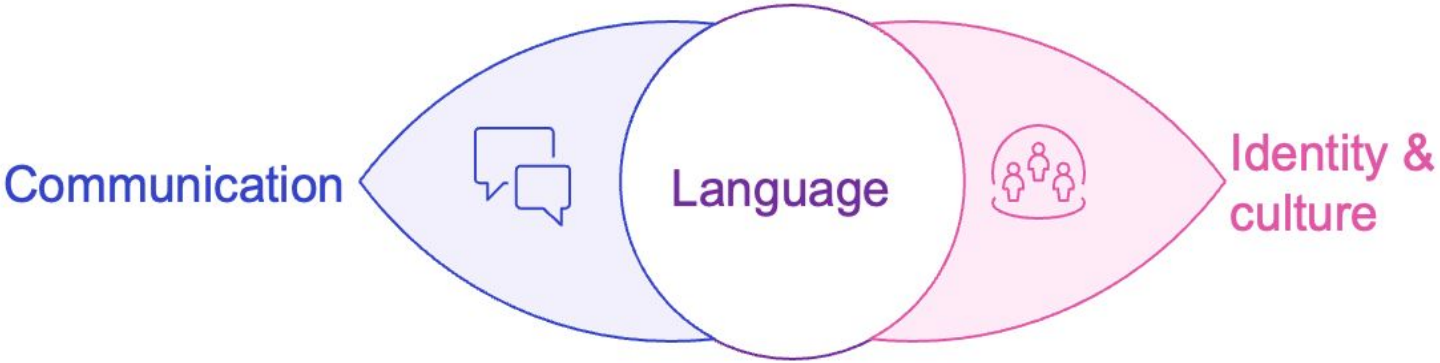
Building Better:

Avoiding Pitfalls in Developing Language Resources when Data is Scarce

Nedjma Ousidhoum, Meriem Beloucif, Saif M. Mohammad
OusidhoumN@cardiff.ac.uk, meriem.beloucif@lingfil.uu.se, saif.mohammad@nrc-cnrc.gc.ca

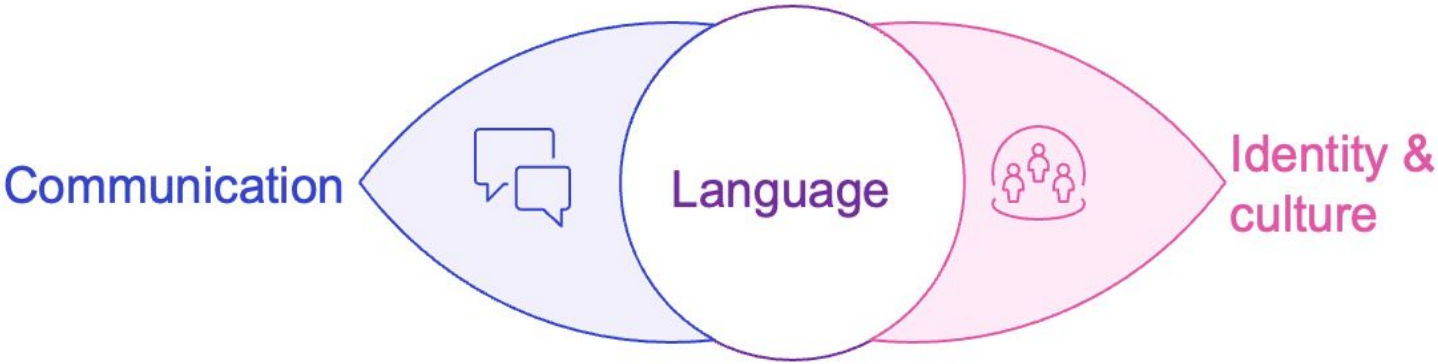
Motivation

- Language is a powerful communication means
- Rigorous data creation practices are essential for developing more human-centered and socially aware language technologies
  - This is challenging when the language is underserved



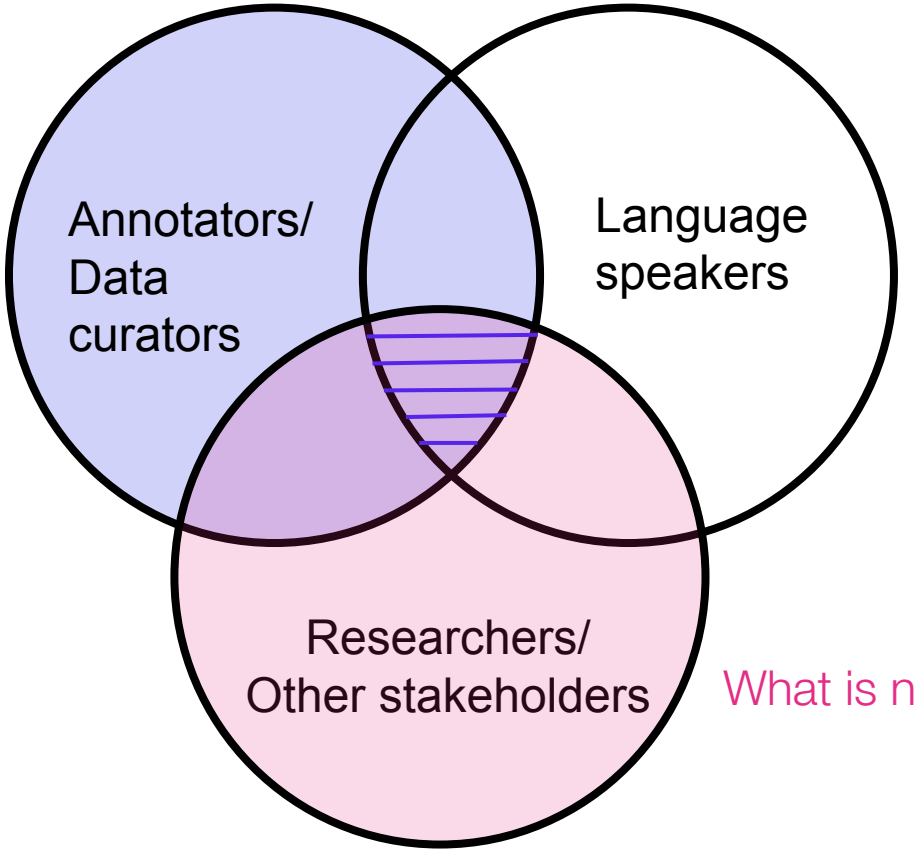
Motivation

We engage with those directly involved and impacted by NLP artefacts for underserved languages



Survey on NLP for Mid- to Low-resource Languages

How is data curated/annotated?



What is needed?

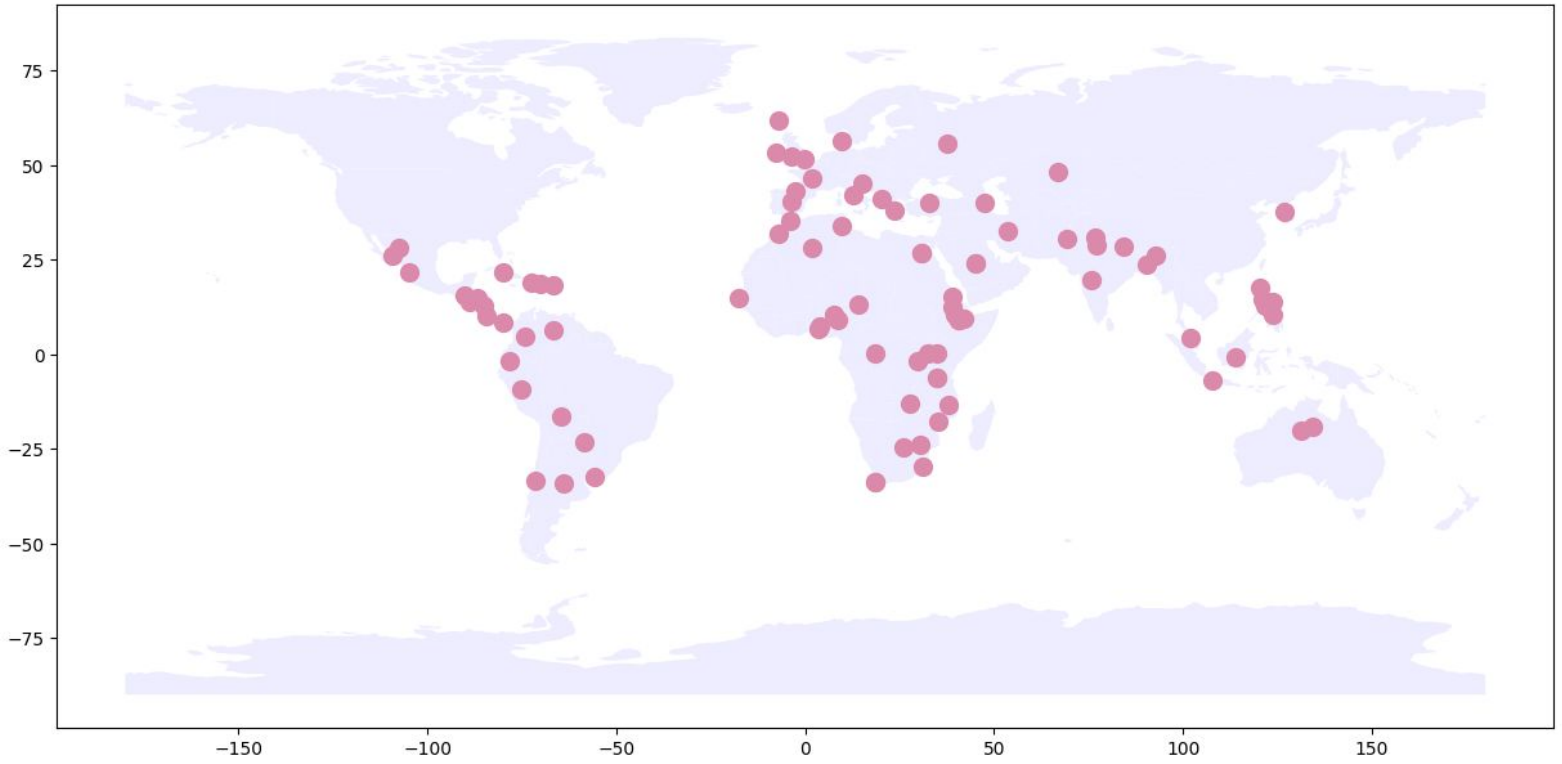
Survey Focus

- Participants work on underserved languages and related projects
  - Affiliation of projects: industry, academia, or both
  - Tasks and languages

- Incentives for participating in particular NLP projects
  - Personal motivations
  - Potential Limitations in NLP artefacts

- Credit attribution practices especially in non-standard settings
  - Involvement in online communities or participatory research
  - Problematic incentivisation

Survey Responses: >70 Mid- to Low-resource languages

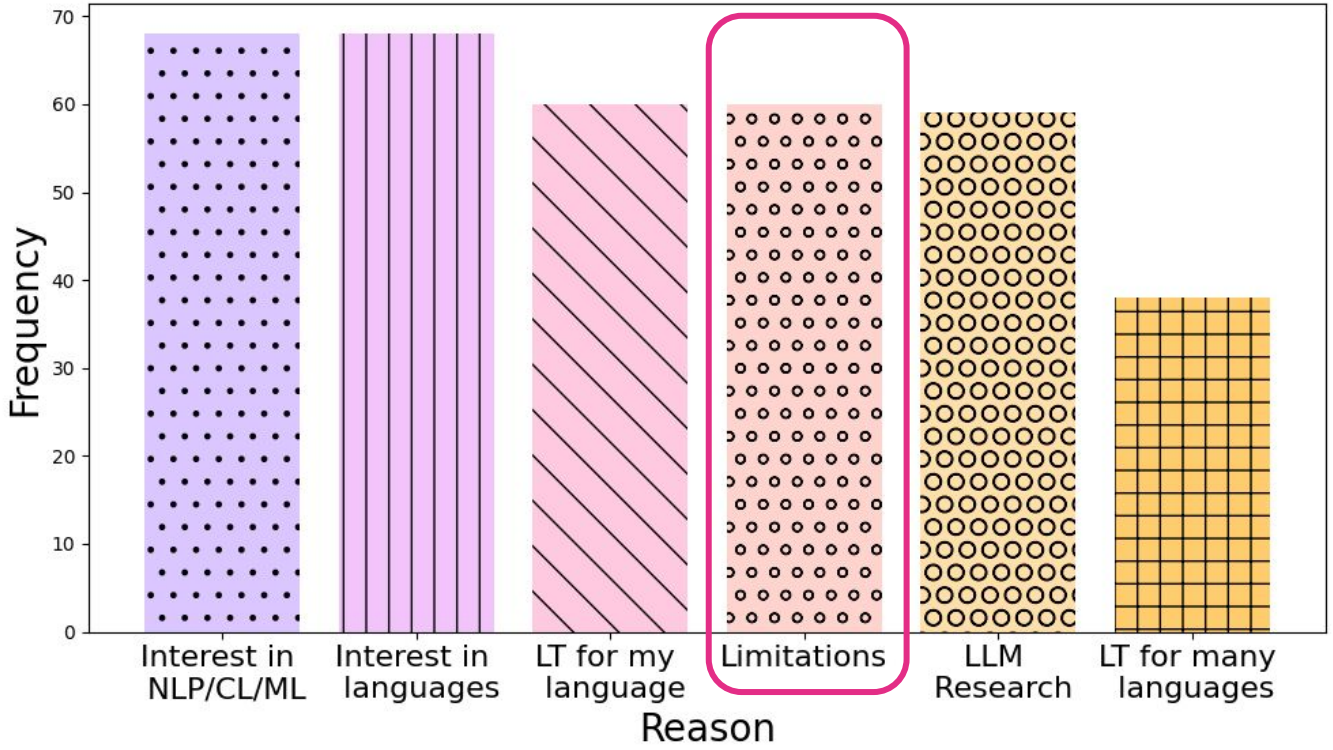


81 Survey Responses

Question	Answer	Percentage
Project Affiliation	Industry	12%
	Academia	57%
	Both	31%
Task Involvement	Data creation	47%
	Data annotation	33%
	Data collection	33%
	Model construction	9%
Motivation / Incentive	Scientific interest	81%
	Building language technologies	72%
	Limitations in language(s) of interest	60%
	LLM research	59%

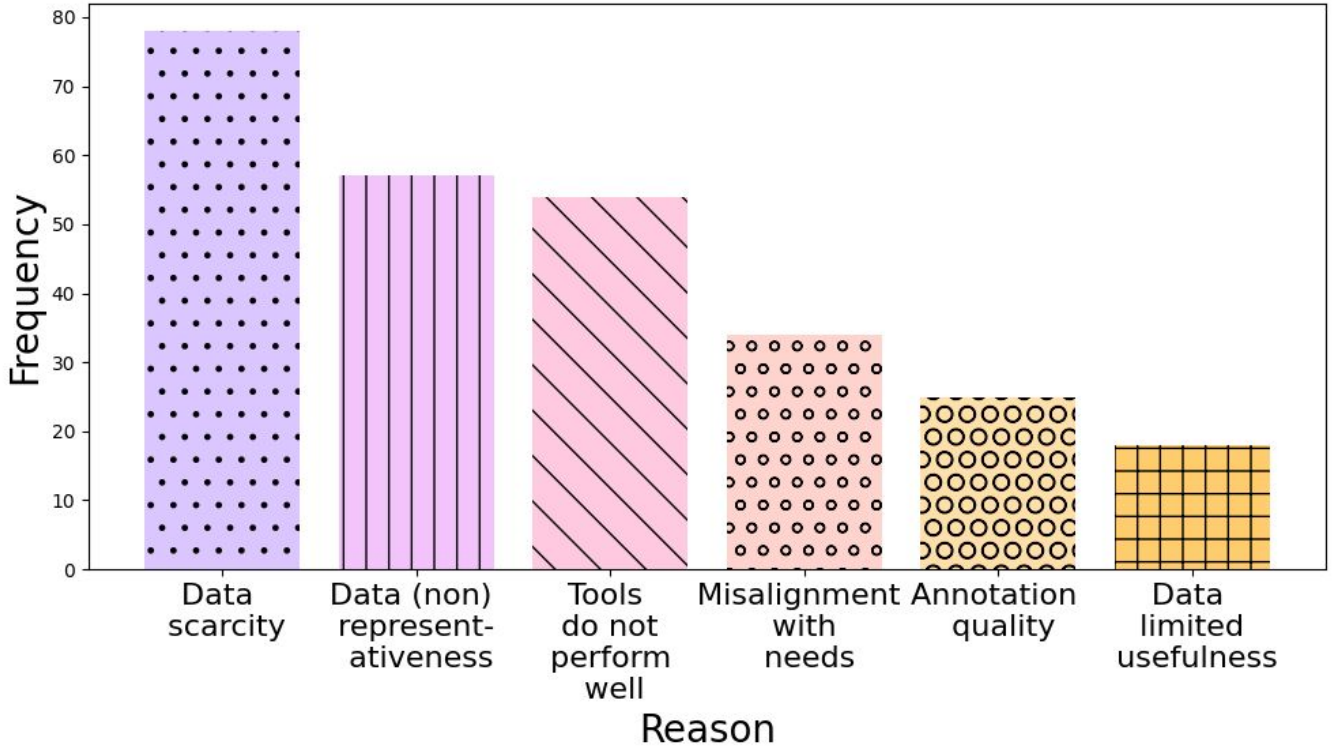
Survey Responses: Incentives

Respondents prioritised building tools/resources for their own language(s) often due to observed limitations



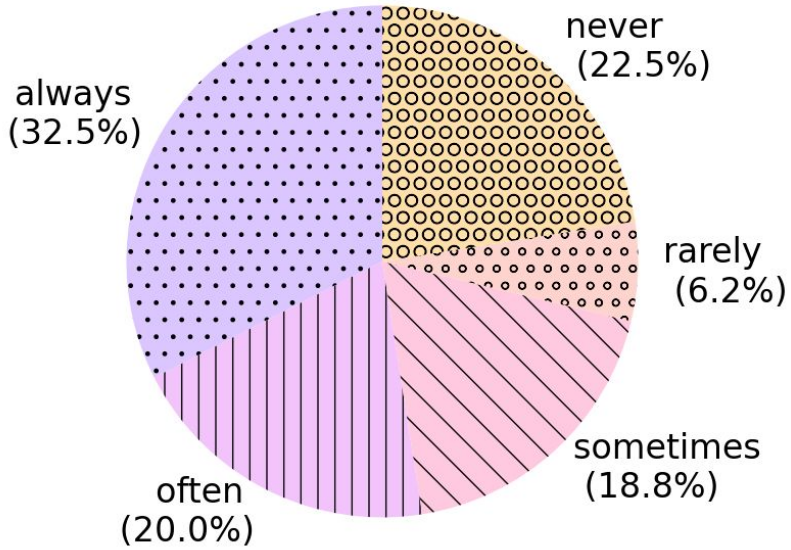
Survey Responses: Reported Limitations

Limitations included data scarcity, poor annotation quality, cultural gaps, and misalignment with community needs



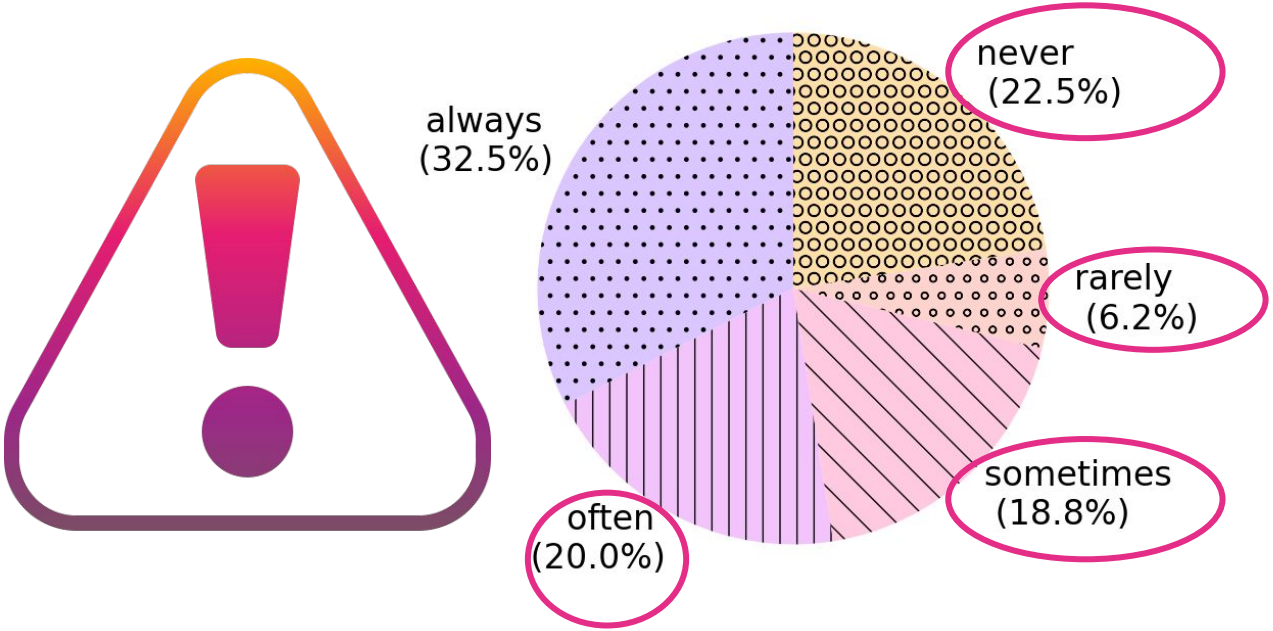
Survey Responses: Credit Attribution

>67% of the respondents reported not being properly credited for their work at least once



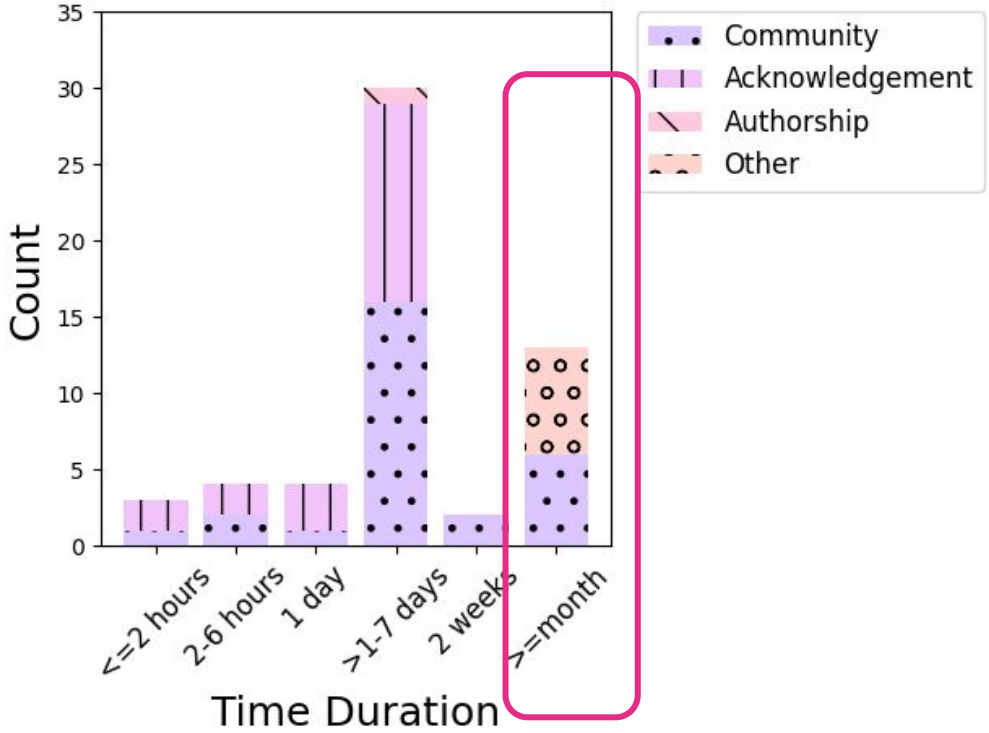
Survey Responses: Credit Attribution

We examined why participants joined these projects and how long the work took



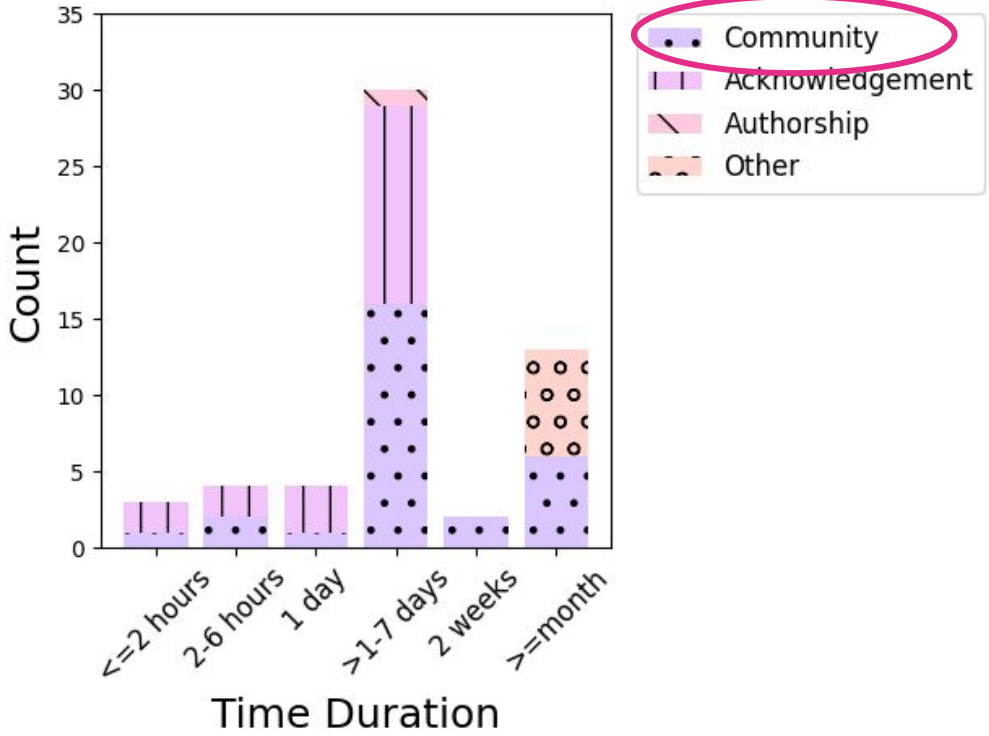
Survey Responses: Credit Attribution

In most cases, the work took over a day to several months

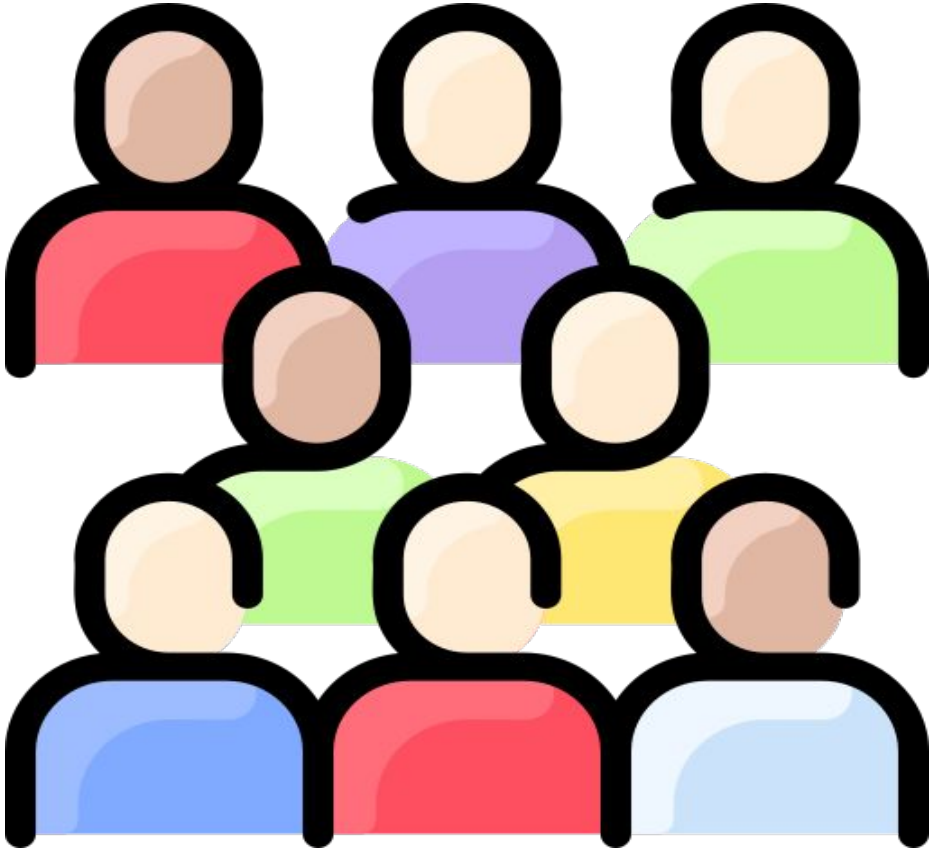


Survey Responses: Credit Attribution

Problematic incentivisation was common including emotional manipulation and misuses of participatory frameworks



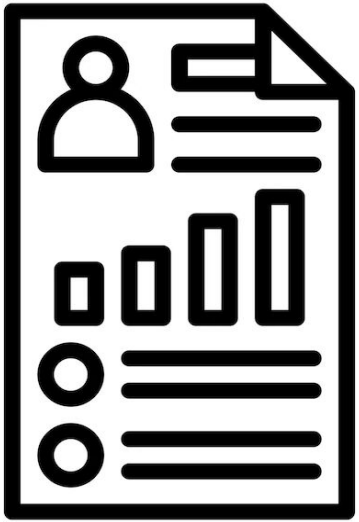
Recommendations : Center the People



Recommendations : Center the People

Speakers

Consider their cultural milieu and who is served



Data Workers

Respect their labour and dignity

Recommendations : Be Fair; Give Credit When Credit is Due

Recommendations : Be Fair; Give Credit When Credit is Due

Fair monetary compensation

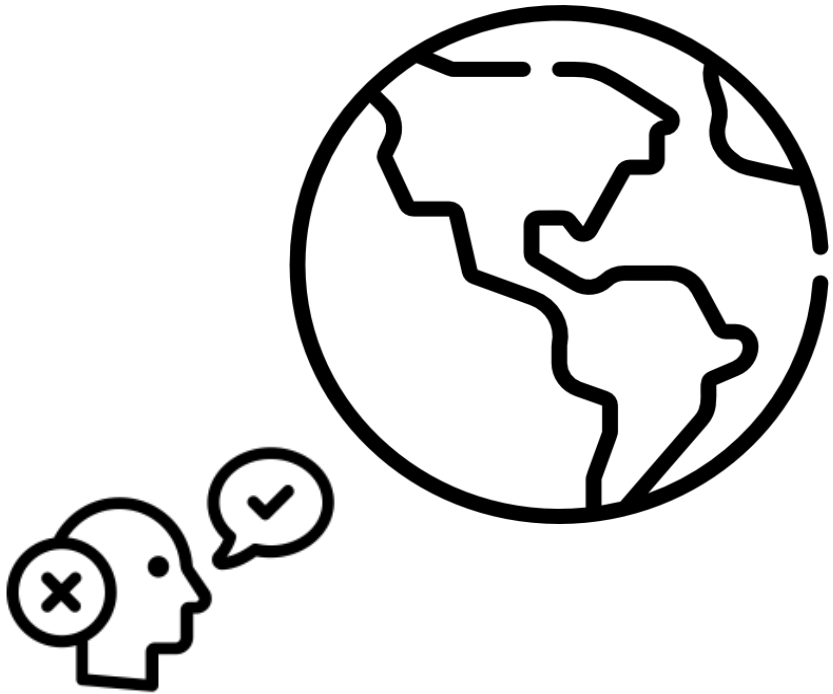
Clear authorship rules

Recommendations : Be Fair; Give Credit When Credit is Due

We provide guidelines for roles that data workers can take to enable co-authorship proper inclusion

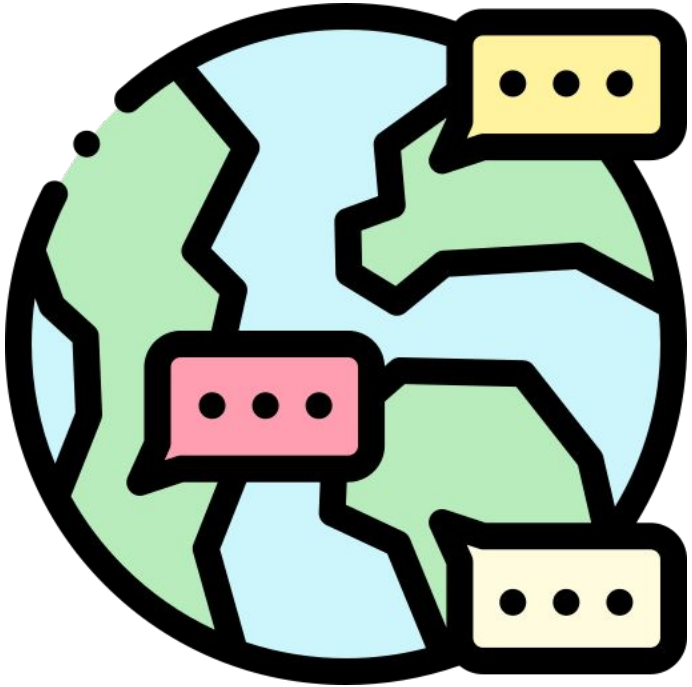
Data Workers as authors?

Recommendations : Avoid False Generalisations



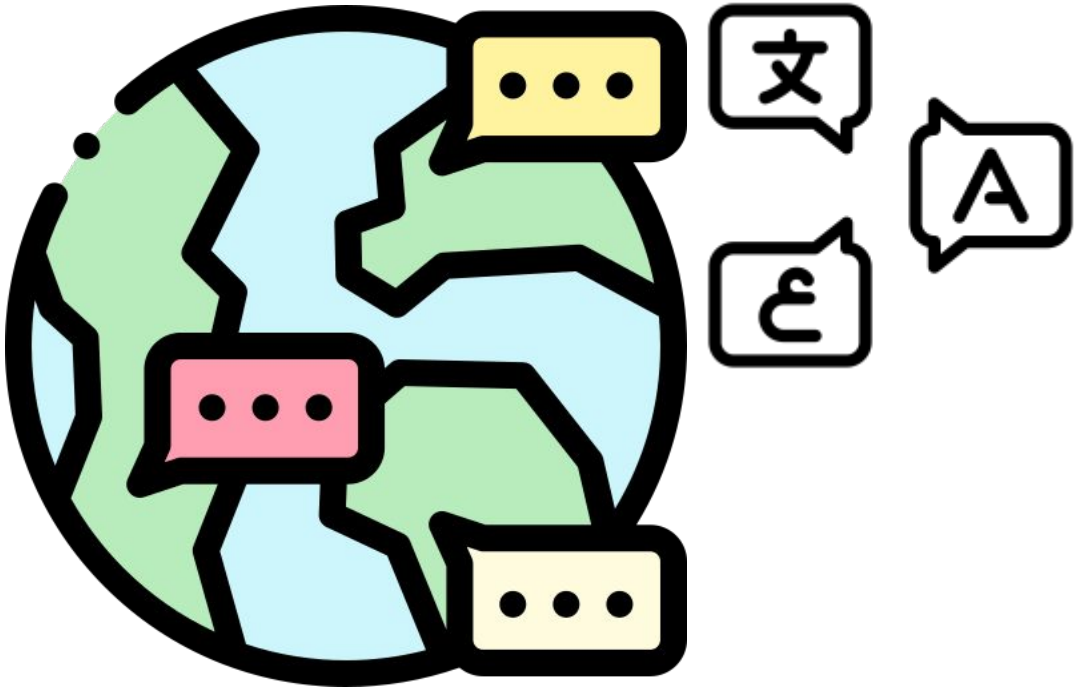
Embrace social awareness and choose country/culture categorisations with care

Recommendations : Set Fair and Realistic Expectations



No prescription on what work should be pursued for underserved languages

Recommendations : Set Fair and Realistic Expectations

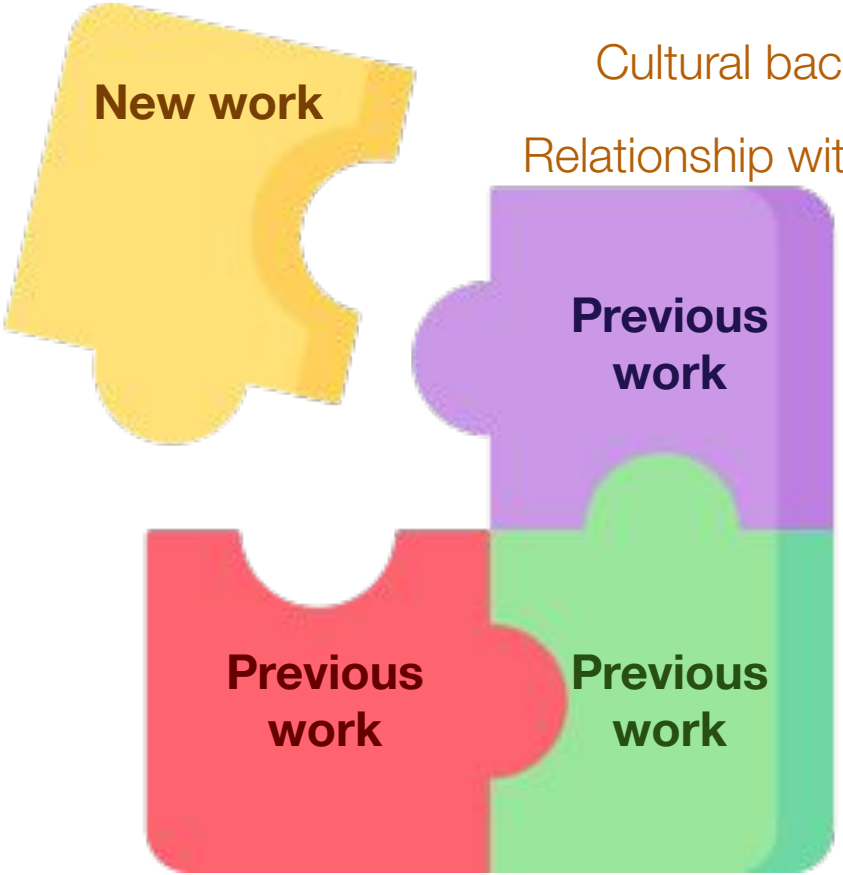


Dealing with a “solved” NLP problem on an underserved language is an actual contribution

Recommendations : Critically assess the source of the data

Even if the language is low-resource

Recommendations : Position Your Contribution



Cultural background

Relationship with the language

…

Many thanks to our survey respondents!

Thank you!

Questions?