Automatic transcription and (de)standardisation

Nina Markl[1], Electra Wallington[1], Ondřej Klejch[1], Thomas Reitmaier[2], Gavin Bailey[2], Jennifer Pearson[2], Matt Jones[2], Simon Robinson[2], Peter Bell[2]

[1]University of Edinburgh, Scotland; [2]Swansea University, Wales

nina.markl@ed.ac.uk

Abstract

In this paper we illustrate the gap between real language use and the language use assumed in ASR development through the example of isiXhosa in Langa, South Africa. Understanding speech and writing practices in context is particularly important when developing speech technologies for minoritised and under-resourced languages, and their communities.

1. Introduction

Advances in automatic speech recognition (ASR) tend to be unevenly distributed between and within languages, with better access to more robust technologies for (monolingual) speakers of prestigious, "high-resource", standard varieties[1] [1, 2]. Minoritised and "under-resourced" varieties, and their speech communities, remain under-served as conventional ASR development pipelines rely on (increasingly large amounts of) specific language resources and particular assumptions about speakers and language use. One such assumption is the existence of one "correct" "gold standard" of how particular words are spoken and transcribed [3]. Systems are furthermore generally developed for users who use one (named) language, rather than drawing on several languages in a single interaction or sentence.

While variation in spoken and written modalities and code-switching between them occurs in all (socio)linguistic contexts [4, 5, 6], it is particularly important in many varieties we might consider "under-resourced" and/or minoritised. Here we focus on a specific case study: the isiXhosa speaking community in Langa, South Africa [7, 8]. We explore users' (i.e., speakers') perspectives regarding how their language should be represented in ASR systems by engaging them in the creation and evaluation of language resources (speech and transcriptions) and ASR systems. Based on these insights, we provide recommendations for speech technology design for minoritised speech communities.

2. Context

2.1. Langa and isiXhosa: Multilingualism

The site of this study is Langa, a township in Cape Town, South Africa. In 2011, about a third of Cape Town residents spoke Afrikaans as a first or main language, while about 30% spoke isiXhosa, and 30% spoke English.[2] While other languages are also spoken in the city, the colonial languages of Afrikaans and English are particularly common in official settings, on public signage and in writing [9, 10]. During apartheid rule, the non-white isiXhosa speakers were only allowed to settle in designated areas like Langa. In Langa, more than 90% of the about 50,000 residents reported isiXhosa, the third official national language, as a first or main language in 2011.[3]

IsiXhosa is closely related to other languages spoken in the region (e.g., isiZulu and isiNdebele) [11]. The variety of isiXhosa spoken in and around Cape Town is characterised by code-mixing with other languages such as English [12]. Many multilingual speakers, especially those living in multilingual communities such as Langa, draw upon all the languages or linguistic resources in their repertoire, often "switching" within the same conversation or even the same sentence [5, 13].

2.2. Langa and isiXhosa: Orthography and "the standard"

Despite the fact that most people in Langa are multilingual, the established, formal written standard of isiXhosa does not account for this variation. Perhaps more critically, the written norm is also strongly associated with its colonial origins.

Orthography can be understood as a "social practice" [4]. Discussions about orthography, just like discussions about other aspects of language [14], are not (just) about language per se but about larger issues of identity, culture and power. Specifically, orthographies can become "iconic representations" of the social groups promoting, using or implementing them, especially in contexts shaped by the enduring legacy of colonialism, such as South Africa [4, 11, 14].

The history of the standardisation of isiXhosa as laid out in [11] highlights this connection. The first dictionaries and grammars of isiXhosa were produced by missionaries. Over the course of the 19th century, they received criticism from isiXhosa writers as part of a wider resistance to the colonial regime, in which the publishing industry was pivotal [12]. In the 20th century, reforms to orthographic systems for African languages developed by European linguists and missionaries were met with vocal resistance by African writers and readers, but embraced and enforced by powerful publishers [11]. After further spelling reforms in the 1950s and 1970s, today's standard orthography remains alien(ating) for many speakers. As [12, p. 250] puts it: "[The written, school-taught isiXhosa standard norm] is perceived as an unchanging artefact which stands in strong opposition to the vibrancy and innovation of the spoken language". As a result, local writing practices diverge from the

[1]We use "variety" as a more neutral term for "dialect" or "language".

[2]https://statssa.gov.za/?page_id=1021&id=city-of-cape-town-municipality

[3]https://statssa.gov.za/?page_id=4286&id=318

formal standard, not just through the inclusion of words and phrases from other languages but also through extensive variation in spelling between different speakers [12].

2.3. An ASR System for Langa

Over the past two years, we have been exploring the potentials and challenges of developing ASR systems for and with the Langa isiXhosa speech community [7, 8]. While some of the most recent state-of-the-art multilingual ASR systems support isiXhosa [15, 16], our approach differs in terms of type and scale of data involved. As discussed in [7] and [8], we have focused on engaging the community in creating a partially transcribed speech dataset and developing an ASR tool to transcribe voice messages – an extremely popular medium of communication in Langa.

In this paper, we reflect on part of this process – in particular a workshop conducted with Langa residents to understand the spelling variation we observed in our community-sourced transcriptions of informal speech. As discussed in [8], we developed a transcription app through which Langa residents could transcribe voice recordings collected through an interface situated in a public space. By engaging the community directly in compiling language resources and evaluating systems throughout development, we were able to take into account local language practices and user needs. We have found that minoritised languages like isiXhosa expose and challenge some of the assumptions underlying ASR development.

In particular, conventional development requires a "gold standard" transcription – and hence an agreed-upon standard variety of the sort which is codified in dictionaries, transmitted through education systems, and often recognised as "prestigious" [17]. Beyond code-switching and variation in speech, we also observe variation in local writing practices, which highlights the need for appropriate language resources, as we discuss below. Where speech technology fails to take account of language variation and the context of the standard variety, it likely performs worse for already marginalised speakers [1]. Beyond this predictive bias, choices of training and test data also have broader impacts on the sociolinguistic context. By training a system using text (and often speech) data in the standard variety and/or formal domains, and subsequently evaluating system performance through comparison with transcripts following conventions of the standard orthography, the status of the standard variety and orthography is reinforced (and other varieties are potentially devalued).

3. The gap between "real language use" and existing language resources

3.1. Available language resources for ASR development

Existing language resources are not representative of the way most isiXhosa speakers in Langa speak and write. The NCHLT isiXhosa speech corpus [18], for instance, consists of 56 hours of read speech (participants were asked to read out short phrasal prompts such as 'omnye ngaphandle kwesizathu'[4]); hence, this corpus contains none of the multilingualism or code-switching we may expect from spontaneous conversational isiXhosa.

Much of the available isiXhosa *text* data is similarly constrained. When [19] compiled the NCHLT isiXhosa text corpus from (all, to their knowledge) isiXhosa textual resources publicly available online, they commented on the lack of breadth

of coverage in terms of domain, style and genre in those resources available. As a result the corpus is mostly drawn from "South African government websites and documents, with some smaller sets of news articles, scientific articles, magazine articles and prose" [19]. This corpus therefore over-represents both the standard variety, and likely also technical and/or legal topics while, crucially, under-representing both informal language use (e.g., code-switching, spelling variation) and topics more commonly discussed among friends in voice messages (e.g., hobbies, conversations about families and friends, etc.). This is reflective of isiXhosa's limited presence on the internet (in 2014 and now), in part due to many multilingual users' preference to use English in this domain and in text-based mobile communication such as SMS, WhatsApp and so on [20].

3.2. Compiling more appropriate language resources

Whilst there are *some* resources more demonstrative of conversational speech – like the Soap Opera corpus [21] – there are too few of these data-sets, and they contain too few hours of speech, for robust model training.

One way to supplement (speech) data in such situations is to turn to the speech/user community and compile a new dataset with their help. We do this in two steps: first, a new speech dataset was collected using a public recording device which prompted local residents to share their experiences of the COVID-19 pandemic in short stories (discussed in [8]). Then, a group of community members were asked to transcribe these stories using a bespoke mobile app [8].

Perhaps because of the personal nature of these stories, this speech data *does* contain clear examples of code-switching and has therefore potential utility for future ASR development in the region. However, we also uncovered challenges, especially with respect to crowd-sourcing transcription from the community. We found extensive variation in the resulting transcripts: see the example in Table 1 for instance. This supports the notion that local writing practices diverge from the formal standard.

On a practical level, this variation is not amenable to immediate integration into ASR development pipelines; recall the requirement for just one "gold-standard" transcription during both training and testing. However, there is no single "gold-standard" if users propose and accept different ways of transcribing the same utterance. This variability, coupled with our prior knowledge about the status of the isiXhosa orthography suggests that users might not *like* a system which uses the formal written norm, both because it represents their language variety poorly and because of its associations with the institutions that shaped it. In [8], we did not probe further to understand these subtle variations and their implications for ASR system design. Here, we set out to do just that by involving local residents exploring appropriate transcription standards for the ASR system.

3.3. Workshop: Understanding transcript variation

Here we present some insights from five participants who were asked to reflect on transcription variants provided by other Langa residents. Importantly, we have no reason to believe that the variation observed in the transcripts would be the result of transcribers misunderstanding or not attending to the task. As the discussions at the workshop confirm, the variation in spelling can also not be simply put down to "spelling mistakes", but rather represents a form of language variation [22]. Participants were asked to comment on which of several transcriptions they deemed "the best" and elaborate on their reasoning for this

---

[4]English Translation: 'another one without reason'.

Table 1: *Examples from [8]'s study, illustrating transcription variability of samples taken from our COVID-19 stories dataset and from the soap-opera dataset [21].*

Transcriber	Transcript
1	kuthiwa abantwana abaninzi bakhula ngaphadle koTata especially boys are more p to the highest behavior...
2	Kuthwa abantwana abaninzi bakhula ngaphandle kotata, especially boys are more prone to high risk behaviour...
3	Kuthw'abantwan'abaninzi abakhula ngaphandle kotata especially boys are more pairing to harsh behavior...
1	then kengoku andakwazi ukuya eskolweni ...nda Quarantiner for ifourteen days ...
2	then kengok andakwazi ukuya eskolweni ...NDA quarantiner for 14 days ...
3	then kengoku andakwazi ukuya esikolweni ...ndakhwaratina for fourteen days ...

evaluation. As expected, the participants did not agree on which transcript variant was "the best" for most examples. However, the discussions did surface some shared concerns around word segmentation, and the treatment of English and non-standard words.

3.3.1. Word segmentation

One notable difference between many transcripts is how words are segmented (see Table 1). The same sound sequences may be chunked into shorter sequences by white space, joined with apostrophes or represented as one uninterrupted character sequence. The workshop participants did not always agree on how words should be segmented, or how important "correct" segmentation would be. One participant explained that segmentation is highly contextual as it depends on speech rate: the same sound sequences should be concatenated when someone is speaking fast, but separated by white space when they are speaking slowly. This type of spelling variation would be quite unusual in (most varieties of) English. The typology of isiXhosa is also relevant here. isiXhosa is an agglutinative language. Relationships between words are indicated through stringing together several units, so-called morphs, which carry specific grammatical meanings (agglutination) and combining several grammatical meanings into one morpheme (fusion). As a result, word segmentation is more flexible as boundaries could be placed between individual morphs.

3.3.2. Non-standard speech

Discussion with participants also surfaced the tension between the desire for transcripts which reflect exactly what has been said and those which follow prescriptive norms. As one participant notes when discussing the difference between <kengoku>[5] and <kengok>: "she's talking slang, she's not saying the full word" but "[some of the transcribers] added a <u> because [that is] how it should be said and written". Here there is a clear notion of how this word "*should be said*". However, this particular participant argues for a faithful transcript, i.e., one in which the phonetic reduction in an informal register is preserved in writing. Other workshop participants disagreed, citing the "missing letter" as the reason why <kengok> is incorrect. This point of disagreement is particularly interesting as it highlights two important choices involved in transcription (manual and automatic). First, there is the question of how to deal with non-standard speech (or "slang"), and secondly, a question of whether the intended meaning or the verbatim speech is more important. This discussion also extends to the

way hesitations, filled pauses and repetitions should be treated in transcripts.

3.3.3. Code-switching

Like in many other linguistic communities, code-switching in isiXhosa is stigmatised [9, 10]. Varieties of isiXhosa spoken in the Eastern Cape, an ancestral homeland of the amaXhosa people, carry a lot of prestige, however [12]. These attitudes were also reflected in our workshop. One participant commented that he believes one of the speakers to be "from the Eastern Cape", adding "so he's speaking clear Xhosa". This speaker, he further says, would likely only use English to borrow words which do not exist in isiXhosa (like "[hand] sanitiser") rather than code-switch like speakers of "township Xhosa".

In code-mixed utterances, English words are often carefully embedded into the utterance according to isiXhosa grammatical rules, for example through affixation of appropriate grammatical markers. One participant explained when discussing the difference between the spelling variants <for fourteen> and <fori ifourteen>: "that's what we do that's our English-Xhosa, so when we're mixing English and Xhosa [...] we add an /i/ just to make it sound as if it was Xhosa when it's not". Not all of this variation was equally accepted by all participants, however. One was unsure about accepting <fori fourteen> since "there's no word such as <fori> in isiXhosa but the person is actually saying /fori/" highlighting again the tension between verbatim transcripts and prescriptive standards. Another participant was certain that <fori fourteen> "[is] completely correct because it's how she's saying it" but qualifies this statement with: "But then I'm Xhosa so I would understand it".

Another point of discussion was the word *quarantine* represented as <quarantiner> or <khwarantina>. One participant explained "the person just used the Xhosa way of writing it – or not the Xhosa way of writing it but how I'd write it if I was writing it in Xhosa without translating it". However, they also argued that "because it's an English word it makes more sense" to spell it according to English norms. Another participant disagreed, arguing instead that only the Xhosa spelling is correct and that the other variants are anglicised. A third participant was more diplomatic, noting that "it depends" and explaining that "Xhosa is more of a sounding language – exactly how it sounds is how you write it. [...] Both of them are correct – I'm not Xhosa though I can write like that but I'd use the <quarantine>."

3.4. Implications for ASR design and evaluation

The insights from this workshop confirm that the variation in spoken and written language use we observe among isiXhosa speakers in Langa must be considered during ASR system de-

---

[5]We adopt the convention of representing writing in <angled brackets> (orthographic transcription) and speech in /slashes/ (phonemic transcription) .

sign and evaluation. The frequent code-switching in informal speech means that an ASR system needs to be able to handle input from at least isiXhosa and English.[6]

Regarding orthographic variation, the workshop feedback also confirms our above intuition: that users might prefer a system which does not strictly reproduce the formal isiXhosa norm. While adherence to the norm is important in some contexts, and participants all made reference to some "spelling errors" in the transcripts, there is also quite a lot of awareness and acceptance of spelling variation. The lack of consensus on how words should be segmented and how to represent English words embedded in Xhosa utterances is particularly interesting from an ASR design perspective.

These clear implications for ASR development highlight the advantages of involving users, or, in the context of language technologies, speech communities, throughout the technology development process. Engagement with user communities is a key area where the technical aspects of speech technology development, and the engineers responsible for it, can benefit from interdisciplinary interaction between human-computer interaction, linguistics and speech technology. This is particularly clear when designing alternative evaluation approaches which centre user perspectives, rather than "objective" metrics commonly used in speech technology development, which break down entirely in the absence of an agreed-upon gold standard and might fail to capture (socially or linguistically) errors.

4. Automatic transcription and (de)standardisation

ASR systems, and other language technologies, interact with existing attitudes and beliefs about language(s). Choices regarding which language varieties to support and how to represent them are never neutral. In the case of isiXhosa, it is clear that the existing written standard is contested. Previous research has shown that many isiXhosa speakers, especially in Cape Town, feel alienated by the written norm. Unlike English and Afrikaans, most Langa residents do not frequently encounter isiXhosa in official writing or on public signs, or through formal education [10, 12]. The "urban" variety spoken in Cape Town is further considered to be very different from the "original" or "traditional", prestigious isiXhosa varieties spoken in the Eastern Cape in particular because of its characteristic code-mixing [12]. Adopting the formal isiXhosa standard to transcribe, in particular, informal speech of Cape Town speakers would not only be inappropriate due to the inability to straightforwardly handle phonetic variation and code-switching but would also be a meaningful intervention in this sociolinguistic context. It would introduce the formal isiXhosa orthography into the new, personal domain of voice messages where it might be particularly unwelcome and alienating. Assuming users would not just reject this because of the use of the written norm, such "domain expansion" could boost the status of the norm and potentially even have a standardising effect, effectively prompting users to accommodate towards the technology.

4.1. Sociolinguistically-informed considerations in ASR development

Based on our experience working on isiXhosa ASR, and working with isiXhosa speakers in Langa, we argue that taking account of language variation, and the way different spoken and written varieties are evaluated by potential users, is crucial for effective language technology design especially for minoritised and/or under-resourced varieties.

Perhaps the most obvious point to start is to establish how most people who might use the ASR system actually use language. In multilingual contexts, this is likely to involve code-switching. The specific application context might also affect users' speech style and degree of accommodation towards the ASR tool. Informal speech, and speech directed at other community members, is likely to be characterised by phonetic variation and a higher density of "dialect" features. Successful ASR systems might therefore require the compilation of new, domain-specific speech datasets.

For systems designed to create transcripts of speech (rather than recognise single commands, for example), understanding if and how local writing practises diverge from the written standard is important. Generally, spelling variation is more likely in computer-mediated, informal writing like text messages and on social media. In contexts where the standard is very contested, not as widely used, or not as widely transmitted to speakers, spelling variation is particularly likely. In this case, users might prefer systems which mirror this variation in their output.

Involving users in the evaluation of ASR systems is particularly important in contexts where "real-world" language use diverges from the kind of language reflected in datasets which would traditionally be used for evaluation. In addition to in-person workshops, online surveys can be useful here. While we only involved a small number of participants in the evaluation process, survey-based evaluation or the use of interactive transcription apps would be scalable to larger sample sizes.

5. Conclusion

In this paper we have shown how understanding local writing practices and language use by multilingual (and multiliterate) language communities of minoritised and under-resourced languages can be incorporated in ASR development. We highlight the importance of returning to users at several points in the development process. Especially in contexts where written standards are highly contested by speakers, relying on standard metrics, rather than user evaluation, can be very misleading. Involving local communities in speech data transcription and evaluation, and drawing on their sociolinguistic expertise and intuitions, ensures that the ASR tool represents their speech in ways that are both readable and appropriate for them.

6. References

[1] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1915768117

[2] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 6282–6293.

[3] S. Bird, "Decolonising speech and language technology," in *Proceedings of the 28th international conference on computational linguistics*. International Committee on Computational Linguistics, 2020, pp. 3504–3519. [Online]. Available: https://www.aclweb.org/anthology/2020.coling-main.313

[4] M. Sebba, *Spelling and Society*. Cambridge University Press, 2007.

---

[6]Ideally a system would be able to accommodate even more local languages, but so far we have limited ourselves to isiXhosa and English.

[5] M. Heller, Ed., *Codeswitching*. De Gruyter Mouton, 1988.

[6] P. Eckert, "Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation," *Annual Review of Anthropology*, pp. 87–100, 2012.

[7] T. Reitmaier, E. Wallington, D. K. Raju, O. Klejch, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Opportunities and challenges of automatic speech recognition systems for low-resource language speakers," in *CHI Conference on Human Factors in Computing Systems*. ACM, 2022.

[8] T. Reitmaier, E. Wallington, O. Klejch, N. Markl, L.-M. Lam-Yee-Mui, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. Association for Computing Machinery, 2023.

[9] A. Deumert, S. Mpazayabo, and M. Thompson, "Cape town as a multilingual city: Policies, experiences and ideologies," in *Routledge Handbook of Translation and the City*, T. K. Lee, Ed. Taylor & Francis Group, 2021, p. 448.

[10] A. M. Dantile, "Language in public spaces: Language choice in two IsiXhosa speaking communities (Langa and Khayelitsha)," Ph.D. dissertation, University of Stellenbosch, Stellenbosch, South Africa, 2015.

[11] A. Deumert and N. Mabandla, "Beyond colonial linguistics," in *Standardizing Minority Languages*, J. Costa, P. Lane, and H. D. Korne, Eds. Routledge, 2017.

[12] A. Deumert, "Imbodela zamakhumsha – reflections on standardization and destandardization," *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, vol. 29, no. 3-4, pp. 243–264, 2010.

[13] R. Otheguy, O. García, and W. Reid, "Clarifying translanguaging and deconstructing named languages: A perspective from linguistics," *Applied Linguistics Review*, vol. 6, no. 3, pp. 281–307, 2015.

[14] J. T. Irvine and S. Gal, "Language ideology and linguistic differentiation," in *Regimes of language: Ideologies, polities, and identities*, P. V. Kroskrity, Ed. School of American Research Press, 2000, pp. 35–84.

[15] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Strohman, B. Ramabhadran, T. Sainath, P. Moreno, C.-C. Chiu, J. Schalkwyk, F. Beaufays, and Y. Wu, "Google usm: Scaling automatic speech recognition beyond 100 languages."

[16] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *arXiv*, 2023.

[17] J. Milroy, "Language ideologies and the consequences of standardization," *Journal of Sociolinguistics*, vol. 5, no. 4, pp. 530–555, 2001.

[18] E. Barnard, M. H. Davel, C. van Heerden, F. De Wet, and J. Badenhorst, "The nchlt speech corpus of the south african languages," in *Workshop Spoken Language Technologies for Under-resourced Languages (SLTU)*, 2014.

[19] R. Eiselen and M. Puttkammer, "Developing text resources for ten South African languages," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 2014, pp. 3698–3703. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf

[20] A. Deumert and S. O. Masinyana, "Mobile language choices — the use of english and isiXhosa in text messages (SMS)," *English World-Wide*, vol. 29, no. 2, pp. 117–147, 2008.

[21] E. van der Westhuizen and T. Niesler, "A first South African corpus of multilingual code-switched soap opera speech," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[22] M. Bucholtz, "Variation in transcription," *Discourse Studies*, vol. 9, no. 6, pp. 784–808, 2007. [Online]. Available: https://doi.org/10.1177/1461445607082580