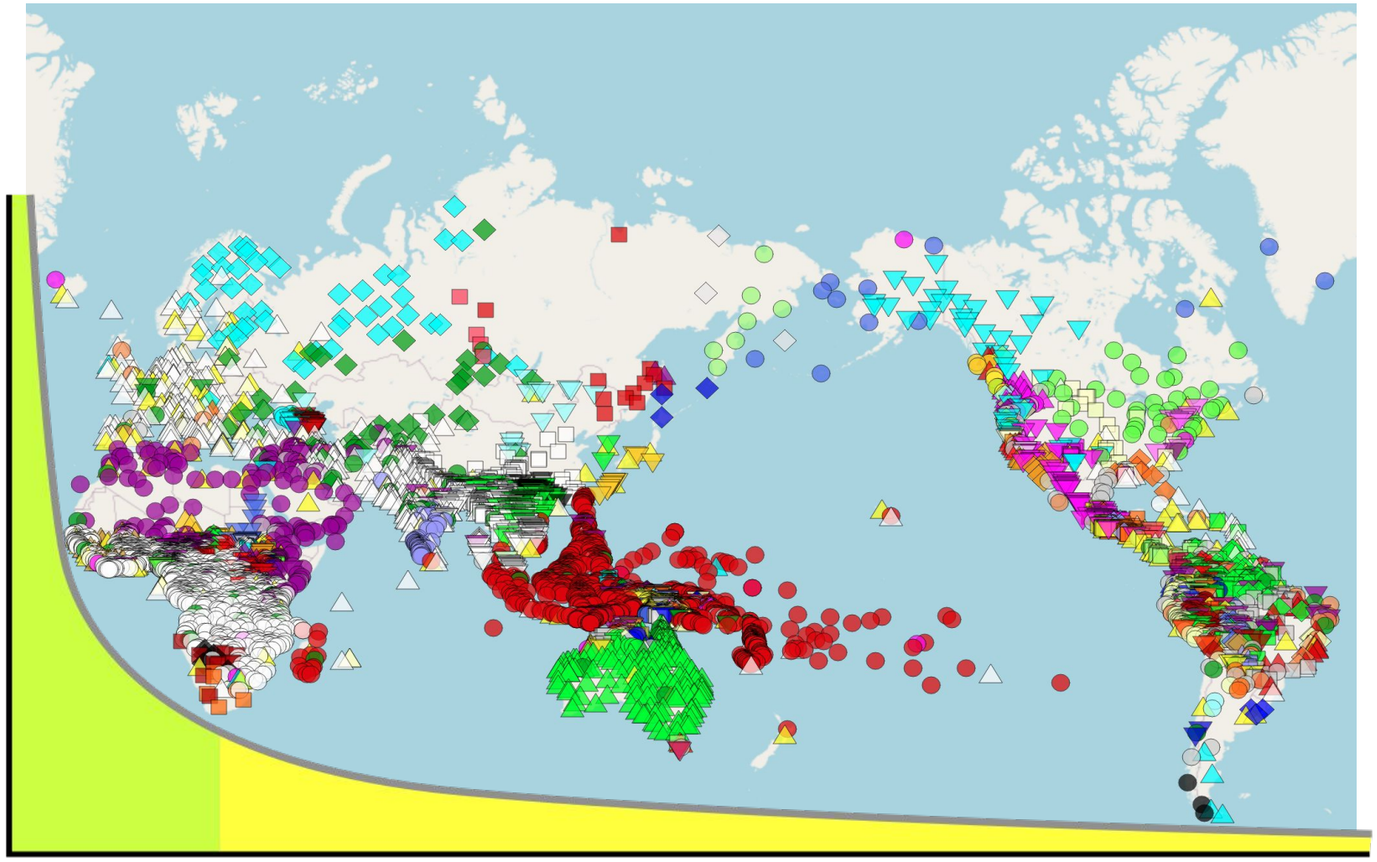
Sec 5.3

Under-resourced Languages

Under-resourced languages

More than 7,000 languages spoken today



Under-resourced languages

What makes a language under-resourced?

- Data availability: labeled data, unlabeled data, quality and representation
- Data domain: coverage and representation
- Noisy and/or opaque orthographies
- Unwritten languages
- Typological coverage:
  - Unique phonetic and phonological systems
  - Dialectal variation
  - Code-switching
  - Representation of non-native speakers

from **SIGUL**, Special Interest Group on Under-Resource Languages

Taxonomy

0. Exceptionally limited resources: pretraining exacerbates situation

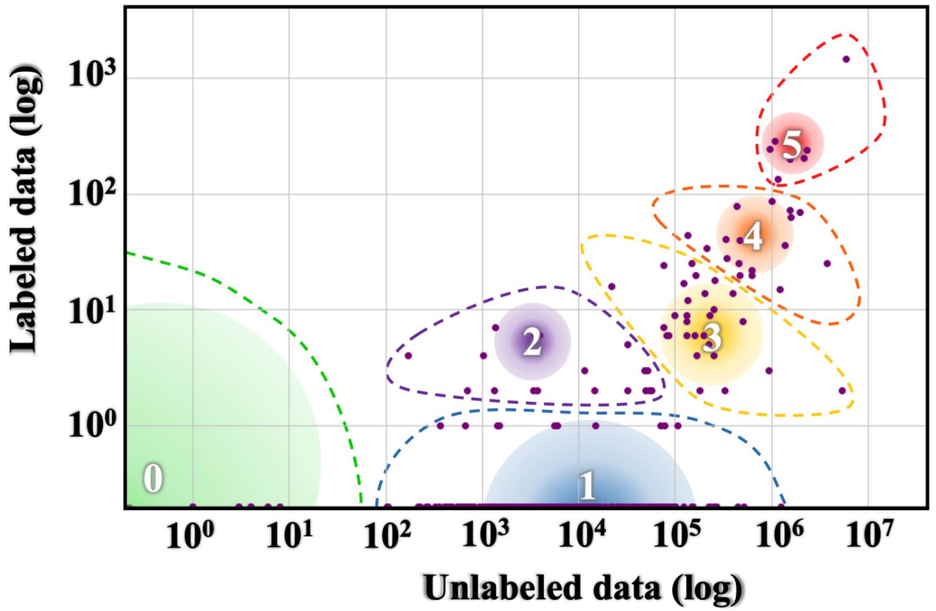
1. Some amount of unlabeled data

2. Small set of labeled data created

3. Unlabeled data enables pretraining, but limited labeled data

4. Large amount of unlabeled data, high quality but limited labeled

5. High-resource languages



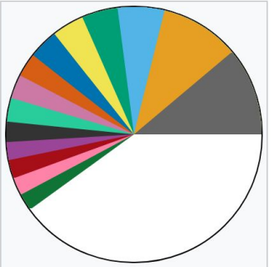
Language resource distribution of Joshi et al. (2020). The size and colour of a circle represent the number of languages and speakers respectively in each category.
Colours (on the VIBGYOR spectrum; **V**iolet–**I**ndigo–**B**lue–**G**reen–**Y**ellow–**O**range–**R**ed) represent the total speaker population size from low (violet) to high (red).

(Joshi et al. 2020)

Languages: Examples

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.0B	88.17%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	1.0B	8.93%
2	Zulu, Konkani, Lao, Maltese, Irish	19	300M	0.76%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.1B	1.13%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	1.6B	0.72%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Number of languages, number of speakers, and percentage of total languages for each language class



Distribution of the 55,991,866 articles in different language editions (as of 9 March 2021);[4] the majority of the articles in Swedish, Cebuano, and Waray were created by Lsjbot.[6]

- English (11.2%)
- Cebuano (9.9%)
- Swedish (6%)
- German (4.5%)
- French (4.1%)
- Dutch (3.7%)
- Russian (3%)
- Italian (3%)
- Spanish (3%)
- Polish (2.6%)
- Waray (2.3%)
- Vietnamese (2.3%)
- Japanese (2.2%)
- Egyptian Arabic (2.2%)
- Other (40%)

0. Dahalo:

Recorded Swadesh list

1. Cherokee:

Bible; 15k sentences parallel text; Tatoeba; Ubuntu

2. Zulu:

Recorded word lists; Tatoeba; Ubuntu

3. Cebuano:

Recorded word lists; BABEL; Bible; Wikipedia; Tatoeba; Ubuntu

4. Korean:

Bible; Wikipedia; OpenSLR 40, 58, 97; Tatoeba; Ubuntu

5. English:

∀

ST: Resources Required

Two steps where resources are required: ① for training and ② for corpus creation

Labeled data:
parallel speech and translations, segmented

Availability:
MuST-C (1); mTEDx (8); CoVoST (21)

Unlabeled data:
monolingual source language speech;
monolingual target language text

Bible (~1000); Wikipedia (285);
linguistic resources often <2 hours

Pronunciation lexicons:
Use: alignment, hybrid ASR models; alternate data representations; CTC loss and/or compression

Hand-created lexicons often unreleased;
Wikipron (117); Epitran (63)

(# source languages)

Pretrained Models

encoder

wav2vec 2.0 — XLSR

decoder

mBART

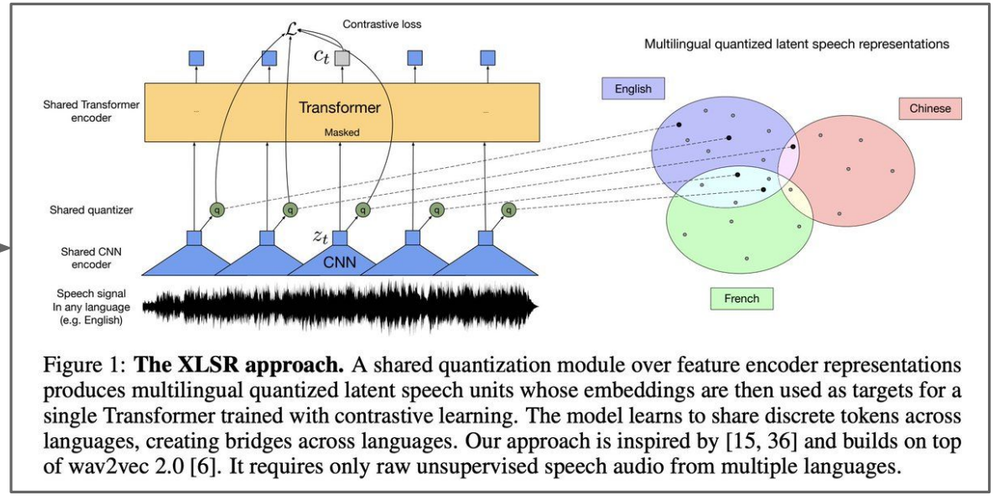


Figure 1: **The XLSR approach.** A shared quantization module over feature encoder representations produces multilingual quantized latent speech units whose embeddings are then used as targets for a single Transformer trained with contrastive learning. The model learns to share discrete tokens across languages, creating bridges across languages. Our approach is inspired by [15, 36] and builds on top of wav2vec 2.0 [6]. It requires only raw unsupervised speech audio from multiple languages.

(Baevski et al. 2020; Liu et al. 2020; Li et al. 2021)

Methods previously discussed:
pretraining + finetuning, knowledge distillation, alternate data representations

Dependences on shared features:
in-vocabulary orthography, phone inventories, use of same model architecture

Unless we assess on under-resourced languages, we will not know how well methods apply!