Towards a General Theory of Linguistic Diversity

Steven Bird
Charles Darwin University, Australia

Abstract

The world's languages are commonly categorised in terms of available language technologies such as speech recognition and machine translation. On this view, "under-resourced languages" suffer from language barriers which cut people off from markets, healthcare, human rights, and AI. The "solution" is more funding for language technologies, opening the way to a utopia of digital language equality and AI-enabled mobility. Yet the world's linguistic diversity is not a set of language objects to be pushed up a cline from emerging to thriving. It consists of polyglossic communities who have long used vernaculars for local functions and dominant languages for external functions. I present a new theory of linguistic diversity which places the world of vernaculars alongside the world of institutional languages, and articulates diverse language technology agendas that lie within and between these worlds.

1. Introduction

Within the expanse of human history, as in the arc of a human life, language is breathed before it is written. Language is embodied as utterance, talk, gesture, sign, story, spoken soul. Linguistic meaning begins with stories embedded in the land and shared for a multitude of human purposes (Basso, 1996; Abram, 1997; Kimmerer, 2013). When we set down language in writing, we perpetuate the "abstraction of linguistic meaning from the enveloping life-world" (Abram, 1997, p101).

In Australia's remote "Top End", Ngalwakadj shares the story of "Wurrbbarn", *Greedy Emu*.[1] She wants to teach children about the consequences of not honouring obligations. She uses Kunwinjku, a variety of "Kunwok". When asked to interpret this word "kunwok", she is likely to say "message" or "talk". There is no local term for this particular group of speech varieties that outside linguists have designated as a single language, assigned the name "Bininj Kunwok" *people's talk* (Evans, 2003), and associated the ISO language code [gup]. When "kunwok" is translated as *language*, English speakers may project their notion of an institutional language (cf. Fig. 1; Bird 2024).[2]

So it was for me when I began learning to speak Kunwok, and an Indigenous academic questioned the possibility that a non-Indigenous person could learn to speak an Indigenous language. I asked myself, what part of the sentence in (1), for instance, could I not understand in principle?

(1) duruk nganang
duruk ŋɑ- na -ŋ
dog 1PS- see -PST
I saw the dog

However, she was thinking that I could never be in the right relationship to the Country to be permitted to hear or share its stories (cf. Gal, 2017, p233).

[1] https://www.youtube.com/watch?v=mVvMkbUD3hg

[2] Oral societies may also have linguistic institutions.

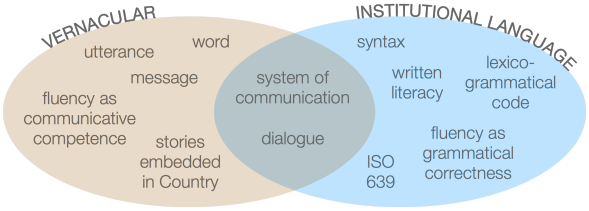


Figure 1: Vernacular versus institutional language

This problem with "language" is pervasive. What does the language resources community mean by "language"? Does "language technology" include *speech* technology or must we specify "*spoken* language technology"? In *Speech and Language Processing*, do Jurafsky and Martin (2000) assume "language" is written? "*Natural* language processing" comes from a world where language is presumed to be artificial. Machine Translation for All (MT4All) typically presupposes *textual* translation, which perpetuates the norm of written language while sidelining or pathologising *unwritten* languages, ie. languages with "*missing* standardisation" (Bird 2022, p7820; Gasparotto 2026).

It is easy to misrecognise the nature of linguistic diversity, leading to "misplaced expert attention on the language as a code rather than language as the conduit and catalyst for social relationships" (Perley, 2012, p134). A consequence is that language technology interventions may undermine the very languages they are intended to support.

In this paper, I theorise linguistic diversity in a way that allows for the epistemological distinction between vernaculars and institutional languages. I present the existing model (§2) before describing vernaculars (§3) and how they inhabit a three-level hierarchy of people-places (§4). Next, I suggest a schematisation for institutional languages (§5), and bring both models together into a general model of linguistic diversity, then articulate several language technology agendas in terms of the model (§6). The paper concludes with a discussion of language technology narratives (§7) and prospects for a basic vernacular resource kit (§8).

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 169–182
11–12 May 2026.

2. Language Objects on a Cline

The age of imperialism and the rise of mass media have brought dominant, written languages into contact with local, oral languages. Economic and technological power have generated markers of linguistic prestige such as writing, books, and media. Over time, dominant languages have come to occupy more semantic domains, gradually restricting the functions of local languages, a process that has been called "language shift" (Fishman, 2001).

This situation is represented in Figure 2. EGIDS 0–4 represents institutional languages with standardised writing and widespread literacy (<10% of the world's languages). These languages range from international to national to provincial and so on. EGIDS 5 represents a language development agenda where an orthography has been established and vernacular literacy is gaining traction. EGIDS 6a–9 represents oral languages from full (6a) and partial (6b) intergenerational transmission, to the situation where the youngest speakers are of the parent (7), grandparent (8a), or great-grandparent (8b) generations, to dormancy (9).

This cline of language vitality suggests a universal agenda of language development, although its proposers stress that each plateau represents a legitimate endpoint. Language development has been pursued under the headings of language endangerment (Krauss, 1992; Roche, 2020), reversing language shift (Fishman, 2001), language documentation and description (Himmelmann, 1998), and mid-century programs of reducing languages to writing and of literacy for preliterate peoples (Pike, 1947; Gudschinsky, 1973).

A parallel narrative has arisen in the language resources community, with its designations of zero-, low-, and high-resource languages, and the cline from "still" and "emerging" languages up to "ascending", "vital", and "thriving" languages (Fig. 3). There is no mistaking the universal development agenda.

In both framings, *languages* are the objects of discourse, and the agenda is for language professionals to intervene, moving languages up the cline. The language development agenda sets up writing and institutional recognition as normative (Fig. 2). To this, the language technology (LT4All) agenda adds text technologies such as keyboards, autocomplete, and spellchecking; more efficient ways to create text through automatic speech recognition (ASR); and access to the world's information through MT. This is what it means for a language to "thrive" in the digital era (Fig. 3). At present, these language objects are not all thriving, and this is seen as an inequality to be fixed with human language technologies:

> *Overcoming language barriers becomes crucial for the EU in the digital era. European citizens need to communicate in their own languages across the borders of Europe in order to increase workers' mobility, access to European public and private services and contents, and seize the opportunities of the Digital Single Market ... HLT are key to overcoming language barriers. (Pastor et al., 2017, p11)*

Purveying language technologies becomes a social good, promising a future in which all languages thrive and people are liberated from their language silos to participate in the common market and enjoy economic prosperity. The proposed solution: fund language engineers to construct the required resources and technologies for each language (Krauwer, 2003; Gaspari et al., 2021).

These models of linguistic diversity treat languages as bounded objects that lie on a cline, and that are subject to universal development agendas. However, so far, neither program has delivered on the promise of sustaining linguistic diversity. In fact the opposite is true: the activities of linguists and language engineers have grown hand-in-hand with the *acceleration* of global language loss (Roche, 2020). To understand why, we need to consider things from the perspective of a local oral society.

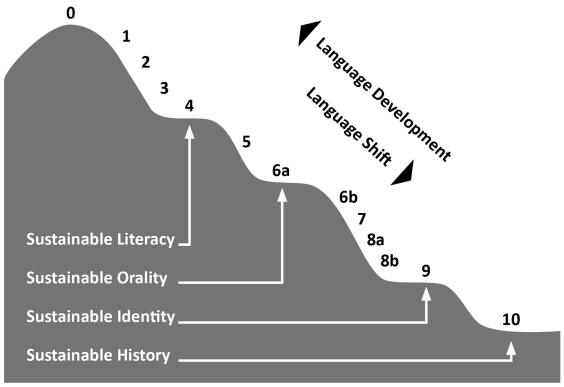


Figure 2: Linguistic Diversity as a cline of language vitality, with intergenerational disruption levels 0–10 (Lewis and Simons, 2016, p116)

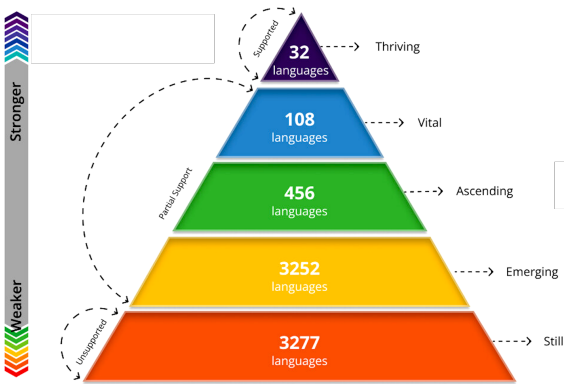


Figure 3: Linguistic diversity as a cline of digital language support (Ethnologue, 2025), using the five categories proposed by Simons et al. (2022)

3. Vernaculars and Polyglossia

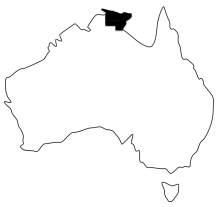
My engagement with communities in Arnhem Land began 10 years ago in a series technocolonial engagements. This included the aspiration of creating the first million word corpus of annotated speech for an Australian Aboriginal language. I based myself in Gunbalanya before moving to the homeland of Mamardawerre, 3 hours' drive to the East (Fig. 4).

In Mamardawerre, locals were not interested in having me record their speech, even for well-known stories. Nawumud told me: "I can't tell you that story, you need to ask Ngalwakadj, that story is from her country". When I found her, she asked "who is it for?" I said it was not for anyone in particular. It was just "data" to be used for making technologies. Perhaps there would be apps for converting speech to text or for translating, I mused.

I had many such conversations. They usually ended with a change of topic, or a child calling out and my interlocutor walking off. In time I was able to make a small corpus (Bird and Yibarbuk, 2024). However, I began to see the disconnected nature of recording, transcribing, and translating.

A couple of years later, I sat with two men over several days to translate the code of conduct of a land management organisation from English into Kunwok. We encountered many terms, like "risk" and "safety", having no equivalent in Kunwok. We discussed real-life scenarios, reframing pages of English legalese into a half-hour video of Kunwok dialogue. It was meaningful to interpret "placeless" English text into Kunwok talk; locals wanted to understand Western rules, assumptions, and concepts which impacted their lives.

These anecdotes align with the distinction between vernaculars and institutional languages (cf. Tab. 1). Elsewhere, this has been labelled as *langue* vs *parole*, as oracy vs literacy, and as primary orality vs primary literacy (de Saussure, 1916/2011; Wilkinson, 1970; Ong, 1982).



(a) Arnhem Land

(b) Principal towns

(c) Gunbalanya (pop 1,200)

(d) Homeland (pop 20)

Figure 4: Arnhem Land, Northern Australia

Vernacular	Institutional Language
Primary orality: • oral storytelling • learning on Country	Primary literacy: • writing, printed books • learning in school
Functions: • identity, participation • knowledge transmission • ceremony	Functions: • economic participation • information access • western work
Extent: • ancestral places • culture area	Extent: • regional • international

Table 1: Archetypal characteristics of vernaculars versus institutional languages

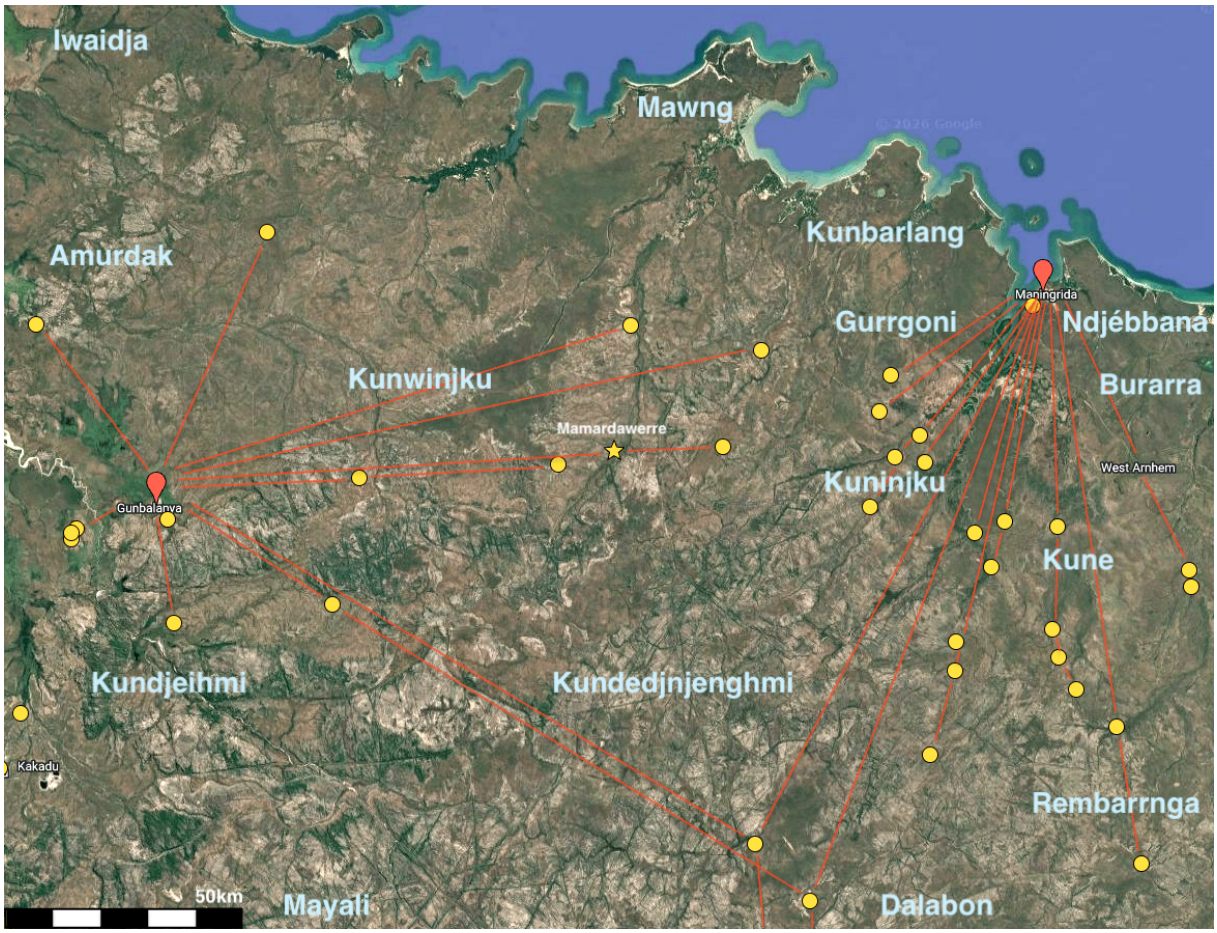
However, this binary distinction is insufficient and we need to consider the situation of "polyglossia", where a speech community uses multiple language varieties that are differentiated by prestige and function (Fishman, 2001; Lewis and Simons, 2016). This is not just the situation of two or more pure languages with the addition of code-switching. Things are more complex thanks to different levels of fusing of languages, or *language mixing* (Grosjean, 1989; Meakins, 2013; Marley, 2020), for example:

(2) Wurdurd kabirrimdolkang Mamardawerre bu ngarrimwohre ngarrini wanjh kabirridjolohre *school*. Mahne *mix* kayime kunbininjbeh dja kunbalandabeh. Wurdurd benbukkang *teacher* bu kabirrikurrme *vegetables, plants*, dja *flower*, kabenyawmarnedi bedberre *little small garden*. [030-20180608-RN-tour]
Children raised in Mamardawerre are coming to this school. It mixes Kunwok and White ways. The teacher shows the children how to plant vegetables, plants, and flowers to make this little garden here.

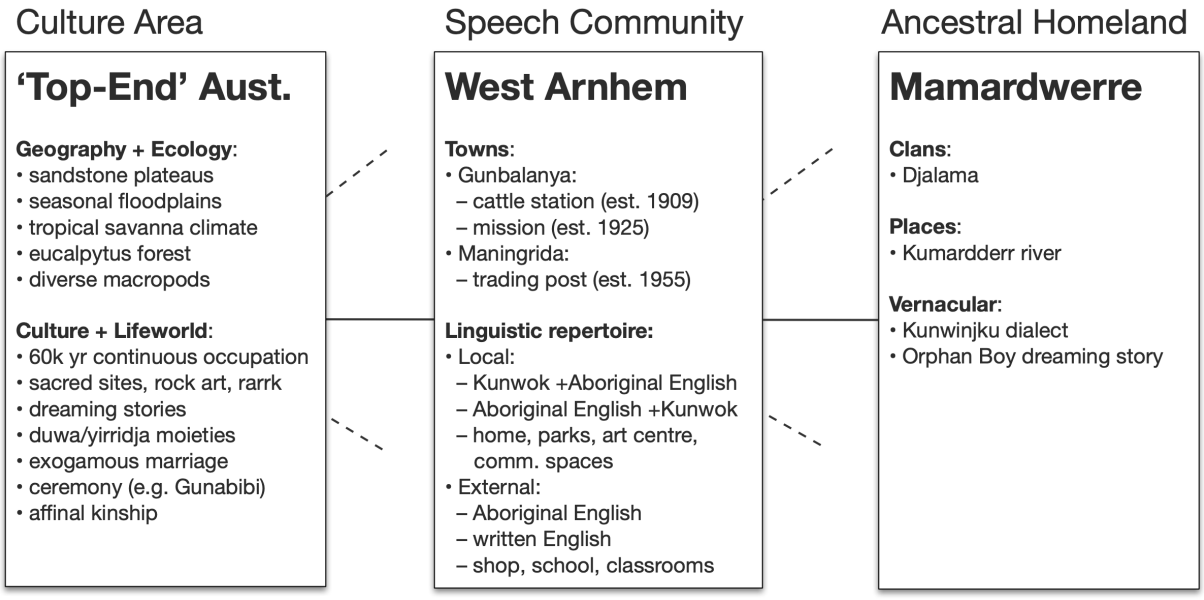
The passage in (2) is largely Kunwok but includes English words, even though Kunwok equivalents exist. The converse is also prevalent: locals speak Aboriginal-accented English and add Kunwok vocabulary which is better adapted for referring to kin, food, seasons, directions, places, and so on:

(3) *Yimray*, this is little *ngarridj*, *bebmeng koyek*, you call him *ngadjadj*.
Come, this is little [subsection name], arrived from east, who you call [mother's brother].

In view of this situation, I believe that we need a more nuanced model of linguistic diversity, one that accounts for speech communities, polyglossia, and mixed languages. To begin, we take a closer look at the cultural geography of West Arnhem (Fig. 5(a)), and observe a three-level hierarchy of "people-places" (Fig. 5(b)). This will inform the design of a new model of linguistic diversity (§4).



(a) West Arnhem homelands (yellow) with single or multiple affiliations to the townships of Gunbalanya and Maningrida (red), with areas associated with vernaculars (light blue). The region contains hundreds of sacred sites, many with ancestral (or "dreaming") stories, e.g. the Orphan Boy story is located near Mamardawerre, starred (cf. Fig. 4).



(b) A three-level hierarchy of "people-places" in which geography, society, culture, and language are co-constituted. At the top level are "culture areas", e.g. 'Top-End' Australia (Fig. 4). These are eco-cultural regions where we find substantial cultural similarities due to shared geography and long-term contact. There is no requirement of shared linguistic history, and as a case in point, Australia's Top End is home to vernaculars from several language families.

Figure 5: West Arnhem Cultural Geography: A three-level analysis, for the purpose of identifying culturally-valid language technology agendas

4. A Model of Vernacular Diversity

We now turn to the task of devising a general model of linguistic diversity. Our purpose is twofold: (a) for expert attention from the world of language technology to be placed more appropriately; and (b) to allow space for local agency which may be resistant to externally-driven language development agendas (cf. Lewis and Simons, 2016, p59).

How can we take seriously this epistemology of language as embodied, situated, and owned, of language as a way of speaking that emerges when one is in a right relationship with the country, of language as stories embedded in the land? Such language ideologies cannot be separated from the local geography, the society which inhabits that place, and the culture they manifest through language:

> *Whenever Apaches describe the land – or as happens more frequently, whenever they tell stories about incidents that have occurred at specific points upon it – they take steps to constitute it in relation to themselves. . . . With words, a massive physical presence is fashioned into a meaningful human universe (Basso, 1996, p40).*
>
> *The stories told within an oral culture [are] deeply bound to the earthly landscape inhabited by that culture. The stories [present] ways in which earthly locales may* speak through *the human persons that inhabit them (Abram, 1997, p182).*
>
> *The community is not defined by, and as, people who speak the language but by, and as, people who observe the connection of the language with the country and share possessory interests in the language (Lee, in Henderson, 2002, p4).*
>
> *What knowledge the people have forgotten is remembered by the land (Kimmerer, 2013, p369).*
>
> *Yolŋu language is in the land and in our Yolŋu story. Bringing my memory back to the origin when and where I was taught by my elders as a whole, I feel emotional satisfaction on Country and feel companionship. I am not alone. I feel brave and fearless. I do not feel afraid or scared when I am on the land with my language. I feel that I am being encouraged by the land and the soundings of my language spoken through the wind, rocks, trees, and water streams. . . The land holds the whole thing – the language and everything talks to us (Wanambi, in Muthamuluwuy et al., 2026, p9).*

Accordingly, we seek notions of place, on various scales, where language is located or embedded. A long-standing and geographically-broad cultural zone is the *culture area*, a geographical region having substantial cultural homogeneity where traditional practices, ceremonies, cultural norms, and material culture are shared (Voegelin and Voegelin, 1964; Newman, 1971; Babaii et al., 2020). This idea grew from ethnography in the colonial era, and also appears in crosscultural psychology and transnational cultural mapping (Grendstad, 1999; Beugelsdijk and Welzel, 2018). Culture areas can traverse linguistic frontiers, e.g. the Nordic region and the Baltic region, which both have Indo-European and Uralic languages. Equally, language families can traverse many culture areas, e.g. Indo-European and Austronesian.

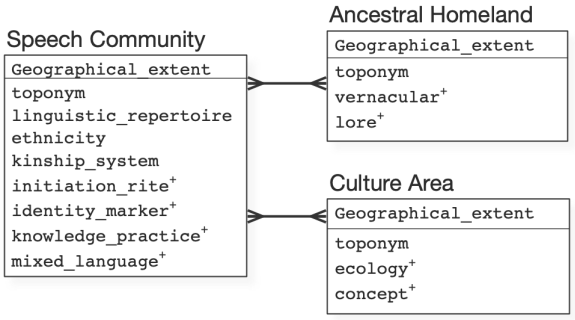


Figure 6: Entity-relationship diagram representing three levels of local culture to which linguistic practices are typically associated (provisional attributes)

A second concept, more localised, is that of a *speech community*, "any human aggregate characterized by regular and frequent interaction by means of a shared body of verbal signs and set off from similar aggregates by significant differences in language usage" (Gumperz, 1968). A speech community is defined by shared linguistic repertoire, linguistic knowledge, linguistic possessions, and linguistic behaviours (Lewis and Simons, 2016, p43).

A third concept, still more localised, is the *ancestral homeland*. This is an anchor of identity, the focus of origin stories, the locus of a multigenerational home and family roots, and of return visits by members of a diaspora (King and Christou, 2011). This is where we locate vernaculars, or dialects, which are "the natural form of language" (Kamusella, 2012, p68).

The three levels – culture areas, speech communities, and ancestral homelands – are exemplified in Figure 5(b) and formalised in Figure 6.

We understand each level as an eco-cultural monad, a "people-place" (following Christie and Verran, 2013). Each level is identified by geographical extent and associated with a principal toponym along with various other linguistic, cultural, and geographical attributes.

There is a non-strict inclusion relationship between levels to allow for overlap, as arises when local and diaspora speech communities share an ancestral homeland, or when a speech community is situated in the borderlands of two or more culture areas.

5. A Model of Institutional Languages

We now approach institutional languages from the perspective of vernaculars, and begin to see attributes that may have been taken for granted. Perhaps the most striking is the notion of a language as "bounded" (Dobrin et al., 2009), with a lexicon, grammar, and orthography that allow us to determine whether any input is contained in the language. Then there is the official definition of a language as a set of mutually intelligible dialects, frequently imposed on vernaculars of the global south while routinely ignored for dominant languages (Kamusella, 2012), which highlights the political processes behind "language making" (Krämer et al., 2022).

Governments enshrine the standard language in law and education. Books are published and sold. Scholars curate lexicons and grammars. Engineers make language technologies. None of these activities is possible without the construct of an institutional language. There are only ∼500 languages where writing is in use (Eberhard et al., 2023), far fewer than the ∼1,600 languages having a digital presence thanks to a notional orthography and a Bible translation but lacking widespread literacy and the associated language functions.

We can represent institutional languages as shown in Figure 7. Here, language resources (top right) and models (bottom right) stand in many-to-many relationships with institutional languages (left). The models represent lexical and phrasal meanings as embeddings in a high-dimensional vector space where similar meanings and translational equivalents are proximate. Language models are derived from formal language resources, but also from web scraping where detecting language identity is critical (hence, the boundedness requirement and the central place of language identifiers; Fig. 7 left).

When we view the model of linguistic diversity in Figure 7 from the standpoint of local oral societies (cf. Fig. 6), we can observe several assumptions.

Lexical overlap. The model assumes that the lexicalised concepts of different languages can usefully be projected into a common space of meanings, enabling what we could call "lexicogrammatical translation", i.e. translation by mapping and re-arranging words and short phrases, as distinct from more general, cultural translation (cf. §3).

Universality. Second is the assumption that each language covers an agreed, universal set of linguistic functions. Speech communities in oral cultures typically manifest polyglossia, i.e. a local linguistic repertoire in which different languages support different functions. Instead, the model in Figure 7 assumes each language has a corresponding community of speakers (ie. "language communities"). As a result, it is not sensitive to the different

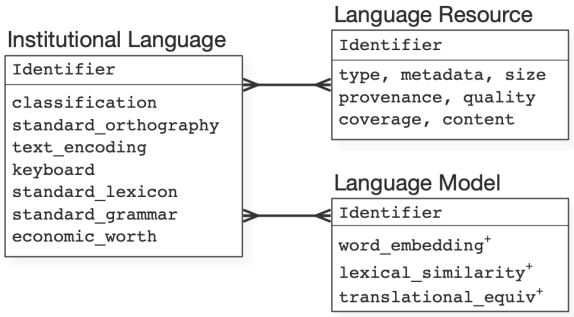


Figure 7: Entity-relationship diagram representing institutional languages and how they are supported with language resources and a shared concept space (ie. word embeddings)

sociopolitical status of, say, Spanish in Mexico versus New Mexico. All we have is language objects, and these are presented as equally valid targets for development (cf. Pastor et al., 2017).

Standardisation. In Figure 7, languages are primarily textual. Standard orthography, literary tradition, and formal education reinforce the idea that the written form is primary. This tends to obscure our view of regional spoken dialects of institutional languages, which are diverse, and usually not well supported by speech technology (e.g. Markl and Lai, 2021). This could be why discourse on the coverage of language technology tends to count languages instead of dialects (cf. §7).

Public property. We see in Figure 7 no representation of language ownership that is explicit in Figure 6. An institutional language is a public resource which anyone is entitled to learn, describe, or model. The act of language making threatens the sovereignty of a speech community over its vernacular, contravening the CARE Principles (Carroll et al., 2020).

Placelessness. We can coarsely indicate the geographical extent of a language resource in metadata using a country code, e.g. `fra-ht` for Haitian French. However, the general assumption is that institutional languages are placeless:

> *Standard languages signal anonymity in a specific sense. Their authority rests on the claim to be the voice of everyone because they are the voice of no one in particular. They seem to exemplify disembodied reason... The "voice from nowhere" and its speakers supposedly share a lack of markedness, a lack of linkage to any social group, a position above them all, representing, in this ideology, not any specific positions or interests, but science and truth itself.(Gal, 2017, pp234ff).*

These assumptions are not presented as problems to be fixed, but as evidence of differences to be navigated when we combine the two models.

6. Theorising Linguistic Diversity

We are now ready to theorise linguistic diversity, through a model that is built on the distinct, co-existing worlds of orality and literacy, and the corresponding spaces of vernaculars and institutional languages. To do this we put Figures 6 and 7 side by side (see Fig. 8). We add a relation between speech communities and institutional languages to represent the virtually ubiquitous situation where a speech community uses an institutional language. This relation is optional: uncontacted speech communities have no institutional language; and extinct institutional languages have no speech community.

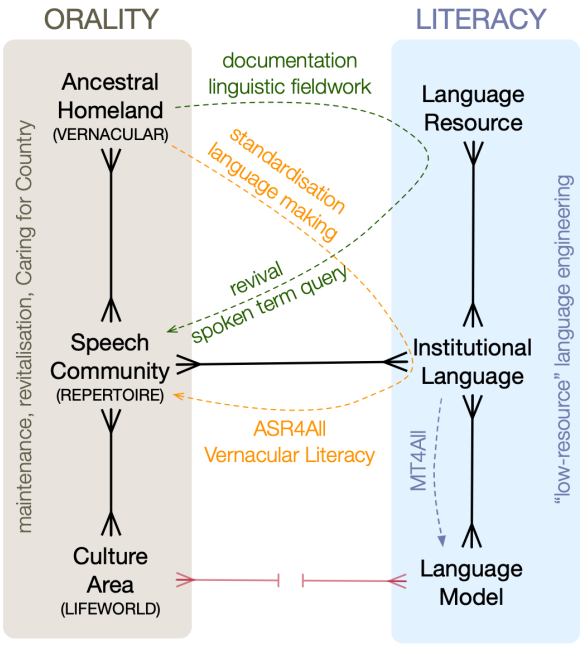


Figure 8: Entity-relationship diagram for linguistic diversity, showing technology agendas including an orality pathway (EGIDS 6, green) and a literacy pathway (EGIDS 5, orange)

Language documentation & linguistic fieldwork usually involve curation of language resources directly from vernaculars, with no-precondition of standardisation (although this may be the goal). Opportunities for language technologies are reported in the *ComputEL* and *Field Matters* workshops.

Language revival from archival records requires delivery of language resources to a speech community. **Spoken term query** would enable members of a speech community to access archived content for their vernacular(s) in the absence of an orthography. The existence of content and technology in the digital realm may support the local prestige of a vernacular (cf. Lewis and Simons, 2016, p52).

Technology may be important in the critical zone of EGIDS 6 where children are shifting to a higher-prestige variety (often an institutional language). The quandary is whether to embrace or refuse technology intrusions in the local lifeworld.

The project of **standardisation** bounds vernaculars into an institutional language (Krämer et al., 2022), a profound shift from orality to literacy and from mixed language to pure language, along with functional levelling. It sets up written languages as "better" and "more developed" than dialects (Kamusella, 2012, p69). It comes with the arduous and contentious program of orthography development (Baker, 1997; Rehg, 2004; Hinton, 2014; Doğruöz and Sitaram, 2022). An opportunity for language technology here is so-called **vernacular literacy** (e.g. Waters, 1998; Littell et al., 2022). This stage of development corresponds to EGIDS 5 (ie., the ascent from sustainable orality at EGIDS 6a to sustainable literacy at EGIDS 4).

ASR4All presents the possibility of speech technology for regional spoken dialects of institutional languages (Markl and Lai, 2021), e.g. ASR for Aboriginal English (Hutchinson et al., 2025). Here there is a precondition of standardised orthography, and so the orange arrow makes explicit the presence of an institutional language.[3]

MT4All presupposes an institutional language, for without it there can be no training data that links speech tokens to consistently-spelled words, and no effective way to represent the fact that two speech tokens correspond to a single lexical type. The above ASR4All functionality may be coupled with MT4All, permitting spoken-term cross-language information retrieval for regional spoken dialects of institutional languages. Observe that this depends on the existence of a universal semantic space implied by the presence of a (multilingual) language model with its word embeddings.

We locate **language maintenance and revitalisation efforts** on the orality side in recognition of their function in strengthening the relationship between a speech community and its ancestral homelands within its culture area. We include **Caring for Country** here, given the fact that vernaculars and their lore are intimately connected to the land (cf. §3,4). Technology might support virtual returns to Country, or re-imagining an ancestral homeland that is no longer physically accessible, or engagement with spatially indexed oral language (cf. Carew et al., 2015).

I place the agenda of **low-resource language engineering** on the literacy side, as it depends on standardised writing, and a widely-held aspiration to see institutional languages with technologies like ASR and MT (ie. language equality).

[3]There is a prospect of speech-to-speech translation between non-institutional vernaculars but I have omitted this as I am not aware of a use case; the local communicative requirement seems to be covered by receptive multilingualism (Singer, 2018).

7. Language Technology Narratives

So far, I have explored the relationship between oral and written languages and its consequences for language technology agendas. I have centered speech communities in place of language communities, and taken seriously the local linguistic ecology including polyglossia, functional differentiation, and mixed languages. This represents a fundamental shift from the framing of bounded language objects, of monolinguals cut off from AI, and the need to conquer language barriers:

> *"Voice recognition, translation, and natural language processing using AI are changing the world, but only for a select few languages," Khudanpur says. "If the language you speak is among the thousands of outliers, AI is not for you. It is a matter of digital equity." ... Khudanpur and his colleagues at CLSP are applying their myriad skills to extend AI's reach to those underserved billions. The cause is one of cultural preservation, but increasingly, Khudanpur says, also a matter of national security, global public health, and human rights (CSLP, 2023).*

This quote illustrates "misplaced expert attention" which can be attributed to misrecognition of linguistic diversity (§1). Faced with such narratives treating languages as objects needing to be pushed up a cline (§2), we point to Figure 8 and, at its centre, the relation between speech communities and institutional languages. We observe that virtually everybody on the planet can function in one of ∼500 institutional languages (or knows someone who can) in order to engage with the world outside their local speech community.

This quote promises benefits for culture, security, health, and human rights, yet by all accounts, such indicators are getting worse and the digital divide is growing (Crawford, 2021). In any case, the promised technologies are rarely delivered or evaluated (Moshagen et al., 2024). The hyperbole exists to drive a self-serving funding narrative (Fig. 9).

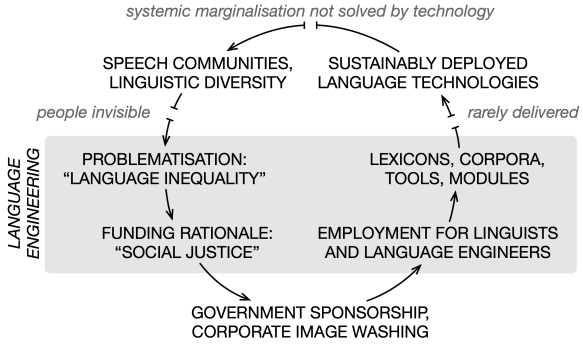


Figure 9: A perpetual funding cycle: Linguistic diversity is problematised as language inequality, "solved" with more funding for language technologies which rarely have the impact claimed for them

Notions like "digital language equality" (Gaspari et al., 2021) and "digital language extinction" (Kornai, 2013) express hopes and fears about the future of linguistic diversity. They see institutional languages as the norm, and do not recognise the legitimacy of the much larger set of vernaculars. They see information-seeking functions of language as the norm, and do not recognise communicative practices centered on identity, country, and knowledge transmission. They see "inequality" of languages and essentialise the difference between vernaculars and institutional languages in terms of resource level, and do not recognise systemic inequalities like poverty, marginalisation, and low digital literacy (Hardy et al., 2019), other than, perhaps, as a stage for showing off language technologies.

This way of seeing language perpetuates epistemic harm (cf. Alvarado, 2023; Srivastava, 2025). It values the left side of Figure 8 as a stepping stone to the right side, through clines of development (Figs. 2,3). It brings further epistemic harms:

Concerning *forms*, when we treat a vernacular as a nascent written language, just needing a phone recogniser and orthographic normalisation, we practice phonocentrism, seeing linguistic communication as no more than a sound pressure wave that can be transliterated into writing with no loss of situational meaning (Tedlock, 1983, p195). The left side allows for nonverbal communication and for "language-as-a-process-of-sustaining-relationality" (Henne-Ochoa et al., 2020).

Concerning *meanings*, we observe that the three eco-cultural levels (Fig. 8, left) are loci of meaning (§3,4). Each culture area corresponds to a local lifeworld. Most locally significant concepts have no equivalent in the embeddings of a language model. Hence the impossibility of translating between lifeworlds (Evans and Sasse, 2007; Helm et al., 2024) as revealed in the anecdote of cultural translation (§3; Fig. 8 broken red line).

How do we move forward? Figure 2 has a *threshold* for literacy (EGIDS 5), where standardised writing is not widely adopted. A Basic Language Resource Kit (Krauwer, 2003) is a *threshold* amount of language resources which could be put into service of sustainability literacy (EGIDS 4). There are many initiatives here (Kazantseva et al., 2018; Littell et al., 2022; Moshagen et al., 2023; Ravindran, 2023; Narayanan et al., 2025; Pine et al., 2025; Moeller et al., 2026), including work to address the transcription bottleneck (Foley et al., 2018; Jimerson and Prud'hommeaux, 2018; Bird, 2020b; Le Ferrand et al., 2023; Liang and Levow, 2025).

Figure 2 also has a *threshold* for orality (EGIDS 6), where children are losing their vernacular as they shift to a higher-prestige language. Here, language technologies may support programs for oral language vitality. How would we begin?

8. A Basic Vernacular Resource Kit?

Language technology engagements which reach into the world of vernaculars often presume the frame of an institutional language, "a bounded, homogeneous, structural system, a unity made primarily for denotation (i.e. reference, labelling the world), with centrally defined norms of grammatical and orthographic correctness" (Gal, 2017, p226). They reveal an impulse to make generative AI available in all the world's languages, even though there will never be enough data for this:

> *For most languages in the world there are too few speakers to produce enough text to build robust models for their language... The worst case scenario is one where the major share of online text in Indigenous languages is produced by generative neural models... The result will be that online text for Indigenous languages cannot be trusted anymore... (Moshagen et al., 2024, p103).*

I believe we should pause to reflect on such engagements with speech communities, and reflect on why we do this, and who benefits (Widder and Kneese, 2025). If most of the world's linguistic diversity is found in the space of vernaculars, why would we treat them as if they were institutional languages? The primary resource in the kit, then, is not language technology artefacts, but the language technologists who engage with speech communities to make them.

I believe that we need to ask about the scientific preparation of language technologists, and how adequately this addresses the phenomenon of human language (e.g. Ong, 1982; Bender, 2013; Grosjean, 2021; Doğruöz and Sitaram, 2022), including the epistemology of vernacular languages (Fig. 1) and the conditions for sustainable orality (Lewis and Simons, 2016). I believe we must also ask about moral development. The rhetoric of liberating underserved billions and the logic of language equality are moral starting points, after all.

We can ask to what extent language technologists take seriously the CARE Principles (Carroll et al., 2020), and the need to respect people's desire for self-determination and relationality (Bird, 2020a; Birhane, 2021; Lignos et al., 2022; Liu et al., 2022; Schwartz, 2022; Mahelona et al., 2023; Bird, 2024; Bird and Yibarbuk, 2024; Cooper et al., 2024; Markl et al., 2024; Moshagen et al., 2024).

The quote in §7 enumerates social benefits in an *instrumental performance* of morality (Bender and Hanna, 2025, ie. AI hype). A higher level of development is seen in the subfield of NLP for Social Good with *moral norms*, yet the expectation is that language technology is the answer to every question (Fortuna et al., 2021). The rhetoric of language equality applied to vernaculars also fits in here. These instrumental and moral positions are the first and second levels of a three-level theory of moral development which culminates in *universal ethical principles* (Kohlberg, 1976). I believe that it is from this third level that we can argue for our field to respect the rights of local oral societies (cf. United Nations, 2007; Patton, 2016). This includes the right to refuse technology interventions entirely. Ours is not a universal panacea.

Once we can see the genius of minoritised communities in maintaining their vernaculars and lifeworlds down to the present day, we find ourselves in a high-resource scenario, a place of "abundant intelligences" (Bird, 2022; Lewis et al., 2024). We can protect disappearing voices while there is still time (Bird, 2010; Reiman, 2010; Bird et al., 2014; Adda et al., 2016; Blachon et al., 2016; Hanke, 2017; Godard et al., 2018). We can support access via spoken term retrieval, the orality pathway in Figure 8, which has no requirement for standardisation. We can work with speech communities to develop language-specific roadmaps (e.g. Mainzinger, 2024). We continue by building on local strengths, an act of *Appreciative Inquiry* (Bushe, 2013). Sometimes this will lead to the making of another institutional language. But not always.

9. Conclusion

To the extent that the language resources community has theorised linguistic diversity, it has produced a one-dimensional model and the single agenda of digital language support (Fig. 3). This agenda is sometimes justified in terms of saving languages (Fig. 2), but there is little evidence that this has impacted the vitality of any vernaculars.

I have proposed a three-level, place-based model to represent the world of vernacular languages, and shown how it ties in with the world of institutional languages and how various language technologies are placed (Fig. 8). This model makes it clear that technologies designed for institutional languages do not in general apply to vernaculars. Speech communities use a vernacular for local participation, and an institutional language to engage with the outside world. The plethora of X4All initiatives must state when they build in assumptions of an institutional language. Only now, once we recognise that languages are not equal, are we in a position to respectfully engage with the diverse opportunities presented by the world's oral cultures.

Acknowledgements

I am profoundly grateful to the Bininj people of West Arnhem for welcoming me into their community, and to Conrad Maralngurra, Rosemary Nabulwad, Lois Nadjamerrek, and Alexandra Marley for sharing in my struggles and epiphanies with patience, love, and humour.

References

David Abram. 1997. *The Spell of the Sensuous: Perception and Language in a More-Than-Human World*. Vintage.

Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Sebastian Stüker, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Innovative technologies for under-resourced language documentation: The BULB Project. In *Workshop on Collaboration and Computing for Under Resourced Languages, International Conference on Language Resources and Evaluation*, pages 59–66. ELRA.

Ramón Alvarado. 2023. AI as an epistemic technology. *Science and Engineering Ethics*, 29(5):32.

Esmat Babaii, Mahmood Reza Atai, and Abbas Parsazadeh. 2020. A call for international recognition of culture-specific words from the Middle East. *Asian Englishes*, 22:106–110.

Philip Baker. 1997. Developing ways of writing vernaculars: problems and solutions in a historical perspective. In Andrée Tabouret-Keller, Robert B. Le Page, Penelope Gardner-Chloros, and Gabrielle Varro, editors, *Vernacular Literacy: A Re-Evaluation*, volume 13 of *Oxford Studies in Anthropological Linguistics*, pages 93–141. Oxford: Clarendon Press.

Keith Basso. 1996. *Wisdom Sits in Places: Landscape and language among the Western Apache*. UNM Press.

Emily M. Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Springer.

Emily M. Bender and Alex Hanna. 2025. *The AI Con: How To Fight Big Tech's Hype and Create the Future We Want*. Penguin.

Sjoerd Beugelsdijk and Chris Welzel. 2018. Dimensions and dynamics of national culture: Synthesizing Hofstede with Inglehart. *Journal of Cross-Cultural Psychology*, 49:1469–1505.

Steven Bird. 2010. A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 5–14.

Steven Bird. 2020a. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 3504–3519. ICCL.

Steven Bird. 2020b. Sparse transcription. *Computational Linguistics*, 46:713–744.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7817–7829. ACL.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14915–14929. ACL.

Steven Bird, Florian Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5. ACL.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Association for Computational Linguistics*, pages 826–839. ACL.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2:1–9.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouaratab, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proceedings of the Fifth Workshop on Spoken Language Technologies for Under-resourced languages*, pages 61–66. Elsevier.

Gervase R Bushe. 2013. The Appreciative Inquiry Model. In Eric H. Kessler, editor, *Encyclopedia of Management Theory*. Sage Publications.

Margaret Carew, Jennifer Green, Inge Kral, Rachel Nordlinger, and Ruth Singer. 2015. Getting in touch: Language and digital inclusion in Australian indigenous communities. *Language Documentation and Conservation*, 9:307–323.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(43):1–12.

Michael Christie and Helen Verran. 2013. Digital lives in postcolonial Aboriginal Australia. *Journal of Material Culture*, 18:299–317.

Ned Cooper, Courtney Heldreth, and Ben Hutchinson. 2024. "It's how you do things that matters": Attending to process to better serve indigenous communities with language technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–211. ACL.

Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

CSLP. 2023. No language left behind. https://www.clsp.jhu.edu/2023/06/29/no-language-left-behind/, Accessed 20260225.

Ferdinand de Saussure. 1916/2011. *Course in General Linguistics*. Columbia University Press.

Lise Dobrin, Peter Austin, and David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. *Language Documentation and Description*, 6:37–52.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97. European Language Resources Association.

David M. Eberhard, Gary F. Simons, and Charles Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26th edition. Dallas: Summer Institute of Linguistics.

Ethnologue. 2025. What is the digital language divide? https://ethnologue.com/insights/digital-language-divide, Accessed 20260215.

Nicholas Evans. 2003. *Bininj Gun-wok: A Pan-Dialectal Grammar of Mayali, Kunwinjku and Kune*. Pacific Linguistics. Australian National University.

Nicholas Evans and Hans-Jürgen Sasse. 2007. Searching for meaning in the Library of Babel: Field semantics and problems of digital archiving. *Language Documentation and Description*, 4:58–99.

Joshua A. Fishman, editor. 2001. *Can Threatened Languages be Saved?: Reversing Language Shift, Revisited: A 21st Century Perspective*. Multilingual Matters.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochví, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209. ISCA.

Paula Fortuna, Laura Pérez-Mayos, Ahmed AbuRa'ed, Juan Soler-Company, and Leo Wanner. 2021. Cartography of natural language processing for social good (NLP4SG): Searching for definitions, statistics and white spots. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 19–26.

Susan Gal. 2017. Visions and revisions of minority languages: Standardization and its dilemmas. In Pia Lane, James Costa, and Haley De Korne, editors, *Standardizing Minority Languages*, pages 222–242. Routledge.

Federico Gaspari, Andy Way, Jane Dunne, Georg Rehm, Stelios Piperidis, and Maria Giagkou. 2021. Digital language equality (preliminary definition). Technical Report D1.1, European Language Equality. https://european-language-equality.eu/deliverables/.

Melissa Gasparotto. 2026. "Missing standardization": Identifying harmful language ideologies in natural language processing work. *Big Data and Society*, 13(1):20539517251406184.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Maynard, Markus Mueller, Annie Rialland, Sebastian Stueker, François Yvon, and Marcely Zanon-Boito. 2018. A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 3366–70. ELRA.

Gunnar Grendstad. 1999. A political cultural map of Europe: A survey approach. *GeoJournal*, 47:463–475.

François Grosjean. 2021. *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press.

François Grosjean. 1989. Neurolinguists, beware! the bilingual is not two monolinguals in one person. *Brain and Language*, 36:3–15.

Sarah C. Gudschinsky. 1973. *A Manual of Literacy for Preliterate Peoples*. Ukarumpa, PNG: SIL.

John Gumperz. 1968. The speech community. *International Encyclopedia of the Social Sciences*, pages 381–386.

Florian Hanke. 2017. *Computer-Supported Cooperative Language Documentation*. Ph.D. thesis, University of Melbourne.

Jean Hardy, Susan Wyche, and Tiffany Veinot. 2019. Rural HCI research: Definitions, distinctions, methods, and opportunities. *Proceedings of the ACM Conference on Human-Computer Interaction*, 3:1–33.

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2024. Diversity and language technology: How language modeling bias causes epistemic injustice. *Ethics and Information Technology*, 26(1):8.

John Henderson. 2002. Language and native title. In John Henderson and David Nash, editors, *Language in Native Title*, pages 1–19. Aboriginal Studies Press, Canberra.

Richard Henne-Ochoa, Emma Elliott-Groves, Barbra A Meek, and Barbara Rogoff. 2020. Pathways forward for Indigenous language reclamation: Engaging Indigenous epistemology and learning by observing and pitching in to family and community endeavors. *The Modern Language Journal*, 104:481–493.

Nikolaus Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–95.

Leanne Hinton. 2014. Orthography wars. In Michael Cahill and Keren Rice, editors, *Developing Orthographies for Unwritten Languages*, pages 139–168. SIL International.

Ben Hutchinson, Celeste Rodríguez Louro, Glenys Collard, and Ned Cooper. 2025. Designing speech technologies for Australian Aboriginal English: Opportunities, risks and participation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 108–124. ACM.

Robert Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 4161–66. ELRA.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Tomasz Kamusella. 2012. The global regime of language recognition. *International Journal of the Sociology of Language*, 218:59–86.

Anna Kazantseva, Owennatekha Brian Maracle, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2018. Kawennón:nis: the wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 53–64. ACL.

Robin Wall Kimmerer. 2013. *Braiding Sweetgrass: Indigenous Wisdom, Scientific Knowledge and the Teachings of Plants*. Penguin.

Russell King and Anastasia Christou. 2011. Of counter-diaspora and reverse transnationalism: Return mobilities to and from the ancestral homeland. *Mobilities*, 6:451–466.

Lawrence Kohlberg. 1976. Moral stages and moralization: The cognitive-developmental approach. In Thomas Lickona, editor, *Moral Development and Behavior: Theory, Research, and Social Issues*, pages 31–53. Holt McDougal.

András Kornai. 2013. Digital language death. *PloS One*, 8(10).

Philipp Krämer, Ulrike Vogl, and Leena Kolehmainen. 2022. What is "language making"? *International Journal of the Sociology of Language*, 274:1–27.

Michael E. Krauss. 1992. The world's languages in crisis. *Language*, 68:4–10.

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first milestone for the Language Resources Roadmap. *Proceedings of the International Workshop on Speech and Computer*, pages 8–15.

Éric Le Ferrand, Fabiola Henri, Benjamin Lecouteux, and Emmanuel Schang. 2023. Application of speech processes for the documentation of kréyòl gwadloupéyen. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 17–22. ACL.

Jason Edward Lewis, Hēmi Whaanga, and Ceyda Yolgörmez. 2024. Abundant intelligences: Placing AI within Indigenous knowledge frameworks. *AI and Society*.

Paul Lewis and Gary Simons. 2016. *Sustaining Language Use: Perspectives on Community-Based Language Development*. SIL International.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR

models for extremely low-resource fieldwork languages. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37. ACL.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics*, pages 523–532. ACL.

Patrick Littell, Eric Joanis, Aidan Pine, Marc Tessier, David Huggins Daines, and Delasie Torkornoo. 2022. ReadAlong studio: Practical zero-shot text-speech alignment for indigenous language audiobooks. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 23–32. European Language Resources Association.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3933–3944. ACL.

Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. OpenAI's whisper is another case study in colonisation. *Papa Reo*. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/, Accessed 20250215.

Julia Mainzinger. 2024. Technology and language revitalization: A roadmap for the Mvskoke language. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 7–12. ACL.

Nina Markl, Lauren Hall-Lew, and Catherine Lai. 2024. Language technologies as if people mattered: Centering communities in language technology development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 10085–99.

Nina Markl and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: Considering user experience and language variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 34–40. ACL.

Alexandra Marley. 2020. *Kundangkudjikaberrk: Variation and change in Bininj Kunwok, a Gunwinyguan language of Northern Australia*. Ph.D. thesis, Australian National University.

Felicity Meakins. 2013. Mixed languages. In Peter Bakker and Yaron Matras, editors, *Contact languages: A comprehensive guide*, pages 159–228. Mouton De Gruyter.

Sarah Moeller, Godfred Agyapong, Jarrod Cruz, Alexis Palmer, and Mans Hulden. 2026. Computational methods for language documentation and description. *Annual Review of Linguistics*, 12:147–170.

Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41:102–116.

Sjur Nørstebø Moshagen, Flammie Pirinen, Lene Antonsen, Børre Gaup, Inga Mikkelsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain-Asikainen. 2023. The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP. In Arvi Hurskainen, Kimmo Koskenniemi, and Tommi Pirinen, editors, *Rule-Based Language Technology*, pages 70–94. Northern European Association for Language Technology.

Brenda Muthamuluwuy, Gawura Wanambi, Emily Armstrong, and Yasunori Hayashi. 2026. Holding and practising yolŋu concepts of märr and ŋayaŋu in northern australia. *Australian Review of Applied Linguistics*, pages 1–23.

R. Karthick Narayanan, Siddharth Singh, Saurabh Singh, Aryan Mathur, Ritesh Kumar, Shyam Ratan, Bornini Lahiri, Benu Pareek, Neerav Mathur, Amalesh Gope, Meiraba Takhellambam, and Yogesh Dawer. 2025. Field to model: Pairing community data collection with scalable NLP through the LiFE suite. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 76–84. ACL.

James Newman. 1971. The culture area concept in anthropology. *Journal of Geography*, 70:8–15.

Walter Ong. 1982. *Orality and Literacy: The Technologizing of the Word*. Routledge.

Rafael Rivera Pastor, Carlota Tarín Quirós, Juan Pablo Villar García, Toni Badia Cardús, and Maite Melero Nogués. 2017. *Language Equality in the Digital Age: Towards a Human Language Project*. European Parliament.

Paul Patton. 2016. Philosophical justifications for Indigenous rights. In Corinne Lennox and Damien Short, editors, *Handbook of Indigenous Peoples' Rights*, pages 13–23. Routledge.

Bernard Perley. 2012. Zombie linguistics: Experts, endangered languages and the curse of undead voices. *Anthropological Forum*, 22:133–149.

Kenneth L. Pike. 1947. *Phonemics: A Technique for Reducing Language to Writing*. Ann Arbor: University of Michigan Press.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech generation for indigenous language education. *Computer Speech and Language*, 90:101723.

Sandeep Ravindran. 2023. Frustrated that AI tools rarely understand their native languages, thousands of African volunteers are taking action. *Science*, 381:262–265.

Kenneth L Rehg. 2004. Linguists, literacy, and the law of unintended consequences. *Oceanic Linguistics*, 43:498–518.

Will Reiman. 2010. Basic oral language documentation. *Language Documentation and Conservation*, 4:254–68.

Gerald Roche. 2020. Abandoning endangered languages: Ethical loneliness, language oppression, and social justice. *American Anthropologist*, 122:164–169.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 724–731.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305. International Committee on Computational Linguistics.

Ruth Singer. 2018. A small speech community with many small languages: The role of receptive multilingualism in supporting linguistic diversity at Warruwi Community (Australia). *Language and Communication*, 62:102–118.

Shashank Srivastava. 2025. Large language models threaten language's epistemic and communicative foundations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28650–28664. ACL.

Dennis Tedlock. 1983. *The Spoken Word and the Work of Interpretation*. University of Pennsylvania Press.

United Nations. 2007. United Nations Declaration on the Rights of Indigenous Peoples. https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-indigenous-peoples.html.

Charles F Voegelin and Florence Marie Voegelin. 1964. Languages of the world: Native America, fascicle one. *Anthropological Linguistics*, 6(6):1–149.

Glenys Waters. 1998. *Local Literacies: Theory and Practice*. Summer Institute of Linguistics, Dallas.

David Gray Widder and Tamara Kneese. 2025. Salvage anthropology and low-resource NLP: What computer science should learn from the social sciences. *Interactions*, 32(2):46–49.

Andrew Wilkinson. 1970. The concept of oracy. *The English Journal*, 59(1):71–77.