Language Technology and the Metacrisis

Steven Bird

Australasian Language Technology Workshop
Canberra, 2-4 December 2024
linktr.ee/stevenbird

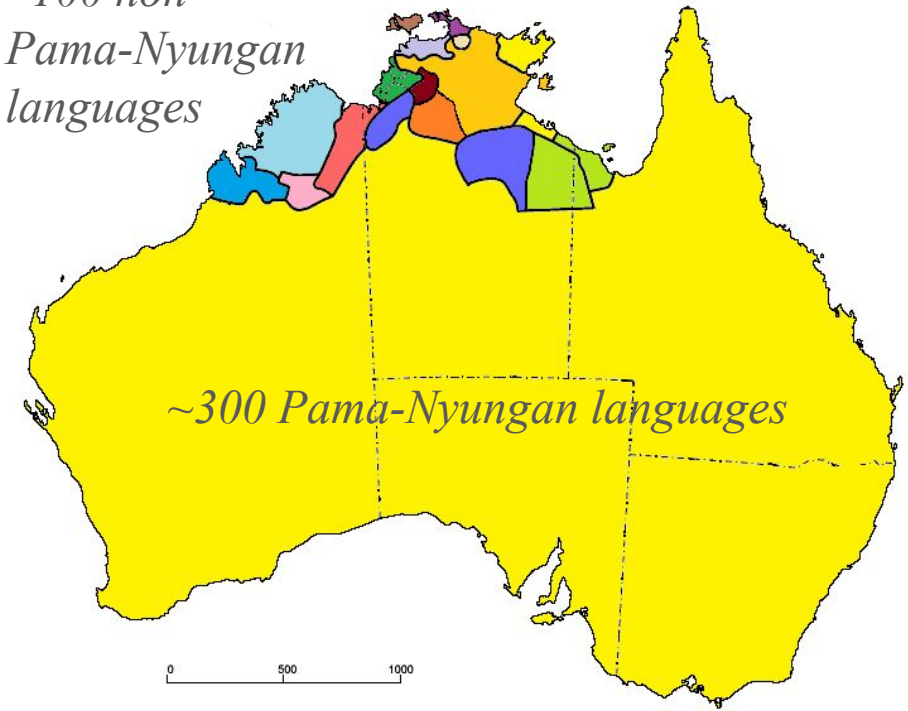
NORTHERN INSTITUTE
People. Policy. Place.

TOP END LANGUAGE LAB
CHARLES DARWIN UNIVERSITY

CHARLES DARWIN UNIVERSITY AUSTRALIA

Australian Language *Families*

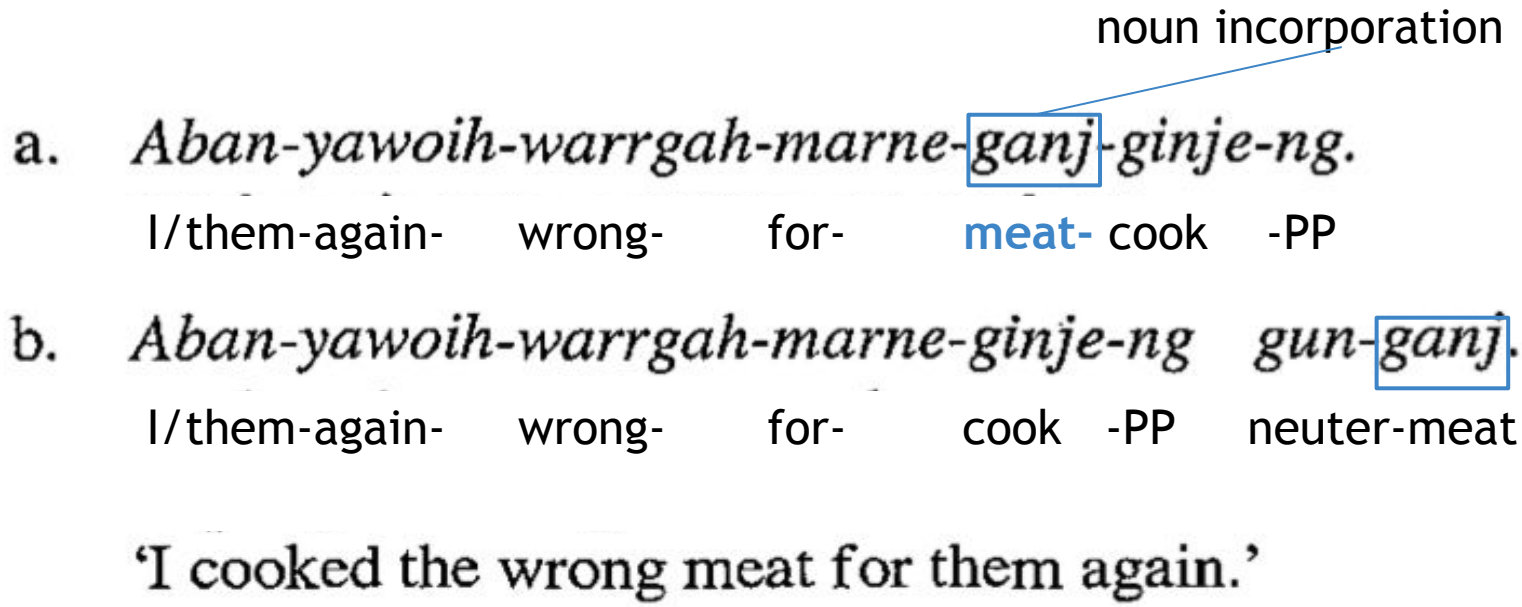
~100 non Pama-Nyungan languages



~10 Tasmanian languages

- Nyulnyulan
- Worrorran
- Bunuban
- Jarrakan
- Mindi (2 areas)
- Daly (4 families)
- Wagiman
- Yangmanic
- Tiwi (offshore)
- Darwin Region
- Iwaidjan
- Giimbiyu
- Arnhem, incl. Gunwinyguan
- Garawan and Tangkic
- Pama–Nyungan (3 areas)

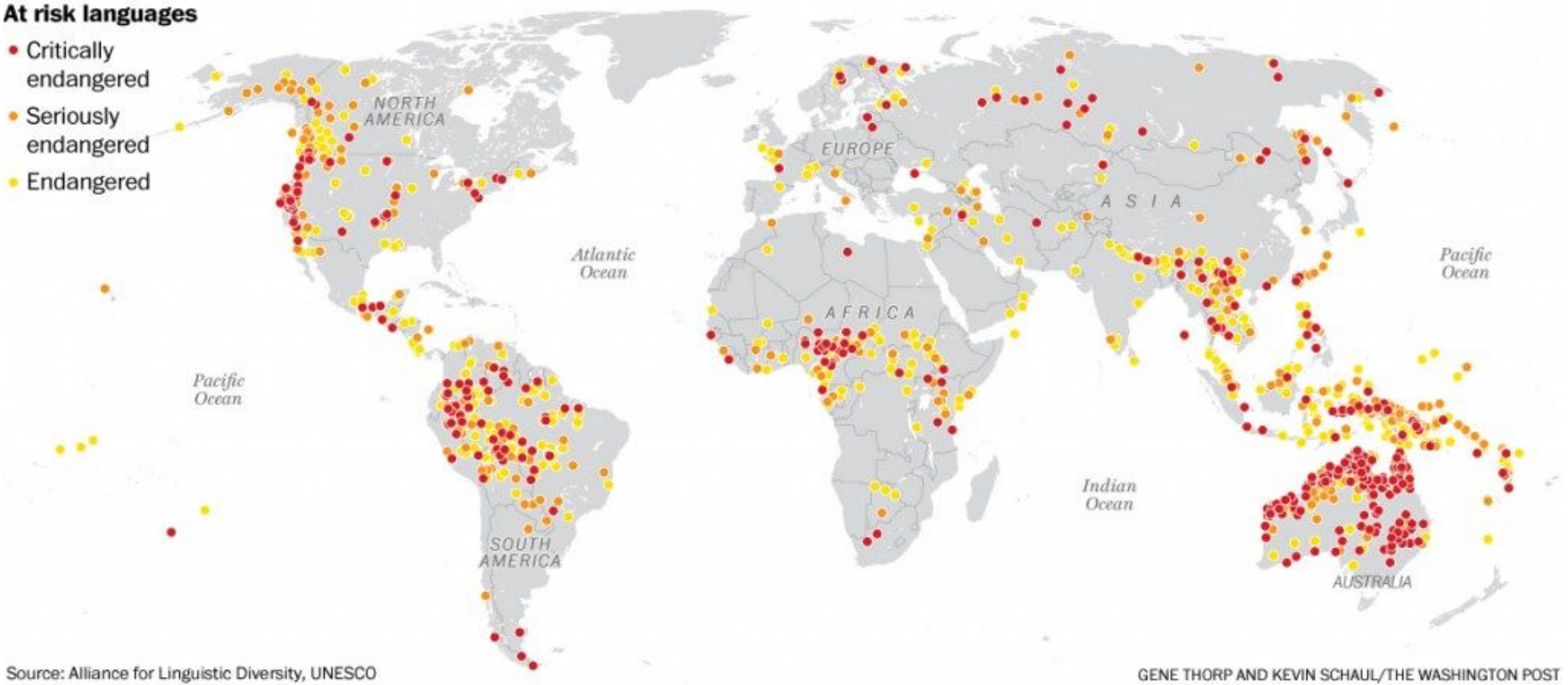
Kunwinjku: A polysynthetic language (pop. 2,000)



Evans, N (2003). Grammar of Bininj Gunwok

Losing languages

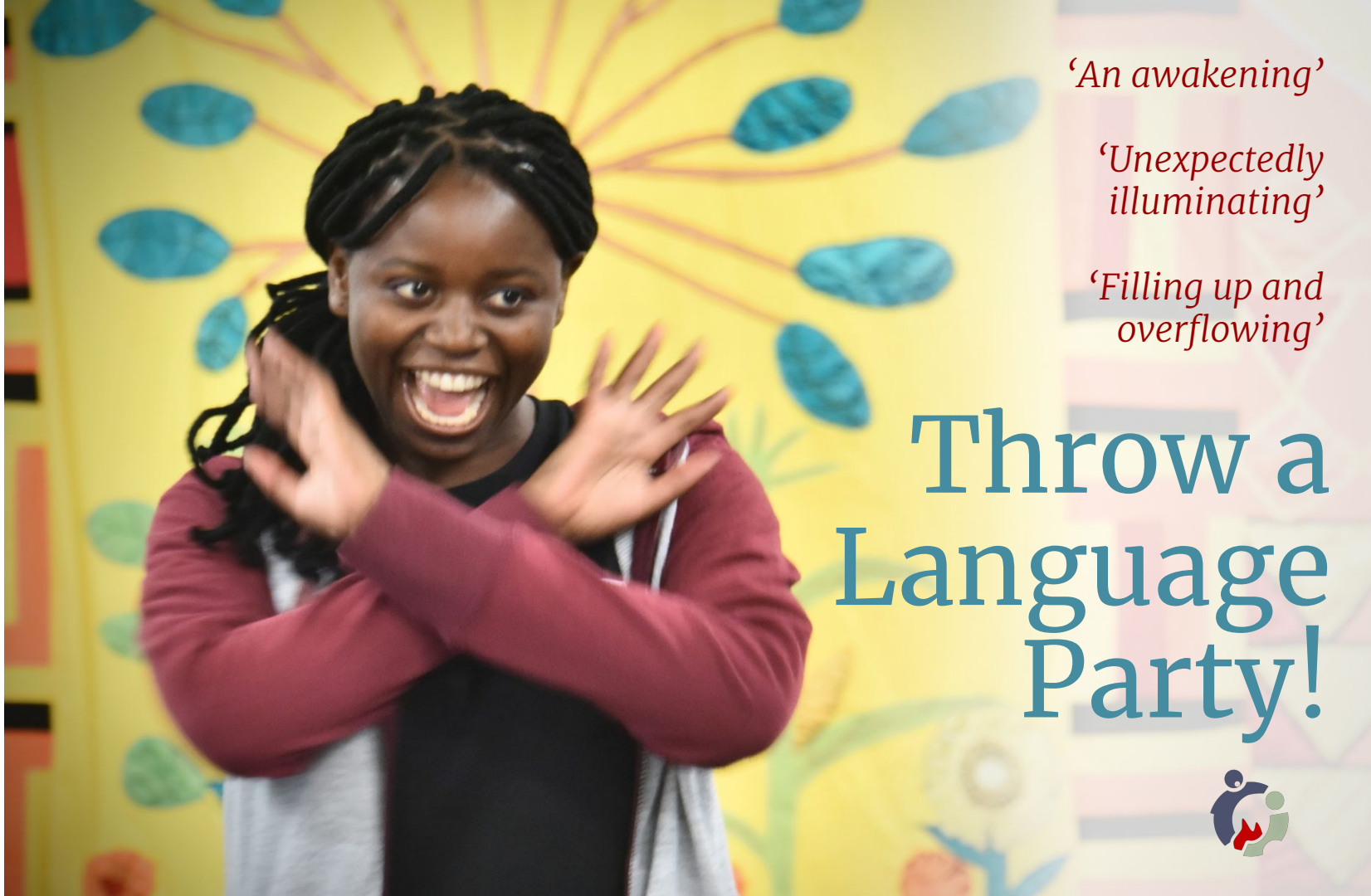
There are currently some 6,000 languages being spoken in the world today. About 96 percent of which are spoken by only perhaps three percent of the world's population. The United Nations estimates that the vast majority of these languages will be replaced by dominant ones by the end of the century.



Source: Alliance for Linguistic Diversity, UNESCO

GENE THORP AND KEVIN SCHAUL/THE WASHINGTON POST

languageparties.org



‘An awakening’

‘Unexpectedly illuminating’

‘Filling up and overflowing’

Throw a Language Party!

Background reading, with citations to related work

https://aclanthology.org/people/s/steven-bird/

1. Centering the speech community
   Bird & Yibarbuk, EACL'24
2. Must NLP be extractive?
   Bird, ACL'24
3. Envisioning NLP for intercultural climate communication
   Bird, Aquino & Gumbula, ClimateNLP'24

BACKGROUND 1:

6,500+ oral languages with diglossia

Typical low-resource NLP agenda ✔

Diglossia ⇓ *functional differentiation* ⇓ *different opportunities!*

Language Vitality Status (EGIDS)	Living Languages	Median Population
(a) ***490 Institutional Languages***		
International (0)	6	263,318,175
National (1)	99	6,260,290
Provincial (2)	44	1,802,500
Wider Communication (3)	172	884,900
Educational (4)	169	277,000
(b) ***5,241 Oral Languages*** *(learnt by children)*		
Developing (5)	1,637	34,100
Vigorous (6a)	1,963	12,900
Threatened (6b)	1,641	2,800
(c) ***1,437 Oral Languages*** *(not learnt by children)*		
Shifting (7)	438	1,500
Moribund (8a)	356	250
Nearly Extinct (8b)	313	12
Dormant (9)	330	

Centering the Speech Community

Steven Bird
Northern Institute
Charles Darwin University
Darwin, Australia

Dean Yibarbuk
Warddeken Land Management
Kabulwarnamyo
West Arnhem, Australia

Abstract

How can NLP/AI practitioners engage with oral societies and develop locally appropriate language technologies? We report on our experience of working together over five years in a remote community in the far north of Australia, and how we prototyped simple language technologies to support our collaboration. We navigated different understandings of language, the functional differentiation of institutional *vs* oral languages, and the distinct technology opportunities for each. Our collaboration unsettled the first author's western framing of language as data for exploitation by machines, and we devised a design pattern that seems better aligned with local interests and aspirations. We call for new collaborations on the design of appropriate technologies for oral languages.

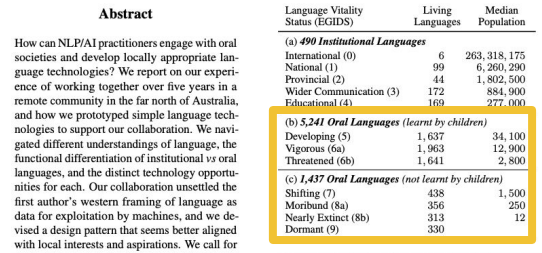


Figure 1: Distribution of Languages by Vitality, as measured using the Expanded Intergenerational Disruption Scale (EGIDS, Simons and Lewis, 2013), with statistics drawn from (Eberhard et al., 2023)

1 Introduction

The world's living languages can be categorised into ∼500 *institutional languages* and a further ∼6,500 local vernaculars, or *oral languages* (Fig. 1). Institutional languages feature standardised orthographies and widespread literacy. Local languages feature 'primary orality' (Ong, 1982), and include ancestral languages with an unbroken history of oral transmission and languages in danger of disappearing. This paper addresses the languages in Figure 1(b), which still play a significant role in intergenerational knowledge transmission, also known as 'languages with sustainable orality' (Lewis and Simons, 2016). In such speech communities, people interact with the outside world using a language of wider communication, often a variety of an institutional language.

For example, the speech community in Gunbalanya in the remote north of Australia relies on Kunwinjku [gup] (pop. 2,000) for local interaction, alongside Aboriginal English as the language of wider communication. The latter is the natural target for the usual suite of language technologies, including speech to text and machine translation, supporting participation in the global information society (cf. Bird, 2022). What do we offer a local language like Kunwinjku? One answer is that we offer it the same technologies as the institutional languages, under the belief that all languages are equal. Yet all languages are not equal, in the sense that *languages are functionally differentiated within the linguistic repertoire of speech communities*. In light of this reality, how might we engage local speech communities in the design of language technologies?

In this paper, we centre the needs, desires and aspirations of a local speech community as we rethink the design of language technologies. What are good ways in from outside, i.e., approaches for 'newcomers' to engage with 'locals'?[1] Our starting point is respect for the agency of local people and a commitment of newcomers to embrace local

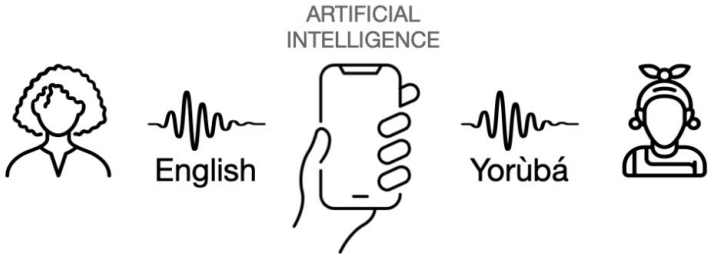
[1]We adopt the terminology of Wagner 2015.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 826–839

BACKGROUND 2:

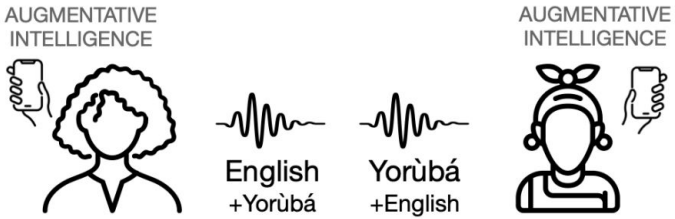
Designs for building human capacity

Build machines to "conquer the language barrier"



(a) Communication is hostage to the machine which must model all layers of communicative interaction

Build human capacity to work interculturally



(b) Communication is amplified by the machine which provides an imperfect but helpful assistant

Must NLP be Extractive?

Steven Bird
Northern Institute
Charles Darwin University
Darwin, Australia

Abstract

How do we roll out language technologies across a world with 7,000 languages? In one story, we scale the successes of NLP further into 'low-resource' languages, doing ever more with less. However, this approach does not recognise the fact that – beyond the 500 institutional languages – the remaining languages are oral vernaculars. These speech communities interact with the outside world using a 'contact language'. I argue that contact languages are the appropriate target for technologies like speech recognition and machine translation, and that the 6,500 oral vernaculars should be approached differently. I share stories from an Indigenous community where local people reshaped an extractive agenda to align with their relational agenda. I describe the emerging paradigm of Relational NLP and explain how it opens the way to non-extractive methods and to solutions that enhance human agency.

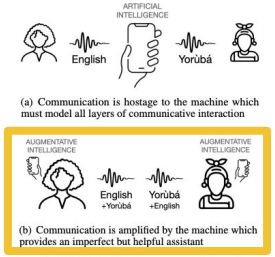


Figure 1: LT4All Design Patterns: machine *vs* human learning; simulating *vs* supporting humans; diminishing *vs* enhancing agency; monolingualism *vs* language mixing, translanguaging, and receptive multilingualism

1 Introduction

For over half a century this community has been developing methods for so-called 'natural' language processing (NLP). By *natural* this community does not mean the kinds of spoken interaction most people would regard as natural. We mean documents containing a textual trace of human language, as distinct from the default kind of language to be processed by computer, which is apparently programming languages. I believe that generative AI and large language models misconstrue the nature of language, and I argue that it is time for the NLP community to take 'natural language' seriously.

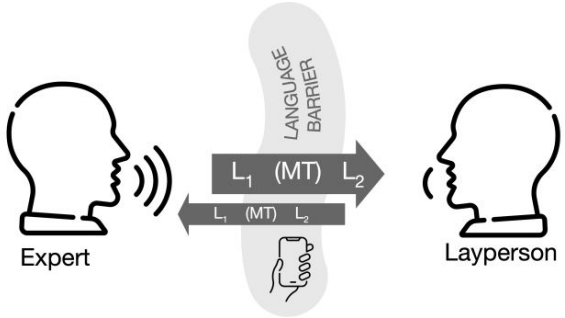
Meta's project "No Language Left Behind" promises to enable people to make "more *meaningful* connections in their preferred or native languages, [bringing] people together on a global scale" (Meta, 2023). Google's Universal Speech Model will "*understand* the world's 1,000 most-spoken languages" (Roth, 2023). The chatbots are going massively multilingual.

So it was that I listened while an African scholar described the prospects for his friend in Switzerland to learn ancestral food practices from her grandmother in Nigeria. A translation app would solve the language barrier, he mused. I sketched the scenario (Fig. 1(a)). Yes, that's it, he said. I asked if this system would need to be trained on familial conversations with an interpreter in the middle, to be replaced by his app. Which of the 20+ dialects of Yorùbá would he pick? It would need to handle words for ingredients and implements that have no translation. And how would this system interpret the kinds of utterance that are common between family members, whose meaning depends on shared knowledge that the system has not been exposed to? We sat in silence. Yes, it's a problem, he said, and even if it was possible, it would take too long. I asked if the woman already knew some Yorùbá and if she adds it to her English. Yes, she already does that, he said, and she wants to learn more. I drew another diagram (Fig. 1(b)).

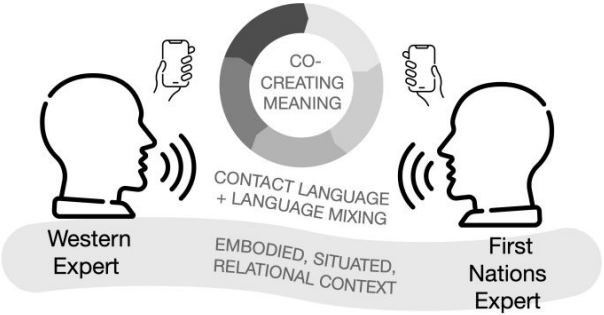
BACKGROUND 3:

Beyond the noisy channel model

Add MT to the noisy channel



Support co-construction of meaning



Envisioning NLP for Intercultural Climate Communication

Steven Bird, Angelina Aquino, and Ian Mongunu Gumbula
Northern Institute, Charles Darwin University
Darwin, Australia

Abstract

Climate communication is often seen by the NLP community as an opportunity for machine translation, applied to ever smaller languages. However, over 90% the world's linguistic diversity comes from languages with 'primary orality' and mostly spoken in non-Western oral societies. A case in point is the Aboriginal communities of Northern Australia, where we have been conducting workshops on climate communication, revealing shortcomings in existing communication practices along with new opportunities for improving intercultural communication. We present a case study of climate communication in an oral society, including the voices of many local people, and draw several lessons for the research program of NLP in the climate space.

1 Introduction

Central to climate action is communication – not only among climate scientists, industry leaders, and heads of state – but across all peoples and levels of society, for understanding, collaboration, and behavioural change. A common assumption is that climate communication consists of broadcast of information from 'experts' to 'laypeople', on the belief that "the public are 'empty vessels' waiting to be filled with useful information on which they will then rationally act", ie. the so-called *information deficit model* (Ockwell et al., 2009, p321). However, effective climate communication calls for engagements that connect with people's values, identities, and motivations, through culturally-appropriate language and modes of discourse (Nerlich et al., 2009).

How do we meet this challenge using language technologies? In particular, how can language techologies support actors from diverse cultures and standpoints to develop mutual understanding and respect for each other's knowledge practices, and to work together in devising effective and sustainable solutions? This is intercultural work in that it exceeds the definition of communication as a mere conduit for the transfer of information from expert to layperson, and of machine translation as mere substitution and rearrangement of word sequences to surmount language barriers (cf. Bird, 2024).

We, all researchers based at Charles Darwin University (CDU), are engaging with remote Aboriginal communities in the far north of Australia. In the course of this early work, we have observed how intercultural communication problems go beyond what can be addressed by *machine translation inside the information conduit* (see Fig. 1). The differences can be traced to linguistic and cultural differences which are not well handled in NLP, as others have also noted (Liu et al., 2021; Hershcovich et al., 2022).

We present viewpoints coming from local communities that point to an alternative approach that involves co-creating meaning amongst participants, leading to new possibilities for language technologies.

This paper is organised as follows. In Section 2 we discuss climate communication as an opportunity for NLP, focussing on oral languages outside the 500 institutional languages. In Section 3 we examine the social geography of Arnhem Land, Aboriginal country in Australia's Top End, including the various institutional and Indigenous stakeholders, and including our own positionality as authors. In Section 4 we report on our field-based research including two workshops on climate communication which brought together these stakeholders in order to document local matters of concern and explore new avenues for more effective intercultural communication. In Section 5 we reflect on the findings and draw out lessons for NLP which are guiding our ongoing field-based research. Finally, Section 6 presents conclusions and future prospects.

Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), pages 111–122
August 16, 2024 ©2024 Association for Computational Linguistics

Why?


behaviours
beliefs,
intentions

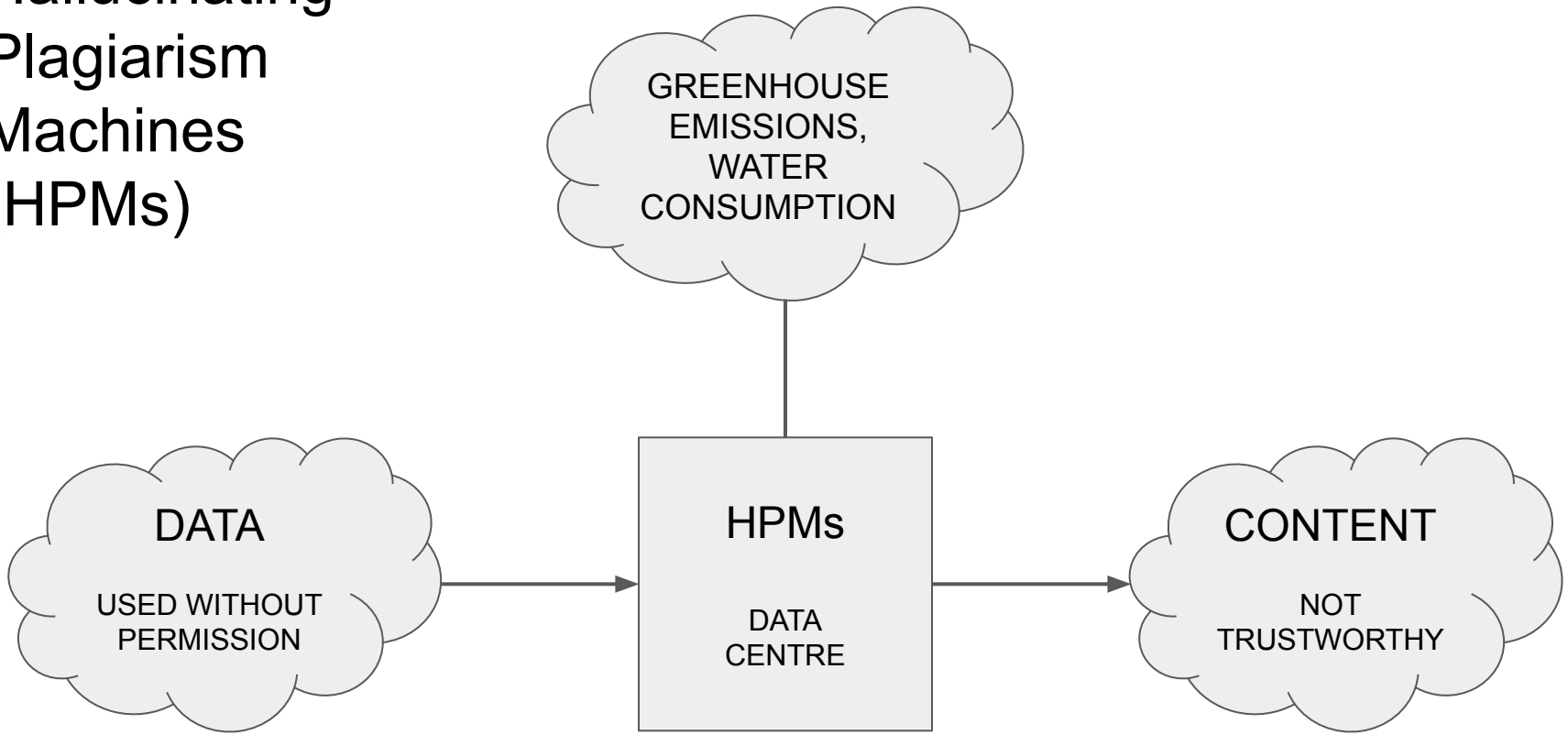
rivalrous & excludable

rivalrous

excludable


behaviours
rewards & penalties
(enforcing public-spirited behavours)
beliefs,
intentions
private interest

Hallucinating Plagiarism Machines (HPMs)



What does technology do?

~~*Technology creates a level playing field*~~

Technology magnifies existing forces

“Technology’s Law of Amplification”

–Kentaro Toyama

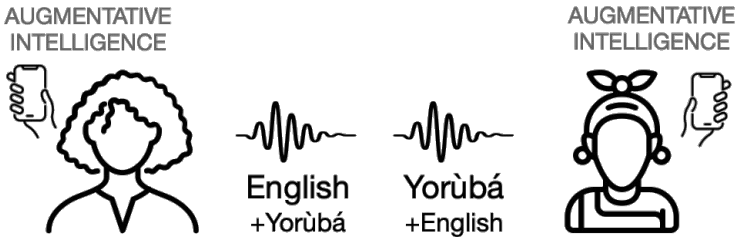
Serina (my adoptive sister)

Ursula von der Leyen (EC President)

Metacrisis?

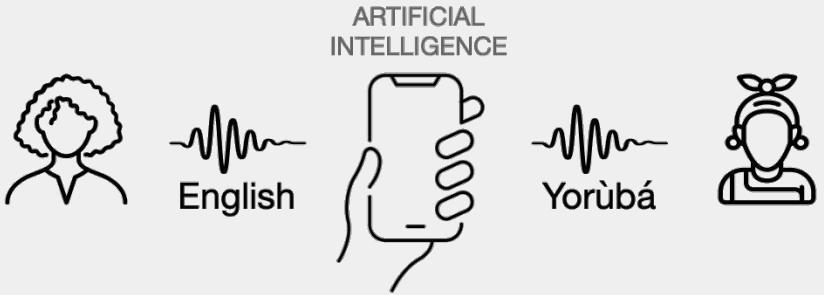

behaviours
language data
beliefs,
intentions
social & cultural meaning

Augmentatitive Intelligence ✔



Doug Englebart
SRI International
1960s

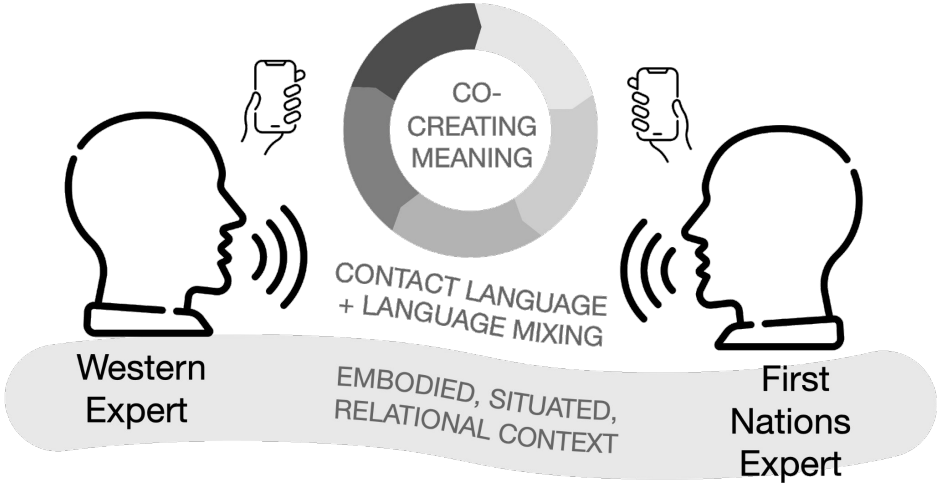
Artificial Intelligence ✗



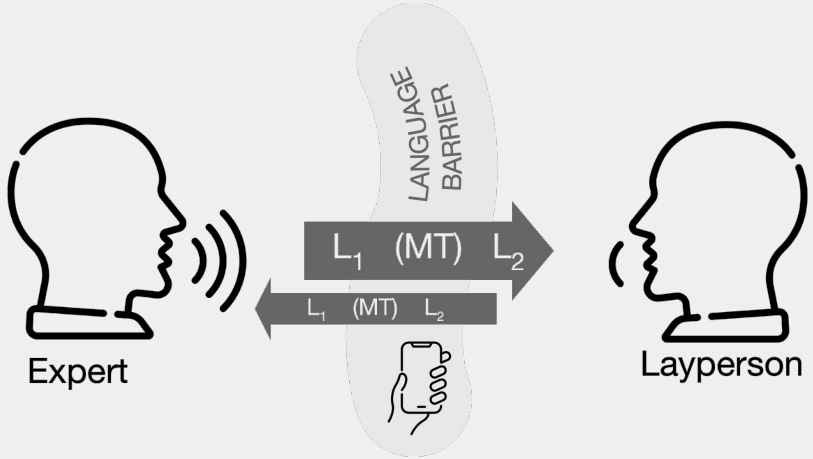
John McCarthy
Stanford University
1960s

Must NLP be extractive? S Bird, ACL'24

Co-creating meaning with language mixing and assistive technologies ✔

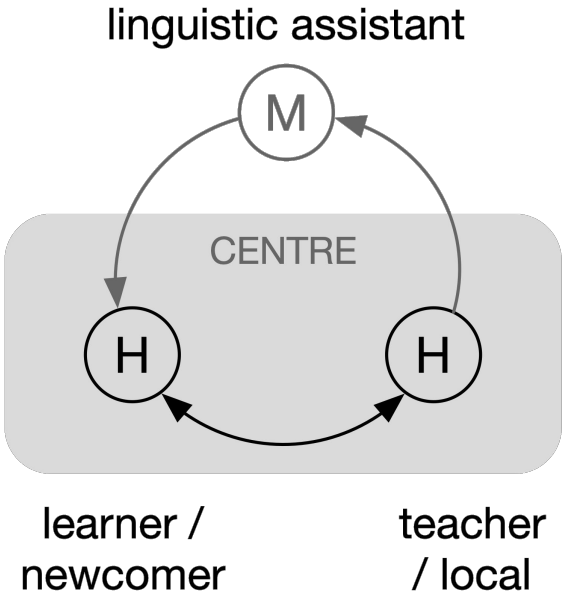


Information transmission by transmitting and translating documents ✗

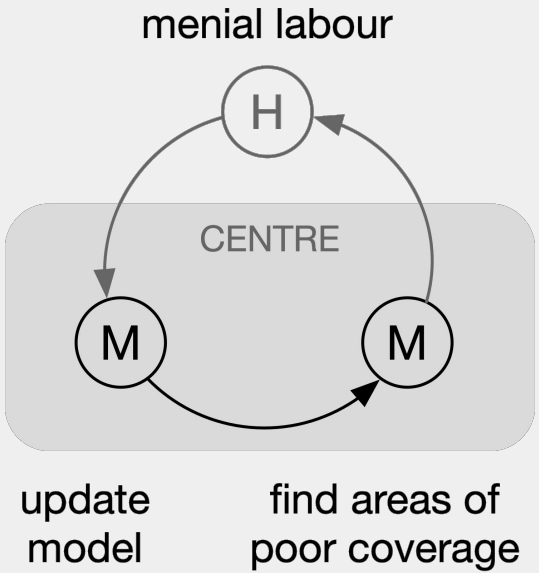


Envisioning NLP for intercultural climate communication, Bird, Aquino, Gumbula, ClimateNLP'24

Language as Situated and Embodied Social Practice ✔



Language as Data ✗



Centering the speech community, S Bird & D Yibarbuk, EACL'24

	CLOSED language-as-code; language-as-data; monolingual mindset	OPEN language-as-situated-embodied-social-practice; diglossia
STATIC	**document the language** (we create rich archival records and preserve the language in perpetuity)	**leave them be** (locals use the contact language to participate in the information society)
DYNAMIC	**develop the language** (we deploy the full suite of language technologies; “all languages are equal!”)	**?**

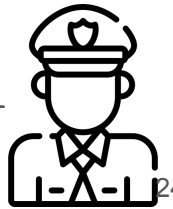
Intercultural communication

Take two pills twice daily for 5 days. Do you understand?

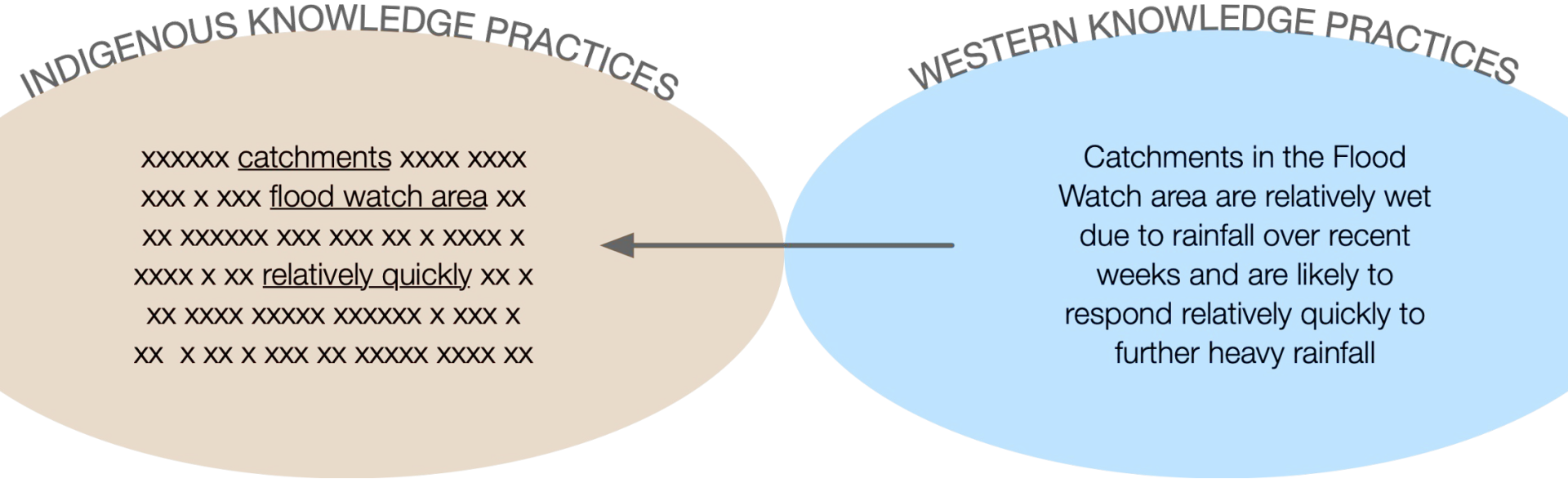
Catchments are likely to respond relatively quickly to new rainfall

Self-isolate if you develop respiratory symptoms

Clear your premises of potential wind-borne missiles

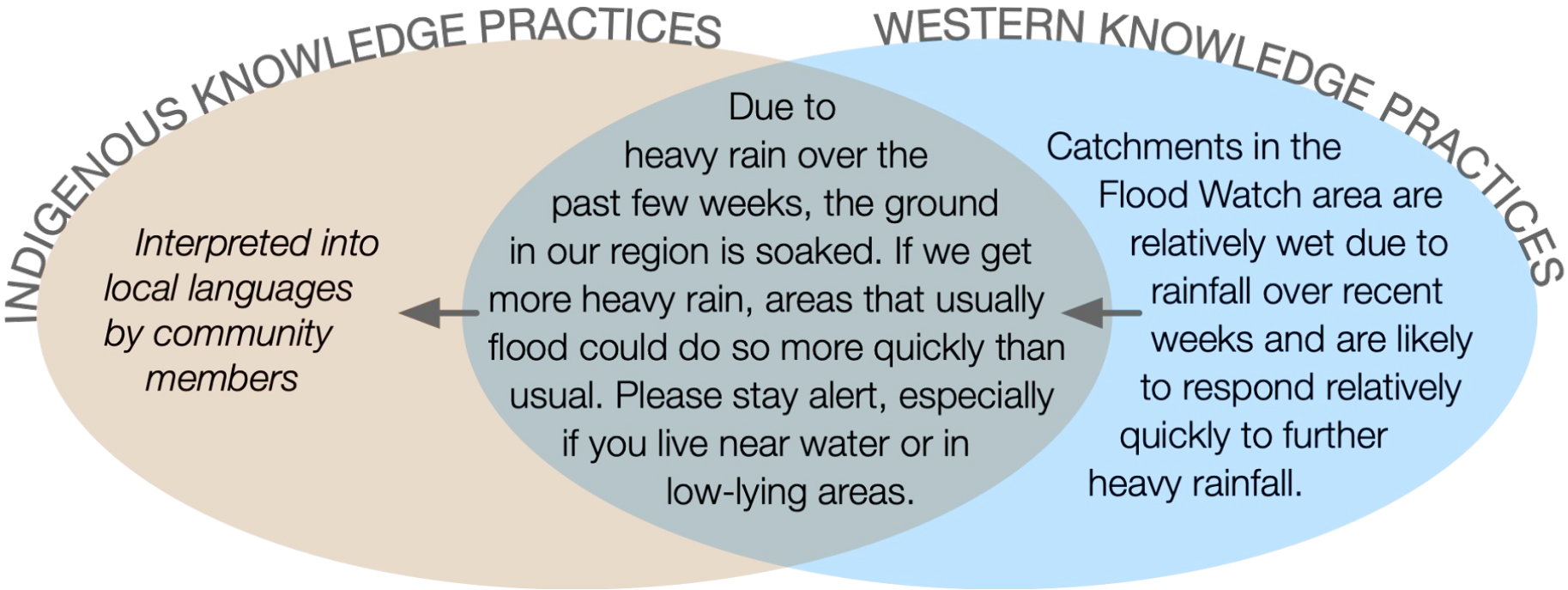


Lexicogrammatical Translation

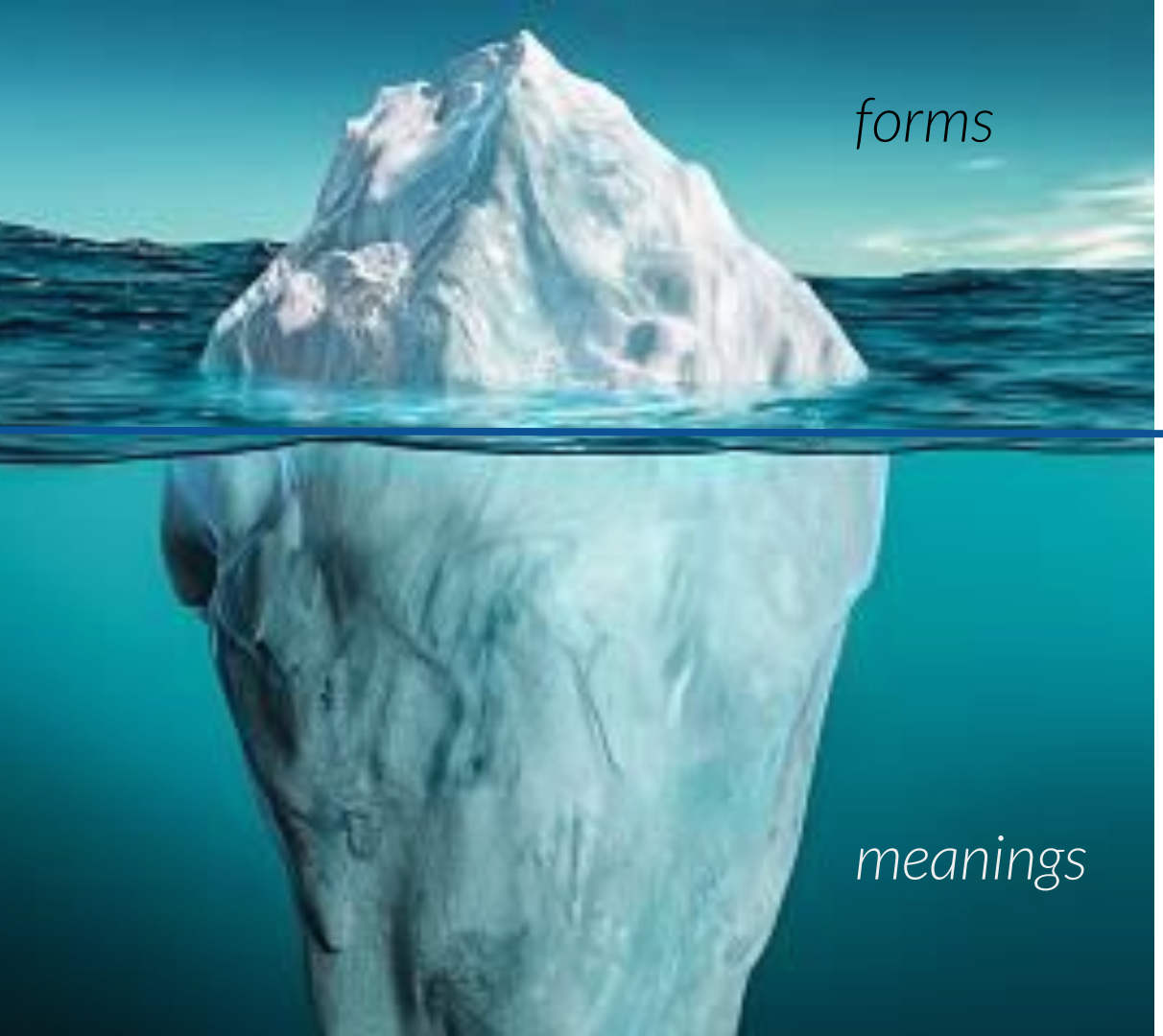


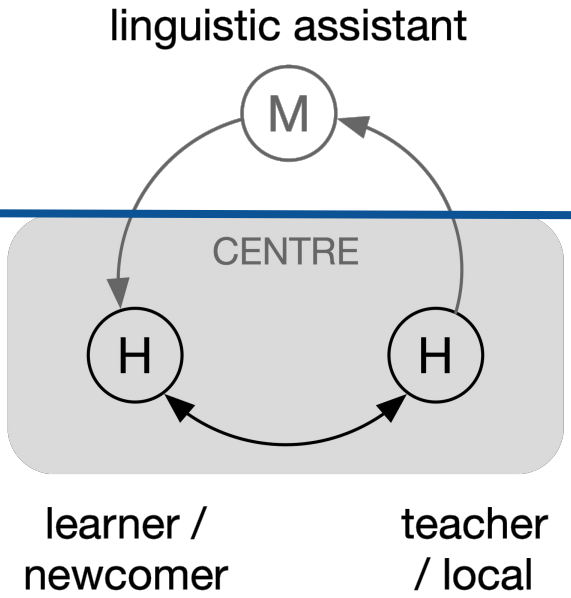
Quotes: Sapir, Evans

Cross-Cultural Translation (English$_1$ to English$_2$)



	CLOSED language-as-code; language-as-data; monolingual mindset	OPEN language-as-situated-embodied-social-practice; diglossia
STATIC	**document the language** (we create rich archival records and preserve the language in perpetuity)	**leave them be** (locals use the contact language to participate in the information society)
DYNAMIC	**develop the language** (we deploy the full suite of language technologies; “all languages are equal!”)	**work together** (locals and newcomers; building capacity in intercultural spaces; assistive technologies)





Centering the speech community, S Bird & D Yibarbuk, EACL'24

STORY ON THE GROUND

1. Diglossia: local language; contact language
2. Functional differentiation of speech varieties (identity, country, knowledge transmission)
3. Diverse positions on if/how to develop the local language (90% oral / “unwritten”)
4. Many situations of locals and newcomers coming together (healthcare, education, construction, land management, …)

✔ activates leaders

✔builds human capacity (scales exponentially)

TECHNOLOGY STORY

1. Language for information access
2. But there’s a “language barrier”
3. Deploy LT to “remove language barrier”

Q.E.D.

✔ “Scaled” (for 1000+ languages!!!!)

✗ Data centres accelerating the climate crisis

✗ Performance on small languages not measured

✗ Literal translations not situationally aware

How can we say LT offers a “scalable solution”?

Resources

workthatreconnects.org

sarahwilson.substack.com

thegreatsimplification.com