

Third, emojis and emoticons, which are crucial for conveying emotional and cultural nuance, are frequently treated as noise by NLP systems.

In the end, our analysis encountered several key technical challenges:

1. Support for identifying indigenous languages in available libraries was unreliable and limited.

2. The ambiguity of language categories in multilingual societies was further highlighted by overlap between closely related languages, code-switching within short sentences, and non-standard orthographies.

3. Code-switches away from English in informal discourse were not identified.

4. Identifying different UNICODE emojis and ASCII-based emoticons requires careful planning.

5. Emojis directly adjacent to words were counted together as a single token.

6. Tokenising informal text at the sentence level is complicated by inconsistent punctuation and sentence boundaries.

We thus confirmed the continued problem of monolingual bias in standard tools for NLP.

The time constraints of the hackathon encouraged us to use a dictionary of English words (cracklib-english) to identify non-English tokens.

Non-English tokens were then either identified as emoji, named entities, or usernames. Finally, using the multilingual knowledge of the whole team, the remaining stretches of discourse were manually coded with tags differentiating between informal English, other South African languages, and taboo uses of language (swearing or potentially hateful and offensive speech). Figure 1 is one of the visualisations produced from this manual coding exercise. In this chart the artificial linguistic categories nonetheless simplify the finely-grained variety of linguistic resources in online discourse.

The process of coding reinforced our awareness of the politics of maintaining and blurring boundaries between languages in contexts where notions of linguistic "purity" are used to maintain racial hierarchies (Haupt, 2017). Commenters blended and juxtaposed linguistic resources while addressing online audiences with complex multilingual and multimodal repertoires.

3.4 Adaptable Interface for Collaborative Coding

Given the limited availability of existing resources and the range of linguistic expertise needed to engage with these corpora of social media comments, we agreed that it was a priority to design an adaptable collaborative interface (also available in the project repository (Shibeshi, 2025)). This prototype would allow team-based manual tagging of tokens that could not be identified automatically, and could be also be used to review lexically ambiguous or misidentified tokens.

Identifying the linguistic repertoires present in even the small snippets of data (cf. Figure 1) required input from all members of the team.

Automated approaches to identifying taboo words, offensive terms and hate speech were also not successful. Nonetheless, we found multilingual strategies being used to evade platform moderation of racist hate speech and of misogynistic language. These uses of taboo language were tagged manually as well, drawing on cultural knowledge and eliciting discussions about context. Developing localised, multilingual dictionaries of offensive terms and hate speech using this interface might help to identify online practices designed to evade English tools for content moderation (Leppänen and Sultana, 2023).

While incorporating the scale of certain NLP techniques, the proposed interface would also allow closer reading and the contextual interpretation needed for content analysis and discourse analysis (Baker et al., 2008).

Overall, the design should not assume that lexical resources belonged to separate pre-defined categories but should allow researchers to approach

Mixing it up
South Africans translanguaging in YouTube comments
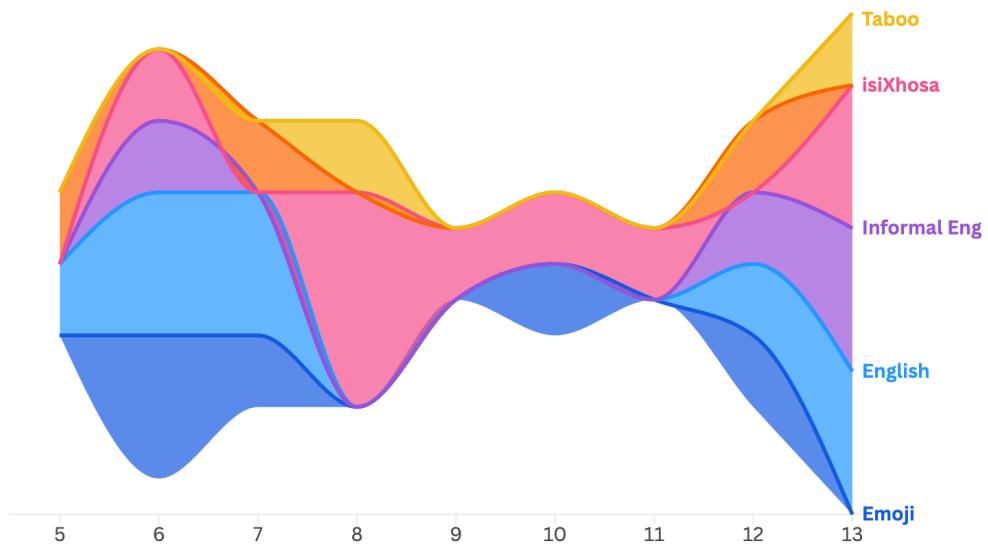
Figure 1: Translanguaging features in 9 lines of manually tagged comments (@cmtvsa dataset)

language identification in flexible ways. We are hopeful that dictionaries created in this way could be used to train more nuanced machine-learning models of multilingual text in future.

3.5 Visualising Emoji

The text was tokenised with Natural Language Toolkit (NLTK) and regular expressions to search for the various items that we wanted to analyse. Note that emojis are found in different ranges of Unicode; as a result, the script needed to use these ranges to identify all these emojis.

Several libraries exist in Python and R to identify and visualise emojis. None were entirely satisfactory as they all provided monomodal visualisations (see Figures 3–2 in Appendix A). Nonetheless they reveal affective differences e.g. conflict and racism in the SABC channel (Figures 4, 5) versus a space of solidarity and support cultivated by a YouTuber discussing the same controversy (Figures 6, 2). We wanted visualisations to show emojis in these affective contexts of translanguaging. Thus, our proposed interface included emoji-cloud visualisations as an interface to explore the significance of emojis not only through frequencies, but also through transmodal collocations, distribution, concordances, and position in co-word and co-commenter networks. Such navigable visualisations should allow for keyword-in-context approaches to sentiment and multimodality.

4   Conclusion

The design and development of the Vibes prototype demonstrates the potential of interdisciplinary collaboration and hackathon methods for studying multilingual online discourse in Southern Africa. The prototype and interface are a first step towards tools for analysing multilingual online communication. Our interactions in the hackathon helped us challenge monolingual bias, document the significance of translanguaging practices, and better understand the dynamics of online publics in the Global South. The interdisciplinary expertise of the team and the hackathon method played an important role in evaluating the NLP tools.

This effort also reflects the practical challenges encountered when analysing multilingual data, such as the inability of standard NLP tools to identify languages in mixed English/isiXhosa text accurately, the difficulty of parsing code-switching and non-standard orthographies, and the need to account for emojis and other multimodal cues that carry meaning in context.

Limitations

This article aimed to address the challenges posed by graphocentric data that standard NLP tools struggle to process due to their monolingual design. While in doing so, we introduced Vibes, a prototype interface developed for analysing and visualising multilingual and code-switched social media data, specifically multimodal paralinguistic forms that incorporate emojis, certain limitations remain.

First, we were unable to overcome parsing challenges, particularly inter-word code-switches, which are typical in the texts used for our experiments. Second, local taboo language, including swearing, insults, and hate speech, was not reliably flagged by standard NLP libraries. This type of analysis is important for the type of data we used in our experiments.

Third, the manual tagging of languages other than English proved time-consuming, highlighting the challenges of extending analyses to multiple languages. Additionally, our analysis was limited in its coverage of Afrikaans data, which may have been better supported by existing tools, restricting the generalisability of our findings across all South African languages.

Acknowledgments

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning Sofia, Bulgaria*, volume 17, pages 183–292.

Neville Alexander. 1989. Language policy and national unity in south africa/azania.

Paul Baker, Costas Gabrielatos, Majid KhosraviNik, Michał Krzyżanowski, Tony McEnery, and Ruth Wodak. 2008. A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the uk press. *Discourse & Society*, 19(3):273–306.

Leslie Barratt. 2018. Monolingual bias. In *The TESOL Encyclopedia of English Language Teaching*, pages 1–7. Wiley-Blackwell.

Tanja Bosch. 2022. Decolonizing digital methods. *Communication Theory*, 32(2):298–302.

Suresh Canagarajah. 2013. *Translingual Practice: Global Englishes and Cosmopolitan Relations*. Routledge, London and New York.

Steven Coats. 2018. Skin tone emoji and sentiment on twitter. *arXiv preprint arXiv:1805.00444*.

Zakeera Docrat and Russell H. Kaschula. 2024. Approaches to interpreting emojis as evidence in south african courts: A forensic linguistic perspective. *South African Journal of African Languages*, 44(2):180–187.

Tessa Dowling, Kay McCormick, and Charlyn Dyers. 2019. Language contact in cape town. In Rajend Mesthrie, editor, *English in Multilingual South Africa: The Linguistics of Contact and Change*, pages 129–150. Cambridge University Press, Cambridge.

Thomas Hylland Eriksen and Riaan Oppelt. 2025. Post-apartheid diversification through afrikaaps: language, power and superdiversity in the western cape. *Ethnic and racial studies*, 48(9):1762–1777.

Henri-Count Evans and Yossabel Chetty. 2023. Mapping emojis across three languages in south africa: Using language as a subtext of culture to understand commonalities and differences in a heterogeneous society. Digimethods Winterschool (19-22 June 2023): Rhodes University. [Seminar presentation].

Vyvyan Evans. 2017. *The Emoji Code: How Smiley Faces, Love Hearts and Thumbs Up Are Changing the Way We Communicate*. Michael O'Mara Books, London.

Nelson Flores and Jonathan Rosa. 2015. Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review*, 85(2):149–171.

Ofelia García and Ricardo Otheguy. 2019. Plurilingualism and translanguaging: Commonalities and divergences. *International Journal of Bilingual Education and Bilingualism*, 23(1):17–35.

Ludwig Gerhardt. 2020. Reflections on the history of african language classification. In Rainer Vossen and Gerrit J. Dimmendaal, editors, *The Oxford Handbook of African Languages (First edition)*, pages 125–136. Oxford University Press, Oxford.

hatebase.org. 2025. Hatebase.

Adam Haupt. 2017. Mix en meng it op: Emile yx?'s alternative race and language politics in south african hip-hop. *M/C Journal*, 20(1).

Ellen Hurst. 2020. Language birth: Youth/town language. In Rainer Vossen and Gerrit J. Dimmendaal, editors, *The Oxford Handbook of African Languages (First edition)*, pages 843–857. Oxford University Press, Oxford.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Sirpa Leppänen and Shaila Sultana. 2023. Digital multilingualism. In Marilyn Martin-Jones and Deirdre Martin, editors, *The Routledge Handbook of Multilingualism*, pages 281–298. Routledge, London.

Indika Liyanage and Suresh Canagarajah. 2023. Lessons for decolonization from pre-colonial translingualism. In Carolyn McKinney, Pinky Makoe, and Virginia Zavala, editors, *The Routledge Handbook of Multilingualism*, 2nd edition, pages 3–16. Routledge, Abingdon, UK.

Baldwin T Lui, M. 202. langid.py: A standalone language identification tool [code].

Sinfree Makoni. 2003. From misinvention to disinvention of language: Multilingualism and the south african constitution. In Sinfree Makoni, Geneva Smitherman, Arnetha F. Ball, and Arthur K. Spears, editors, *Black Linguistics: Language, Society, and Politics in Africa and the Americas*, pages 132–153. Routledge, London and New York.

Sinfree Makoni and Alastair Pennycook. 2005. Disinventing and (re)constituting languages. *Critical Inquiry in Language Studies*, 2(3):137–156.

Raj Mesthrie. 2006. Language, transformation and development: a sociolinguistic appraisal of post-apartheid south african language policy and practice. *Southern African linguistics and applied language studies*, 24(2):151–163.

Walter D Mignolo. 2012. *Local histories–global designs: Coloniality, subaltern knowledges, and border thinking*. Princeton University Press, Princeton, NJ.

James Milroy. 2001. Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, 5(4):530–555.

Respect Mlambo, Valencia K Wagner, and Nomsa J Skosana. 2025. Comparing machine translation and human translation for south african languages. *Stellenbosch Papers in Linguistics Plus*, 69(1):1–16.

Nkazimlo Ngcungca and Johannes Sibeko. 2024. Unmasking deception: An exploratory study of viewers' attitudes towards romantic betrayal. *Journal of the Digital Humanities Association of Southern Africa*, 5(1):1–15.

Monika Obrocka, Charles Copley, Themba Gqaza, and Eli Grant. 2019. Prevalence of code mixing in semi-formal patient communication in low resource languages of south africa. *arXiv preprint arXiv:1911.05636*.

Ricardo Otheguy, Ofelia García, and Wallis Reid. 2015. Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, 6(3):281–307.

Stijn Peeters and Sal Hagen. 2022. The 4cat capture and analysis toolkit: A modular tool for transparent and traceable social media research. *Computational Communication Research*, 4(2):571–589.

Bernhard Rieder. 2015. Youtube data tools. `https://ytdt.digitalmethods.net`. Version 1.42 [Software].

Zelalem Sintayehu Shibeshi. 2025. Emoji and multilingual social media text analysis code. `https://github.com/zelalems/codeswitching/`. GitHub repository.

Alexia Smit and Tanja Bosch. 2020. Television and black twitter in south africa: Our perfect wedding. *Media, Culture & Society*, 42(7-8):1512–1527.

Gabriele Sommer. 2020. Pidgin and creole languages. In Rainer Vossen and Gerrit J. Dimmendaal, editors, *The Oxford Handbook of African Languages (First edition)*, pages 889–898. Oxford University Press, Oxford.

Christopher Stroud. 2020. Multilingualism in south african education: A southern perspective. In Rajend Mesthrie, editor, *English in Multilingual South Africa: The Linguistics of Contact and Change*, pages 311–330. Cambridge University Press, Cambridge.

Marion Walton. 2024. Visual methods and multimodal digital research: Building skills, infrastructure and supportive communities of practice. Centre for Film and Media Studies, Cape Town. [Seminar presentation].

Quentin E Williams. 2016. Ethnicity and extreme locality in south africa's multilingual hip hop ciphas. In *Raciolinguistics*. Oxford University Press, New York.

Kaya Yurieff. 2021. Sorry, millennials. The emoji isn't cool anymore | CNN Business.

5 Appendix A: Visualisations

5.1 Clicks racist ad controversy comments

### 5.1.1 Emoji use in Clicks dataset
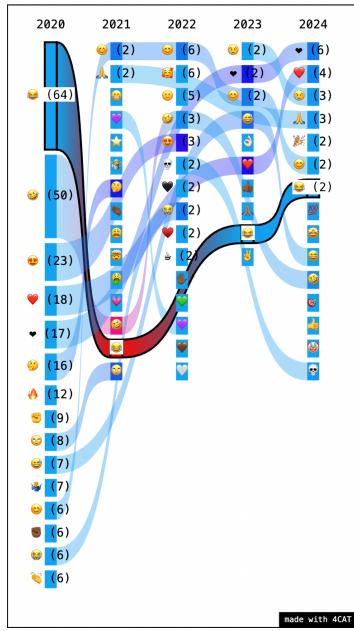
Figure 2: Frequently used emojis in the Clicks dataset, produced with `4cat.uct.ac.za`

5.1.2 SABC channel

Figure 3: Emojis from all SABC videos (Clicks)
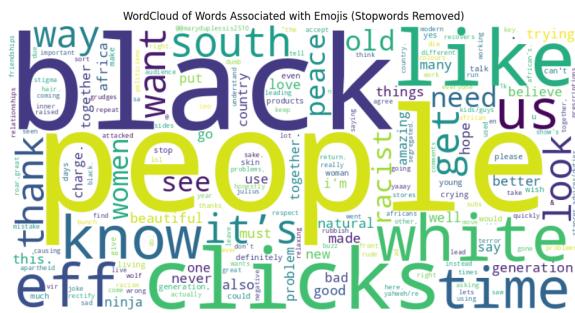
Figure 4: Words most associated with emojis from all SABC videos (Clicks)

5.1.3 South African Youtuber channel

Figure 5: Emojis from SA YouTuber video (Clicks)
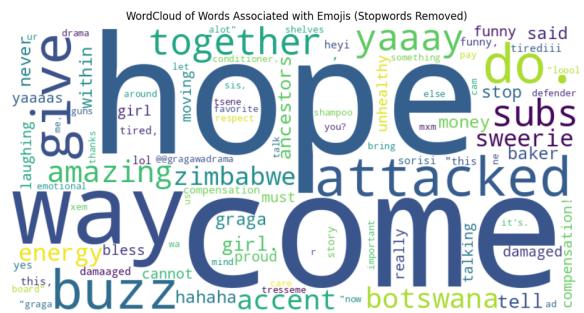
Figure 6: Words most associated with emojis from South African YouTuber channel (Clicks)