A Minimal Computing Approach to Southern African Language Resources

Charumbira, Ruramisai
Department of History, The University of Western Ontario
rcharumb@uwo.ca

Turkel, William J.
Department of History, The University of Western Ontario
wturkel@uwo.ca

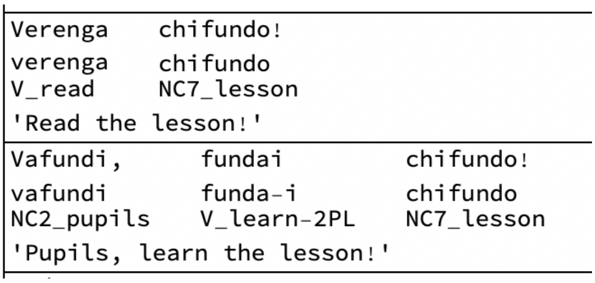
Abstract

This new collaboration between a historian of Southern Africa (RC) and a specialist in computational methods (WJT), is designed to draw on our respective backgrounds and provide opportunities to enlist students and other collaborators in research and teaching. Our goal is to create tools that can be used to help explain unfamiliar languaging in historical contexts. We follow the tenets of minimal computing (Risam & Gil 2022) and take the perspective of language as a complex adaptive system (Kretzschmar 2015). We also situate our work within the postcolonial digital humanities generally (Risam 2018) and the specific critique of knowledge production and racism that Fields & Fields (2012, pp. 5-6) identified as 'racecraft', which "highlights the ability of pre- or non-scientific modes of thought to hijack the minds of the scientifically literate". As practitioners of academic language research and computing, we must be attentive to the history of colonizers trying to not only kill 'native languages' but their speakers and cultures (Ngũgĩ wa Thiong'o 2009). To date, we have partially implemented one prototype for automating interlinear morphemic glossing of chiShona and English as shown in Figure 1 (Charumbira et al 2023). Here our intent is speculative design: to imagine a more inclusive space of computational tools and practices that jettisons some of the assumptions that have shaped the digital cultural record in the Global North.

Keywords: language as complex adaptive system, minimal computing, racecraft, speculative design

1 The Four Questions of Minimal Computing

Minimal computing asks us to consider four questions when developing projects in the digital humanities: "1) 'what do we need?'; 2) 'what do we have?'; 3) 'what must we prioritize?'; and 4) 'what are we willing to give up?'" (Risam & Gil 2022). The need that we address is communicating about languaging in historical contexts for audiences that are partially or completely unfamiliar with vocabulary, dialects, or languages that were in use. What we have are the basic resources of literacy and descriptive linguistics—especially orthographic and phonetic representations and descriptive categories like person, tense, and parts of speech—organized into texts, inventories and word lists, and interlinear morphemic glosses. These are the tools that linguists use when they communicate with one another about unfamiliar languages, regardless of whether they share a theoretical orientation.



```
Verenga    chifundo!
verenga    chifundo
V_read     NC7_lesson
'Read the lesson!'

Vafundi,      fundai         chifundo!
vafundi       funda-i        chifundo
NC2_pupils    V_learn-2PL    NC7_lesson
'Pupils, learn the lesson!'
```

Figure 1: Computer-generated morphological parse displayed as interlinear morphemic gloss (chiShona examples from Fortune 1967).

We wish to prioritize decolonial, antiracist, and inclusive techniques. Citing Syed Mustafa Ali, Roopika Risam argues that decolonial computing in the digital humanities should work from the margins, linking situated and embodied knowledges across the Global South, and reimagining a digital cultural record which has been shaped by systemic racism and white supremacy (Risam 2018). Engaging with raciolinguistics allows us to critique colonial logics of race and language in contexts where they were historically used to oppress (Rudwick & Makoni 2021). As Christopher M. Hutton (1998, p. 3) wrote, "Linguistics is both the parent and the child of race theory". Physical anthropologists in the 19th century sought anatomical features to characterize speakers of various language families. In the wake of degeneracy theory, phrenology, and eugenics, "linguistics has reclaimed its role as the premier science in the classification of human diversity, elaborating a 'characterology' or 'typology' of the world's languages, and therefore of the world's ethnic groups" (cf Stoler 2016). Despite contemporary consensus in the social sciences and humanities that there is no biological basis for race, these reified ideologies continue to serve social agendas, which is where the concept of racecraft can play an important role (Stoler 2016, Sabino 2018). Through imagination and action, racial ideas are continually re-enacted and remade. "[T]he very "relevance" of racial distinctions, what makes them speakable, common sense, comfortably incorporated, and ready to be heard, may derive from the dense set of prior representations and practices on which they build and that they in turn recast," Ann Laura Stoler writes in *Duress* (2016, p. 249). In the American context that Karen and Barbara Fields (2012, p. 24) address, the media reports daily on medical, social, and cultural phenomena systematized 'by race' "constantly churning out factitious evidence for an ever-expanding American immensity, the so-called racial divide".

Minimal computing asks us, finally, what we are willing to give up. By embracing constraint, we resist the easy identification of the digital with the newest, fastest, largest, or most expensive. In our case, that means that one of the first things to abandon is the highly complex engineering approach that characterizes computational linguistics for English. This is partly a matter of necessity, since tools for analysing many Southern African languages simply do not exist (Moors et al 2018, Agić & Vulić 2019) and there is question, for example, about the suitability of traditional word models for Southern Bantu languages which are agglutinative (Kambarami et al 2021). But more than that, the vast number of edge cases and exceptions that form the long tail of natural language are exceedingly costly to account for at scale, and models that attempt to do so inevitably become overcomplicated (Arbesman 2016). We also recognize that the history of computing in Southern Africa is attached to the coloniality of European languages in the region, and Afrikaans, no less.

2 Languaging

Complex engineering models also assume a bounded or static ideal of individual languages which is problematic, especially in the Southern African context where English is sometimes conceived as a more neutral choice than alternatives (Rudwick & Makoni 2021). Comparing the difficulty of identifying markers of linguistic identity with the absence of biological markers for race, Robin Sabino (2018, p. 4) writes "Like all ideologies, the belief in the existence of grammatical systems that are widely shared, uniform, clearly delimited, and autonomous crucially depends on (re)enactment made possible by comfort with familiar incongruities". In creating digital tools to explicate languaging in historical settings, we can deal with beliefs about language (as we deal with beliefs about race) without necessarily committing to the real existence of such entities. We are also still attentive to individual languages as a way of prioritizing reparative work in settings where colonial and post-colonial experiments have rendered some languages better endowed with linguistic resources than others.

Sabino (2018, p. 4) concludes that "in displacing attention from languaging to languages, linguists engage in circular logic, assuming that which we are attempting to establish". Whether we are willing to abandon the utility of named individual languages altogether remains to be seen, but we recognize the issues raised by the proliferation in sociolinguistics of prefixes like *hetero-, metro-, multi-, poly-, pluri-, and trans-* to describe languaging in contemporary and historical settings of mobility and globalization (Lanza in de Bot 2015, p. 78). Social networks, whether computer-mediated or not, provide a rich source of resources (varieties, registers, styles) for individuals and communities to draw upon, fluidly crossing boundaries between named languages (Lee & Wei 2020, Tagg 2020). Rudwick and Makoni (2021, p. 261) write, "Increasingly also, competency in what is thought of as 'standard English' might no longer provide the comprehensibility needed to participate in African metropolitan English lingua franca communication. Rather it might be a complex and skillful polylingualism, mixing, switching and translanguaging strategies which are needed to successfully communicate in African urban spaces".

Writing about the postcolonial context, Rey Chow (2014, pp. 14-15) notes "the colonized is arguably more closely in touch with the reality of languaging as a type of prostheticization" and that this becomes an advantage in undoing and remaking language. She emphasizes "consideration of such illegible and often unconscious elements of languaging as accent, tone, texture, habit, and historicality as well as what is partially remembered, what is erroneous but frequently reiterated, and, ultimately, what remains unsaid and unsayable—all of which bear on transactions of the most basic meanings but tend to elude more positivistic or even scholarly ways of handling translating. (As we know, proper scholarly tools such as etymologies, dictionaries, thesauruses, encyclopedias, archives, databases, and the like are always necessary but never sufficiently helpful)" (Chow 2014, pp. 65-66). If we imagine building inclusive digital forms of the traditional scholarly tools that Chow mentions, and our goal is to facilitate links between situated and embodied knowledges across the Global South, then such tools must present both the thin description or etic perspective, and the thick description or emic perspective. Crucially, it must also be possible to gesture to everything that remains unrepresented in any explicit representation.

These tools will be necessarily shaped by this perspective of prostheticization. First, they must focus on the speaking individual's repertoire—a crucial concept in translanguaging—rather than on the linguistic resources of named languages. Second, rather than trying to establish general patterns that can be used for prediction, the emphasis will be on the moment-by-moment unfolding of languaging in a specific context "to appreciate the epiphanies of creativity and criticality in multilingual settings" (Lee & Wei 2020, p. 408). Shifting attention to the momentary, ephemeral, and contingent aspects of languaging has a historiographic parallel in microhistory (e.g., Ginzburg 1986) with its focus on the clue, the telltale detail, and its metonymic relationship to the surround. In languaging, such 'clues' often become charged with meaning when they creatively disturb or transgress normative linguistic structures (Lee & Wei 2020). Tong King Lee (2023, p. 14) adds three further design features for digital tools that follow from creative multilingualism: such tools should be multimodal and multisensory, they should capture an entire range of language-based performances, and they should create "holistic, transformative social spaces".

We see collaboration on tool design starting with first principles as an important early step in instrumentalizing digital humanities to effect decolonization. As a concrete example, take the automatic creation of a morphological parse and interlinear morphemic gloss as shown in Figure 1. While implementing this tool, we drew on standard techniques in natural language processing such as using a lexicon to store morphemes and transforming representations with context-sensitive rewriting rules. We also accepted the conventions of interlinear glossing: that there are two distinct named languages (one is the 'source' and the other the 'target'), that alignment is used to indicate how a morpheme in one language is translated into the other. Creation of this minimal prototype allowed us to think through some of the ways that it was inadequate for conveying the richness of situated languaging. One example that Charumbira uses in her Southern African history classes is the Zulu greeting *Sawubono* (Northern Ndebele *Sakubona*). A simple translation to English is 'Hello' (or more literally 'I see you') but in the history of the language, it has the deeper meaning of expressing to someone that they are seen and or beheld by a community rather than only one individual greeting another individual. And as an aside, an individualized greeting might sound like *Ngiyakubona*, whose meaning would change to one of seeing someone (at a distance or someone hiding) rather than the meaning of beholding someone and all their relations. To capture this kind of detail we need to engage with the assumptions of descriptive linguistics at the foundational level, with Saussure's distinction between syntagmatic and paradigmatic relations, for example. We also need to explore the affordances that digital systems allow for dynamically representing more complex relations than simple alignment.

The computational aspects of the digital tools that we envision follow from our perspective that language is a complex adaptive system (CAS), and that order emerges through learning and adaptation (Kretzschmar 2015). There are many definitions of CAS, but most agree on there being a large network of interacting components "that exhibits nontrivial emergent and self-organizing behaviors" with no central controller (Mitchell 2009, p. 13). We imagine embedding our digital versions of traditional linguistic tools (like word lists or interlinear morphemic glosses) in bottom-up, stochastic models. These models will be inspired by classic work in CAS simulation by people like John H. Holland (1975, 1995), Stuart Kauffman (1993), and Joshua M. Epstein and Robert Axtell (1996), but our focus on the historical specificity of instances of languaging shifts attention from simulation using simple rule-based interactions among identical agents, to a space that permits exploration and partial annotation of interactions among agents with heterogenous repertoires. Neither the historical agents nor their repertoires can be annotated with any presumption of certainty or completeness. Kretzschmar (2015, pp. 2-3) argues that our own intuitions about our own language do not give us a firm foundation for understanding the alternatives that other speakers can draw upon. "The most basic assumption of generative and structural linguistics, that we speakers all share the system of a language, share the rules for a language, is simply wrong". In this view, whatever generalizations we make about language are post-hoc rather than generative or structural.

3 Reflexivity in Lieu of a Formal Conclusion

The complex adaptive systems perspective that we embrace is also important because it allows the historian to ask what, how, and why minimal computing can do more than the literacy and schooling of yesteryear. That historical system produced the haves and the have-nots among the colonized Africans. In those historical times, though not always the case, it was often that where there were more colonial mission or government schools, the result was a better endowed region whose people often (though not always) became winners on the losing side of colonization. That system privileged some ethnic groups over others in ways that produced a privileged language with ethnographic and linguistic resources that continue to be valuable research resources in a discipline (like History that is) still attached to the written document as the first among equals in terms of proving the rigor of one's scholarship. Through this project, we see the potential of minimal computing to afford us—and our future collaborators—an opportunity for transformative dialog and historical practice. This way, we can both keep up with the changing technology while reflecting on the legacy of a linguistics historiography that focused on "tribes" and their "languages" rather than the languaging that was (and still is) the people's everyday way of being. How do we have conversations with "everyday people" in a way that is neither patronizing nor dismissive of the reality that for some, a focus on languaging without languages may well be understood as a form of coloniality that is dismissive of what they hold dear or experienced. Just as important is asking how the historian of Southern Africa can show the power of minimal computing for the ordinary person hypnotized by the machine that few know how to program or understand its technical language (and languaging) practices. Indeed, "without deep understanding of the politics of language use, translanguaging can devolve into cognitive reification (Orr, 1997) or prejudicial mimicry (Hill, 2007)" (Staats & Halpert 2002, p. 27).

The foregoing point is particularly important for our (the authors') project whose optics fall into the historical stereotypes of the tech whiz white man and the storytelling black woman. Just as important is the historical reality of the native informant and the technical westerner; and southern Africa is replete with such examples many of which now feed into interdisciplinary historical research in the face of the long cold shadow of white supremacy and colonization. The young people crying #RhodesMustFall—and all the complexity that went with those cries—tells us that the emperor still has no clothes, but the historian averts the eye, and therefore the necessary witnessing. The potential of our project also lies in our willingness to hold the both-and, rather than the either/or of doing historical scholarship. We want to hold these contradictions together as we search for generative possibilities that people can both welcome into their lives and use to transform those lives wherever they are lived. The contemporary ubiquity of the cellphone comes to mind: once upon a time, only a privileged few had access to computer machines that took up whole rooms, then a desk, then a lap, and now a palm. In this long and short development, Africans have been (late or the last) consumers, but less the producers or programmers of those machines, even as some of the key components of making those machines were mined in Africa with frightening environmental ruination, and untold human and more-than human suffering. To that end, how can a minimal computing project like ours be both technical and tell human and Earth stories of retrieving our collective humanity through language and languaging practices? To read some of the literature is to realize that if the varied waves of decolonial scholarship and practice are to mean more than an academic discourse, then those interested in inclusive and transformative practices must engage "the people" in ways that the value of what is gained is shared. If the decolonial practices of African nationalisms brought forth political liberation in Southern Africa, how can a project like ours reckon with what has turned out to be a wretched postcolonial experience for the many, especially the poor, as political leaders have practiced their own forms of coloniality using their countries as cash machines, and African languages as weapons in the name of tradition to shut everyone up. To reclaim that voice and space, we return to the initial four questions to lead us forward: "1) 'what do we need?'; 2) 'what do we have?'; 3) 'what must we prioritize?'; and 4) 'what are we willing to give up?'" (Risam & Gil 2022). This is a dialog worth having, a practice worth doing, and an intellectual project worth pursuing.

References

Agić, Ž & Vulić, I 2019, 'JW300: A wide-coverage parallel corpus for low-resource languages', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204-3210.

Arbesman, S 2016, *Overcomplicated: technology at the limits of comprehension*, Current.

Charumbira, R, Edwards, J & Turkel, WJ 2023, 'A minimal computing approach to building computational language resources for Southern African history', Global Digital Humanities Symposium, Michigan State University.

Chow, R 2014, *Not like a native speaker: on languaging as a postcolonial experience*, Columbia, 2014.

Epstein, JM & Axtell, R 1996, *Growing artificial societies: social science from the bottom up*, Brookings Institution.

Fields, KE & Fields, BJ 2012, *Racecraft: the soul of inequality in american life*, Verso.

Fortune, G 1967, *Elements of Shona (Zezuru dialect),* Longman Zimbabwe.

Ginzburg, C 1986, 'Clues: roots of an evidential paradigm', in *Clues, Myths, and the Historical Method*, Johns Hopkins.

Holland, JH 1975, *Adaptation in natural and artificial systems*, MIT.

Holland, JH 1995, *Hidden order: how adaptation builds complexity*, Basic.

Hutton, CM 1998, *Linguistics and the third reich*, Routledge.

Kambarami, F, McLachlan, S, Bozic, B, Dube, K & Chimhundu, H 2021, 'Computational modeling of agglutinative languages: the challenge for Southern Bantu languages', *Arusha Working Papers in African Linguistics,* vol. 3, no. 1, pp. 52-81.

Kauffman, SA 1993, *The origins of order: self-organization and selection in evolution*, Oxford.

Kretzschmar, WA Jr. 2015, *Language and complex systems*, Cambridge.

Lanza, E 2015, Quoted in de Bot, K *A History of Applied Linguistics*, Routledge.

Lee, TK 2023, *Kongish*, Cambridge Elements.

Lee, TK & Wei, L 2020, 'Translanguaging and momentarity in social interaction,' in De Fina, A & Georgakopoulou, A eds., *The Cambridge Handbook of Discourse Studies*, Cambridge.

Mitchell, M 2009, *Complexity: a guided tour*, Oxford.

Moors, C, Wilken, I, Calteaux, K & Gumede, T 2018, 'Human language technology audit 2018: analysing the development trends in resource availability in all South African languages', *Proceedings of 2018 Annual Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT 2018)*, pp. 296-304.

Ngũgĩ wa Thiong'o, 2009, *Something torn and new: an African renaissance*, Basic Books.

Risam, R 2018, *New digital worlds: postcolonial digital humanities in theory, praxis, and pedagogy*, Northwestern.

Risam, R & Gil, A 2022, 'Introduction: the questions of minimal computing," *Digital Humanities Quarterly*, vol. 16.

Rudwick, S & Makoni, S 2021, 'Southernizing and decolonizing the sociology of language: African scholarship matters," *International Journal of the Sociology of Language*, nos. 267-268, pp. 259-63.

Sabino, R 2018, *Languaging without languages: beyond metro-, multi-, poly-, pluri- and translanguaging*, Brill.

Staats, S & Halpert C 2002, 'Comparisons and their magic: a commentary on "Diversity of mathematical expression: the language of comparison in English and isiXhosa early grade mathematics texts" by Ingrid Mostert and Nicky Roberts', *Research in Mathematics Education*, vol. 24, no. 1, pp. 24-27.

Stoler, AL 2016, *Duress: imperial durabilities in our times*, Duke.

Tagg, C 2020, 'Chapter 30: English language and social media' in Adolphs S & Knight, D eds. *The Routledge Handbook of English Language and Digital Humanities*, Routledge.