A Comparative Approach for Auditing Multilingual Phonetic Transcript Archives

Farhan Samir[1,3] **Emily P. Ahn**[2] **Shreya Prakash**[2]
Márton Soskuthy[1] **Vered Shwartz**[1,3] **Jian Zhu**[1]
[1] University of British Columbia, Canada
[2] University of Washington, USA [3] Vector Institute for AI, Canada
fsamir@mail.ubc.ca

Abstract

Curating datasets that span multiple languages is challenging. To make the collection more scalable, researchers often incorporate one or more imperfect classifiers in the process, like language identification models. These models, however, are prone to failure, resulting in some language partitions being unreliable for downstream tasks. We introduce a statistical test, the Preference Proportion Test, for identifying such unreliable partitions. By annotating only 20 samples for a language partition, we are able to identify systematic transcription errors for 10 language partitions in a recent large multilingual transcribed audio archive, X-IPAPack (Zhu et al., 2024). We find that filtering these low-quality partitions out when training models for the downstream task of phonetic transcription brings substantial benefits, most notably a 25.7% relative improvement on transcribing recordings in out-of-distribution languages. Our work contributes an effective method for auditing multilingual audio archives.[1]

1 Introduction

The development of NLP systems has been largely uneven, mostly dominated by a handful of Western European languages and some non-Indo-European ones (Blasi et al., 2022). In some regards, this is improving—only in the last few years, researchers have open-sourced a number of notable multilingual archives,[2] such as MADLAD-400 (Kudugunta et al., 2024), VoxPopuli (Wang et al., 2021), and OWSM (Peng et al., 2024).

The acquisition of these large multilingual archives is complex, however. The data collection pipeline needs to scale to a large number of languages and a large volume of data for each language. Enabling this scaling is the use of predictive models: From language-identification models to sentence-embedding based bitext mining methods (Kreutzer et al., 2022; Koehn et al., 2020) to speaker diarization models (Wang et al., 2021).

These models, however, are prone to failure: Major data collection errors in well-known archives have already been reported. In a comprehensive study over 5 major archives and over 70 languages, Kreutzer et al. (2022) report systematic failures in language identification and bitext mining, with greater error rates for lower-resourced languages. Further, recent studies in multilingual and long-form speech recognition found that prominent speech archives contain substantial chunks of untranscribed content, leading to a high rate of deletion errors in speech recognition models trained on this data (Tian et al., 2024; Fox et al., 2024).

These studies demonstrate the complexity of acquiring high-quality multilingual data. In this light, the data collection pipeline itself can be considered an imperfect approximation of the data distribution we wish to sample from. As with any approximation, common wisdom suggests that we should aim to evaluate the quality of the approximation. Unlike the wealth of empirically and theoretically established metrics and hypothesis tests for comparing two models (Dror et al., 2018), there is a remarkable dearth of methods for evaluating the reliability of a semi-automatically scraped dataset that may serve as ‘‘gold-standard’’ for future downstream applications.

In this work, we ask: How can we efficiently identify languages or dialects where the data-collection pipeline may have failed systematically? At a high level, we want to analyze a small

[1] https://github.com/smfsamir/audit-ipa.

[2] We refer to the entire multilingual dataset as an archive, and each language variety's subset as a partition. See Section 2 for discussion of this terminology.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 595–612, 2025. https://doi.org/10.1162/tacl_a_00759
Action Editor: Jonathan Clark. Submission batch: 10/2024; Revision batch: 1/2025; Published 6/2025.
© 2025 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

subset of the partition for a language from a multilingual archive, and determine whether its samples are high quality. This analysis has two important components. First, how do we determine if a sample is high quality? Second, how large should the subset be to make our determination?

To answer these questions, we draw on theory from statistical power analysis (Cohen, 1992; Card et al., 2020). Specifically, we are interested in accurately estimating *preference* (Bradley and Terry, 1952): How much better does the partition capture the relationship of interest than an existing baseline model? In particular, we elicit preferences from human annotators to ground the partition's quality with judgments from domain experts.

We ground our method in the task of phonetic transcription, where the input is recorded speech from any language, while the output is a transcription into the International Phonetic Alphabet (IPA). Phone recognition models have important applications in language documentation, especially for oral languages (Bird, 2021; Lane and Bird, 2021). The training datasets for this task are often semi-automatically generated, so that many languages can be represented (Section 2).

Consider the partition for one language in such a semi-automatically populated multilingual archive. If a knowledgeable user consistently prefers the output of an existing (imperfect) baseline model over the ''ground-truth'' transcripts in the partition under audit (see Figure 1), then this indicates that this language partition is unreliable. We refer to this statistical test as the Preference Proportion Test, or the PPT, which we introduce in Section 3.2. Critically, we assert that only a small fraction of examples needs to be annotated to attain a high-powered test of whether a language partition should be flagged as unreliable.

To illustrate the effectiveness of the PPT, we perform a case study on a recent large multilingual phonetic transcript archive—the X-IPAPack (Zhu et al., 2024), comprising transcribed audio for 78 languages. Applying the PPT, we efficiently identify 10 language partitions in the archive that have unreliable transcripts. We find that a model finetuned on the filtered version of the archive—without the unreliable 10 language subsets—generalizes *better* to a test set (comprising 5 held out languages) than a model trained on the complete archive.

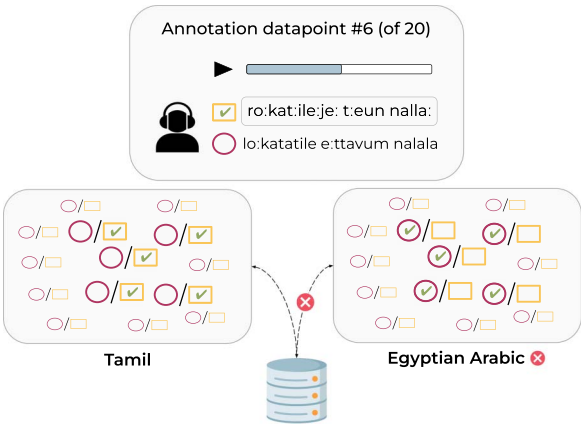

Figure 1: Multilingual archives have been reported to contain a high degree of quality-control issues, especially for lower-resourced languages. We audit the quality of each language partition in a multilingual archive by annotating a small sample from it. Specifically, expert annotators select between the gold-standard transcript and one generated by a baseline model prediction. When the baseline model predictions are consistently preferred over the gold-standard, we conclude that the language partition should be flagged for re-labeling.

We find two ways in which low-quality data can be especially pernicious. First, we find the largest improvement on the Punjabi partition of our held-out evaluation dataset (an error reduction of 20.3%) possibly due to omitting the unreliable transcripts from the Sindhi partition, suggesting that the effects of poor data quality can tamper with performance in related language varieties. Second, we also find a 25.4% improvement on out-of-distribution languages after training on the PPT-filtered archive, suggesting that low-quality data has a considerable impact on lower-resourced languages. Our empirical results add nuance to the purported benefits of data-scaling (Hoffmann et al., 2022, for example).

Finally, we emphasize that filtering out low-quality data, while highly effective, is not a panacea for building robust multilingual models. Some works have suggested that high-quality data on a small number of language varieties is sufficient to obtain a ''universal'' or ''language-agnostic'' model (Taguchi et al., 2023; Li et al., 2020). Our empirical results in phonetic transcription, however, do not support this position. Leveraging a phone segment-level error metric, we find that existing universal phonetic transcription models are instead highly attuned to sounds that are more common in their respective training datasets,

while making more errors on unfamiliar sounds. Overall, this suggests that more diverse and high-quality data collection is required for equitable performance across languages and their varieties.

2 Related Work

Desai et al. (2024) analyze large-scale NLP datasets through the lens of traditional archival studies. They argue that NLP datasets can be interpreted as power-laden (informal) archives. These archives not only afford the capabilities of future generative models, but—in the case of multilingual archives—also mediate our understanding of different cultures. Thus, appraising the quality of different language partitions of these multilingual archives is highly important.

However, appraising the quality of NLP archives has proven challenging due to their immense scale. Researchers have taken to algorithmic filtering methods (Desai et al., 2024, Table 2), but a number of studies have argued that these methods reproduce if not exacerbate existing societal inequalities. Dodge et al. (2021) found that the C4 corpus derived from CommonCrawl applied algorithmic filters that excluded different dialectal varieties of English. Recently, Hong et al. (2024) found that CLIP-filtering removes data pertaining to LGBTQ identities and non-Western regions at higher rates.

When it comes to multilingual audio archives, we can consider the Whisper model's training data (Radford et al., 2023). In compiling their training dataset, they report filtering out audio-transcript pairs where the transcripts were entirely upper-cased. In doing so, the authors sought to remove transcripts generated by speech recognition models. However, many languages do not encode case. This filtering rule thus could only serve as quality assurance for some language partitions, in particular those that employ the Roman alphabet. Such filtering rules are a product of a researchers' language knowledge and proficiency, which is known to be skewed towards Western institutions—specifically, the languages spoken and studied there (Held et al., 2023). There is thus a systemically greater risk for language varieties from certain regions to contain unrepresentative data that was generated by a speech recognition model. This contributes another means by which data processing pipelines recapitulate societal and structural inequities (Hong et al., 2024; Bender et al., 2021; Dodge et al., 2021; Desai et al., 2024; Benjamin, 2019).

By contrast, in a comprehensive study on several prominent multilingual text archives, Kreutzer et al. (2022) demonstrate the importance of intentional manual appraisal. Through a collaborative effort in annotating multiple language partitions in these archives, they found that low-resource language partitions in archives tended to contain high rates of text that was from a different language, or non-linguistic content altogether. The low-resourced status of a language is influenced by the sociopolitical history that defined its marginalization (Nigatu et al., 2024). Its poor representation in a multilingual archive thus unfortunately reproduces inequity in a digital format.

In accordance with Kreutzer et al. (2022), we thus advocate that researchers intentionally (rather than algorithmically) appraise the quality of individual language partitions before leveraging a multilingual archive for a downstream task. We describe a novel comparative approach for this appraisal in the next section, through a case study on a recent multilingual archive. We look at the X-IPAPack archive (Zhu et al., 2024), comprising phonetically transcribed audio for a number of language varieties. Critically, we argue that our comparative appraisal procedure can be completed on a small fraction of the samples in the language partition, making the appraisal process both feasible and effective.

Although Zhu et al. (2024) reported having manually appraised each language partition by inspecting 10 of its audio samples, our comparative procedure identified subtle yet systematic transcription errors in several partitions (Section 3) that were previously overlooked. Further, we find that removing these partitions yields major improvements in the downstream task of automatic phone recognition (Section 4).

3 Case Study: Auditing X-IPAPack

In this section, we introduce the Preference Proportion Test for efficiently auditing the quality of a multilingual archive. We ground our explanation in a case study of the X-IPAPack archive. We first describe the X-IPAPack archive and the preprocessing of the text in the phonetic transcripts (Section 3.1). Then, we introduce the test (Section 3.2) and apply it to the X-IPAPack archive (Section 3.3).

Category	Ex.	#{Type}	#{Tokens}
Valid primary	p	107	21.8M
Valid one diacritic	v^j	282	2.1M
Valid two diacritics	k^{hj}	67	62.5K
Unlikely / invalid	$t^{ʕʕ}$	330	758K

Table 1: Documenting the frequency of phones in X-IPAPack-Fleurs. A phone is considered valid if it is contained in the `panphon` database (Mortensen et al., 2016).

3.1 X-IPAPack Contents

Overview. The archive comprises phonetically transcribed speech for 77 languages. The recordings and orthographic transcripts were provided by Fleurs (Conneau et al., 2023), and the conversion of the orthographic transcripts to phonetic ones was done by Zhu et al. (2024).[3] Each language partition contains at least 3 hours of recordings ($M = 10.12$H, $SD = 2.74$H). Individual recordings are at most 30 seconds ($M = 12.14$s, $SD = 1.68s$). We next document and preprocess the contents of the transcripts that are paired with the recordings.

Transcript Composition. We segment each phonetic transcript into individual phones, using the `lingpy` tokenizer (List and Forkel, 2016). We count the occurrence of each phone in each language and categorize the phones into the taxonomy in Table 1.

We find that the majority of the tokens (97.0%) are valid ones. The majority of these specify a primary place and manner of articulation. Moreover, there are a number of phones that have one or two diacritics, for example v^j, indicating palatalization. Overall, the archive contains a high degree of phonetic diversity.

Transcript Normalization. There is, however, a long tail of 330 unrecognized phonetic strings (according to the `panphon` database) out of a total 786 phone types. This represents 3.0% of the total tokens in the corpus (approx. $758\text{K}/25\text{M}$). Some of these are invalid unicode representations (e.g., ASCII g is different from the IPA velar plosive ɡ; 31,525 occurrences). Other times, there are repetitions of diacritics, e.g., $t^{ʕʕ}$ (129 occurrences). Another common error is non-standard diacritics, e.g., $o^{ŋ}$ (instead of oŋ; 264 occurrences). We correct these invalid phones manually. After our vocabulary cleaning, we arrive at an archive with 473 unique, valid phones. We also document the mapping from an invalid phone to a valid phone.[4]

The occurrence of invalid or implausible phones in the archive may be an artifact of the G2P models applied for converting the orthographic transcripts in Fleurs to the phonetic transcripts in X-IPAPack. For example, one of the Grapheme-to-Phoneme (G2P) models, CharsiuG2P (Zhu et al., 2022), was a byte-level neural model with no constraints mandating that only valid phones be predicted.

3.2 Preference Proportion Test (PPT)

We now introduce the statistical test for assessing whether a language partition in X-IPAPack is of reliable quality. Specifically, we ask whether the transcripts are reasonably descriptive of the recordings in the partition. The phonetic transcripts were automatically generated using G2P models from the orthographic transcripts and thus there is a distinct possibility that the phonetic transcripts do not closely reflect the recording. As seen in prior work on multilingual G2P conversion, performance is far from uniform across language varieties (e.g., the error rate for Egyptian Arabic is more than quadruple that of Spanish in some models; Zhu et al., 2022).

Problem Setup. Formally, we have an archive $\mathcal{D}$ with L partitions, one for each language in X-IPAPack – $\mathcal{D}_1, \ldots, \mathcal{D}_L$. The archive D comprises pairs $(x, G(t))$, where x is an audio recording while $G(t)$ is a phonetic transcript generated by applying a G2P model to an orthographic transcript t. However, some of the partitions may be corrupted from systematic G2P conversion errors, making it unreliable for downstream tasks where a tight correspondence between audio (x) and phonetic transcript ($G(t)$) is important. We would like to efficiently identify highly unreliable partitions,

[3] The X-IPAPack builds on data from three prior datasets: MSWC (Mazumder et al., 2021), DoReCo (Paschen et al., 2020), and Fleurs (Conneau et al., 2023). We focus on auditing the Fleurs portion since it is the largest in terms of the number of languages.

[4] The mapping is available here. We note that there are rare cases of some phonemes such as z^h that are attested (Jacques, 2011) (though rare), but considered invalid in `panphon`. However, such occurrences are rare in both type and token frequency. For example, z^h only occurs 6 times.

annotating only a small sample of $\mathcal{S}_i \subset \mathcal{D}_i$, where $|\mathcal{S}_i| << |\mathcal{D}_i|$. We first describe the setup of the annotation for each datapoint $(x, G(t))$, followed by the construction of $\mathcal{S}_i$.

Annotating the Quality of a Transcript. Directly annotating the quality of a phonetic transcript $G(t)$ in its correspondence with the audio x is challenging, as there is no reference baseline for what makes a transcript high quality. Instead, we turn the task to one of pairwise comparison, by asking an annotator to choose between two transcripts: the ground-truth ($G(t)$) or one generated by a reasonably-good quality phone-recognition model ($\mathcal{M}(x)$), like that of Xu et al. (2022) or Taguchi et al. (2023). Eliciting preferences through comparisons rather than absolute judgments has been championed in other work, most notably in Reinforcement Learning from Human Feedback, where others have commended the strategy for providing consistent choices (Christiano et al., 2017). In Figure 1, the transcripts deviate from one another—for example, the top transcript begins with a rhotic while the bottom begins with a lateral—and the annotator can listen to the recording to determine which transcript is more faithful.

Hypothesis Testing. In order to efficiently determine whether $\mathcal{D}_i$ is an unreliable partition of X-IPAPack, we annotate only a random sample $\mathcal{S}_i$. Intuitively, if the annotator consistently prefers the model-generated transcript over the X-IPAPack ground-truth version, ($\mathcal{M}(x) \succ G(t)$), then we may want to discard $\mathcal{D}_i$ from applications in downstream tasks until the transcripts are improved.

Although $G(t)$ is considered the gold-standard transcript in X-IPAPack, both $G(t)$ and $\mathcal{M}(x)$ are essentially (error-prone) predictions for the phonetic transcript of x. We can thus assess the reliability of $G(t)$ by performing a model comparison hypothesis test (Card et al., 2020) between the two approximations. Specifically, the **null hypothesis** is that the annotator has no preference for the ground-truth transcript $G(t)$ over the model-generated one $M(x)$, and the **alternative hypothesis** is that $G(t)$ is significantly unfavorable. We can model the degree to which the annotator prefers the ground-truth X-IPAPack transcripts as θ_G, where $0 \leq \theta_G \leq 1$. When $\theta_G << 0.5$, we can conclude that $\mathcal{D}_i$ is an unreliable language partition.[5] We refer to this as the Preference Proportion Test, or the PPT.

```
1 from scipy.stats import binom
2 def ppt_sample_size(alpha, theta_null,
    theta_alt):
3     for n in range (5, 100, 5): # num.
    trials
4         k = binom(n, theta_null)
5             .ppf(alpha) - 1
6         power = binom(n, theta_alt)
7             .cdf(k)
8         print(f"Number of samples: {n};"
9             f"Critical value: {k};"
```

Listing 1: Prints the critical value (k) and statistical power of a hypothesis test under a hypothetical effect size (`theta_alt` vs. `theta_null`) and tolerance for false positives (α), for varying numbers of trials (n).

For setting the sample size of the subset to annotate $\mathcal{S}_i$, we perform a power analysis (Cohen, 1992). We provide the code for the analysis in Listing 1. The power analysis is a function of three arguments. First is the tolerance for false positives α. Next is the effect size, which is defined by two parameters: the difference between the preference ratio under the null hypothesis (θ^0_G) and the alternative hypothesis (θ^A_G). We can then supply all three parameters, for example, `ppt_sample_size(0.05, 0.5, 0.2)`. We can then test a wide range of potential sample sizes (`for` loop on line 4). For each sample size, we can determine the critical value (k; line 5), under which the null hypothesis (equal preference for ground-truth or model-generated) would be rejected. We must also compute the power of this test (line 6), which tells us how likely this outcome is under the alternative hypothesis (model-generated is preferable).

As illustrated in Figure 2, our statistical power increases with the effect size ($\theta^0_G - \theta^A_G$) or the number of samples. When we suspect some

[5] We choose the null hypothesis to be no preference ($\theta_G = 0.5$) because $M(x)$ can, for certain languages, serve as a very strong baseline. For example, as we show empirically in Section 4, Xu et al. (2022) and Taguchi et al. (2023) train very strong models for English and Japanese, respectively. Thus, it is plausible that the annotator would be unbiased with respect to $G(t)$ compared to $M(x)$, even if $G(t)$ was reliable for these two languages.

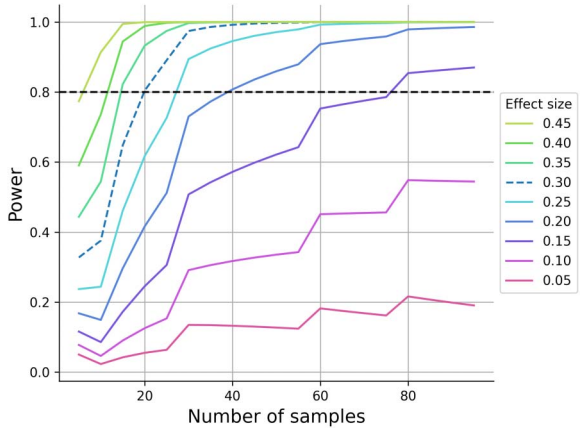

Figure 2: Number of samples that have to be annotated for the Preference Proportion Test (PPT; Section 3.2) for different effect sizes, assuming a fixed false positive tolerance of $\alpha = 0.05$. We aim to achieve at least 80% statistical power. We use an effect size of 0.3 for auditing the X-IPAPack archive; see Section 3.3.

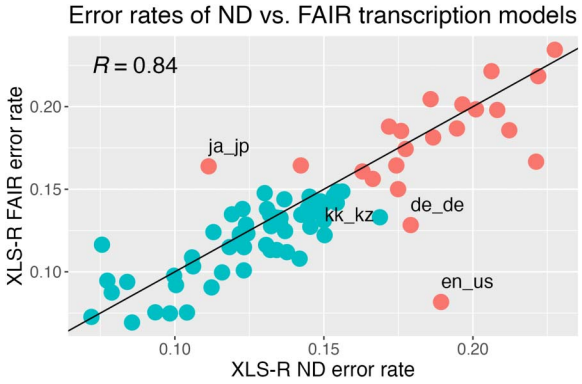

Figure 3: Benchmarking the models of Taguchi et al. (2023) (x-axis) and Xu et al. (2022) (y-axis) on X-IPAPack. Error rates for both models are measured by phonetic feature discrepancies (Mortensen et al., 2016) between the model-generated transcripts and the transcripts in X-IPAPack. In Section 3.2, we audit the languages with relatively high-error rates (top-right quadrant, in red), to determine whether the errors may be attributed to poor-quality transcripts in X-IPAPack.

language partitions to have considerable quality impairments, we can use a large effect size (small θ_G^A) and detect such partitions by annotating only a few samples.

3.3 Applying the PPT to X-IPAPack

We now aim to identify any unreliable language partitions D_i of the X-IPAPack archive with the Preference Proportion Test (PPT). We describe the selection of the language partitions to audit, the number of samples to annotate in each partition, and how we format the datapoints to elicit annotations.

Selecting Languages to Audit. We first select some language partitions in X-IPAPack to audit. To do so, we apply leverage existing phone recognizers, and select languages where the recognizer predictions have a high rate of discrepancies compared to the gold-standard transcripts. To compute the discrepancy, we follow Taguchi et al. (2023) and use the Phonetic Feature Error Rate (PFER) with phonetic feature vectors from `panphon` (Mortensen et al., 2016). We normalize the error rates by the length of the X-IPAPack transcript in terms of number of phones, ensuring that the variation in error rates is not an artifact of length. In Figure 3, we plot the error rates for all languages in X-IPAPack using the phone recognizers of Taguchi et al. (2023) and Xu et al. (2022), which we denote XLS-R ND and XLS-R FAIR, respectively. We use these recognizers as they demonstrated competitive performance against human transcribers in recent work (Taguchi et al., 2023).

We find that there is a substantial amount of variation in performance across languages for both models. In the top right of Figure 3, language datapoints coded in red (e.g., Malayalam; length-normalized PFER: 0.20) achieve error-rates more than double that of languages on the bottom left (e.g., Swahili; length-normalized PFER 0.08). As shown in Figure 3, this variation is mostly robust to the choice of the recognition model, which have a correlation of $r = .84$. We thus select all languages with an error rate in the third quantile of either model (.15 for Xu et al. (2022) and .17 for Taguchi et al. (2023)) for annotation with the PPT, $L = 22$ languages in total. For each language, we generate transcripts using the model that performs better for that language, for example Xu et al. (2022) for English and Taguchi et al. (2023) for Japanese.

Setting Parameters for PPT ($\mathbf{n = 20}$; $\mathbf{k = 5}$). We are auditing for D_i where there are systematic discrepancies between the audio x and the ground-truth transcript $G(t)$, so the value of θ_G^A for the alternative hypothesis should be much lower than the null hypothesis θ_G^0 value of 0.5. We thus set $\theta_G^A = 0.2$ for the alternative hypothesis, an effect size of $\theta_G^0 - \theta_G^A = 0.3$.[6] With a

[6]Since this is a one-sided hypothesis test, our test will detect $\theta_G \leq 0.2$.

false positive error tolerance of $\alpha = 0.05$, we find that we can achieve a test with a statistical power of 80.4% through annotating $n = 20$ samples (Figure 2), which is considered a high-powered test (Card et al., 2020). Specifically, we draw $n = 20$ random samples from the partition (without replacement); then we annotate the samples and reject the null hypothesis when only $k = 5$ times or fewer do we prefer the X-IPAPack transcript $G(t)$ over the model-generated one $M(x)$.[7]

Aligning Gold-standard and Baseline Transcripts. Unlike an X-IPAPack transcript $G(t)$ that contains space-delimited phone strings, the corresponding phonetic language model transcript $\mathcal{M}(x)$ is a phone string with no spaces. In order to facilitate the comparison of the two transcripts, we induce spaces in the model-generated transcript using the Needleman-Wunsch alignment algorithm (Needleman and Wunsch, 1970; Kleinberg and Tardos, 2006) to align the model transcript with the transcript.[8] Since the algorithm requires computing the similarity between pairs of phones, we employ articulatory feature vectors from `panphon`. That is, we encode phones as binary feature vectors – for example, whether the phone is voiced or unvoiced) – enabling a graded measure of similarity by computing the hamming distance between the vectors (Mortensen et al., 2016; Taguchi et al., 2023).

Annotation Process. For every datapoint, the annotator selects between $G(t)$ and $M(x)$ as the preferable transcript for an audio recording from X-IPAPack. We demonstrate the annotation interface and instructions in Appendix B. They could replay the recording as many times as desired before making their choice, and could configure the recording to play at normal speed or at $0.25/0.50/0.75$ speed. Both $G(t)$ and $M(x)$ reflected a broad or phonemic transcription style, as they are either the result of G2P conversion (in the case of $G(t)$) or were trained on G2P transcripts (in the case of $M(x)$). The annotator completed all 20 samples for a language partition S_i from X-IPAPack before moving onto the next annotation subset S_j ($j \neq i$).

[7] It would also be plausible to select a slightly smaller or slightly larger θ_G for the alternative hypothesis—this would only require slightly fewer or more samples to be annotated, respectively. The overall point is that for detecting large quality dropoffs for $G(t)$ (relative to $M(x)$), only $|S_i|$ rather than $|D_i|$ samples need to be annotated.

[8] We could also remove spaces from both transcripts, but this makes the transcripts very difficult to read.

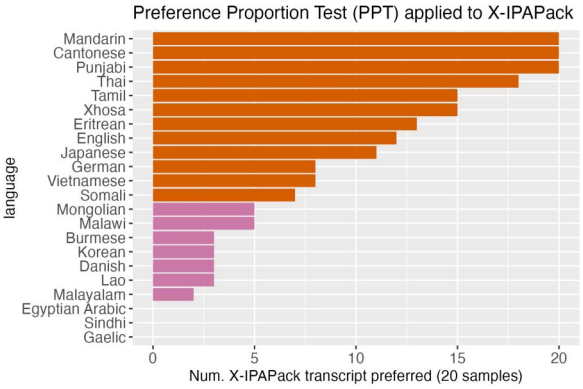

Figure 4: Number of samples where the X-IPAPack transcript was preferred to one generated by a phone recognizer. The bottom 10 languages have $\leq k$ samples (where $\mathbf{k = 5}$) where the X-IPAPack transcript is preferred over the phone recognizer transcript, and thus they fail the Preference Proportion Test.

Language	$G(t)$	$\mathcal{M}(x)$
Egyptian Arabic	taʃtahir	teʃteher
Malayalam	ʧaːrinte	ʧaːrɪɳḍi
English	tuflaɪ	təflaɪɪ

Table 2: Examples of cases where the model-generated transcript $\mathcal{M}(x)$ is preferable to the ground-truth transcript $G(t)$ from X-IPAPack. Note that transcripts and predictions are generally longer (full sentences); see Figure 1 for a realistic example of the annotation task.

Identifying Unreliable Language Partitions. Figure 4 shows that out of the $L = 22$ languages, we reject the null hypothesis for 10, indicating that the X-IPAPack transcript for these languages is unreliable. We demonstrate examples where we prefer the model in Table 2. For example, in Egyptian Arabic ($\theta_G = 0/20$), we find that the vowels are systematically mis-transcribed, often substituting open vowels (a) for close-mid ones (e). In Malayalam ($\theta_G = 2/20$), we find that the transcripts regularly mistake retroflex consonants for alveolar ones, in addition to misidentifying voicing (t instead of d). We also find that the model-generated transcript $\mathcal{M}(x)$ may be preferable over the ground-truth transcript even when

the ground-truth transcript $G(t)$ reasonably represents the audio recording. For example, for American English ($\theta_G = 12/20$), while the X-IPAPack transcripts are generally reliable, $\mathcal{M}(x)$ often provides a more faithful transcription since it can identify cases of co-articulation or reduction (ə instead of e).

Agreement on the PPT Annotations. In order to assess the agreement that the flagged languages were indeed of low quality, we had a second co-author independently annotate 50 datapoints. Specifically, they annotated 10 datapoints for 5 of the flagged languages: Egyptian Arabic, Danish, Gaelic, Malayalam, and Sindhi. We find that the annotators largely agree, providing the same preference on $43/50$ datapoints.

Summary. We demonstrate that annotation of 20 samples per language can enable identifying unreliable language partitions in multilingual datasets. Using our PPT procedure, we efficiently identify a number of languages with low quality transcripts (Malayalam, Egyptian Arabic, among others) in the X-IPAPack dataset. We thus recommend that these partitions be omitted from use until their phonetic transcripts are remediated.

4 Data Quality Effect on Downstream Performance

We now assess the downstream effect of removing the $L' = 10$ unreliable language partitions $\mathcal{D}_1, \ldots, \mathcal{D}_{L'}$ from the X-IPAPack archive $\mathcal{D}$. We do so by training two phone recognition models, one on $\mathcal{D}$ and one on $\mathcal{D} - (\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{L'})$. More specifically, we finetune two Whisper (`small`; Radford et al., 2023) models on these datasets. We refer to the former as Whisper and the latter as Whisper-PPT. Further, we train an additional model, also with L' partitions removed, but on partitions that were not flagged by the PPT. We randomly selected and removed `Tamil`, `Xhosa`, `Cantonese`, `Hindi`, `Mandarin`, `Bengali`, `Finnish`, `Hungarian`, `Italian`, `Kyrgyz`, and `Maori`; we refer to this model Whisper-Anti-PPT. We provide hyperparameter details in Appendix A. To contextualize our results, we also compute performance for two other models that demonstrated strong performance on multilingual phone recognition: Taguchi et al. (XLS-R ND; 2023) and Xu et al. (XLS-R FAIR; 2022).

Model	Th.	En.	Ja	Pa.	Am.	Avg
XLS-R FAIR	20.3	**7.0**	15.9	15.1	18.2	15.3
XLS-R ND	19.0	15.2	**10.5**	15.2	18.6	15.7
Whisper	**15.1**	13.7	11.5	13.8	16.6	14.1
Whisper-PPT	15.2	12.3	11.2	**11.0**	**16.4**	**13.2**
Whisper-Anti-PPT	34.8	14.8	14.8	14.0	17.3	19.1

Table 3: Performance on held-out languages from IPAPack-Fleurs, as measured by Phonetic Feature Error Rate (Median). We use the *small* variant of the Whisper model. The macro-average is computed by averaging the error rate across all 5 languages, with each language given equal weight.

For our evaluation dataset, we select 5 language partitions from Section 3.3 that passed the PPT, indicating they are trustworthy for evaluation. Specifically, we evaluate on a test set $\mathcal{D}_{\text{test}}$ comprising the X-IPAPack evaluation partitions for Thai, English, Japanese, Punjabi, and Amharic. We hold these languages out from training of the two Whisper-based phone recognition models. To evaluate our models, we again compute the PFER using phonetic feature vectors from `panphon` (Mortensen et al., 2016).

4.1 Results

In Table 3, we demonstrate the performance of the four models (XLS-R FAIR, XLS-R ND, Whisper, Whisper-PPT) on the 5 held out languages. Averaging across the 5 partitions, we find that Whisper-PPT achieves the best performance. Importantly, it improves upon Whisper, despite the latter being trained with more datapoints (drawn from the low-quality transcript languages identified in Section 3.2) and 12% more optimization steps (3982 vs. 4466, respectively, for completing 2 epochs of training on their respective datasets).

Low-quality Data Impairs Performance. Comparing Whisper and Whisper-PPT, we find the largest improvement on Punjabi (13.82 vs. 11.01 PFER), a relative improvement of 20.3%. One likely reason for this sizeable improvement is that through the PPT (Section 3.2), we pinpointed in X-IPAPack a related language (Sindhi) that contained low-quality transcripts, and removed it from the training set for Whisper-PPT. Both Punjabi and Sindhi are Indo-Aryan languages (Dryer and Haspelmath, 2013). Malayalam may have also had an impact. Although it is from a

different language familiy (Dravidian), it is also spoken in the Indian subcontinent and may share some common phonetic features from broad areal effects (Everett et al., 2015). We also observe improvements (albeit smaller ones) on English, Japanese, and Eritrean, with Thai being the only language where Whisper achieves a slightly higher PFER over Whisper-PPT.

Importantly, we find that specifically removing language partitions identified by the PPT results in improved performance. Removing randomly selected language partitions (Whisper-Anti-PPT), results in far worse performance than any of the baseline models, with a macro-averaged PFER of 19.1, a full 6 points below Whisper-PPT.

Multilingual Phone Recognition Models are not Language-agnostic. Comparing Whisper-PPT with XLS-R FAIR and XLS-R ND, we find that the former achieves better performance in all but two: English for XLS-R FAIR and XLS-R ND for Japanese. This is likely due to XLS-R FAIR having English in its training set (Xu et al., 2022) and XLS-R ND having Japanese in its training set (Taguchi et al., 2023).[9] Moreover, when we train another Whisper-PPT on all languages, including X-IPAPack-English, we obtain better performance on English than XLS-R FAIR (PFER of 6.95 vs. 6.32). Thus, we find that despite all the models having been trained on a reasonably diverse set of languages, performance still varies depending on the exact training data composition.

4.2 Evaluation on IPAPack-DoReCo

Next, we evaluate the models on another partition of X-IPAPack, the DoReCo partition (rather than the Fleurs partition we have been using up to this point, Zhu et al., 2024) to further assess model capabilities on an out-of-distribution set of languages. X-IPAPack-DoReCo comprises phonetically transcribed speech for 44 endangered languages (Paschen et al., 2020). The utterances tend to be shorter than X-IPAPack-Fleurs (Zhu et al., 2024). Note that we don't apply the PPT on X-IPAPack-DoReCo, since its construction had oversight from expert linguists for each language (Zhu et al., 2024; Paschen et al., 2020). This evaluation sheds further light on model performance since these languages were never seen during finetuning for any of the models; they are also highly unlikely to have been observed during the multilingual pretraining stages for Whisper (Radford et al., 2023) and XLS-R (Babu et al., 2022).

Model	PFER (Median)	Var. (IQR)	Best
XLS-R FAIR	**5.19**	4.21	**29/44**
XLS-R ND	5.48	4.18	5/44
Whisper-PPT	5.60	4.39	10/44
Whisper	7.54	66.89	0/44
Allosaurus	7.53	5.04	0/44

Table 4: Evaluation results on X-IPAPack-DoReCo (44 languages). PFER: Phonetic Feature Error Rate (Mortensen et al., 2016). IQR: Interquartile Range. Highest refers to the number of times the model performed the best within one of the X-IPAPack-DoReCo languages.

Poor Data Quality Impairs Out-of-distribution Performance. In Table 4, we see that Whisper-PPT significantly outperforms Whisper (PFER 5.60 vs. 7.54, representing a 25.7% error rate improvement). We emphasize that the improvement arises solely from having removed the low-quality language partitions from finetuning (Section 3.3), as no other factors were manipulated. We find that trained on the unfiltered X-IPAPack archive, Whisper performs no better than a much older model, Allosaurus. We also find that Whisper is prone to entirely degenerate predictions, such as empty strings and predictions of the same character ad-nauseam, leading to some predictions that incur an extremely high error and a high variance (Interquartile Range; IQR) of 66.89. Whisper-PPT is also susceptible to degenerate predictions, but to a much lesser degree, given its reasonable IQR (4.39). Overall, our results suggest that degenerate predictions are exacerbated by low-quality training data.[10]

XLS-R FAIR Achieves Strongest Out-of-distribution Performance, though there is Language-conditional Variance. In Table 4, taking the median across all datapoints in all 44 languages, we find that XLS-R FAIR achieves the best performance, with a median PFER of 5.19. Whisper-PPT and XLS-R ND achieve similar performance (PFER 5.60 and 5.48, respectively).

[9] It's worth noting that Japanese is also in the XLS-R FAIR training set. Despite this, its performance compared to Whisper-PPT and XLS-R ND is considerably worse.

[10] All of our experiments use greedy generation for all models. We did not find substantive improvements from beam search.

Click	Freq.	Whisper-PPT	XLS-R FAIR	XLS-R ND
ǁ	35	**.83**	0.00	0.00
ǀ	87	**.73**	0.00	0.00
ǃ	102	**.89**	0.00	0.00

Table 5: Recall on predicting click consonants. Only the model trained on X-IPAPack is able to predict clicks, which manifest in Zulu and Xhosa in the X-IPAPack data.

Overall, while XLS-R FAIR exhibits dominance in this test split, we note that there is language conditional variability. There are $10/44$ and $5/44$ languages where Whisper-PPT or XLS-R ND obtain better performance.

5 Dataset Coverage Limits Downstream Performance

Our analyses in training phone recognizers in Section 3 demonstrated that poor-quality data impairs generalization. However, filtering out low-quality data from multilingual archives does not necessarily guarantee multilingual generalization. Prior work had claimed universal phone recognition capacity from training on a small number of languages on high-quality data (Taguchi et al., 2023), but we demonstrate that multilingual generalization remains limited by the training data composition. The number of attested sounds across the world's languages is large (Moran and McCloy, 2019), and their frequencies have a Zipfian distribution (Macklin-Cordes and Round, 2020), making them challenging to learn from limited data. We demonstrate these challenges through two phone segment-level error analyses.

Purported Universal Phone Recognizers Cannot Transcribe Clicks. Since XLS-R FAIR and XLS-R ND were not trained on any Bantu languages or any other language or dialect containing click consonants, we find they are incapable of predicting clicks in (Table 5). By comparison, X-IPAPack contains Zulu and Xhosa transcripts. Moreover, they pass the PPT test (Section 3.2), indicating that they are reliable. Indeed, we find that Whisper-PPT performs well at identifying clicks in the evaluation dataset for these languages when they are present in the recording. This demonstrates training a universal phone transcription model is more challenging than previously thought, since it requires accurate identification of typologically rare sounds that may not be prevalent in prior archives.

Tgt. phone	Lang.	EPR (Whisper-PPT/FAIR/ND)	Maj. label
ʈ	pa_in	**0.09**/0.25/0.21	ʈ, t, ʈ
ʈʰ	pa_in	**0.14**/0.21/0.36	ʈʰ, t, k
ɖ	pa_in	**0.15**/0.26/0.19	ɖ, d, ɖ
ð	en_us	0.25/**0.08**/0.32	d, ð, d
ɹ	en_us	0.23/**0.14**/0.36	r, ɹ, -
ɪ	en_us	0.19/**0.11**/0.25	ɪ, ɪ, i
ɕ	ja_jp	0.25/0.20/**0.16**	ʃ, ʃ, ɕ
eː	ja_jp	**0.04**/0.13/0.10	eː, e, eː
e	ja_jp	0.07/0.14/**0.06**	e, e, e

Table 6: Models vary in their ability to predict certain phones, with Whisper-PPT better at Punjabi (pa_in); XLS-R ND at Japanese (ja_jp); and XLS-R FAIR at English (en_us). EPR: Expected Phone Error for the three models **(Whisper-PPT/XLS-R FAIR/XLS-R ND)**. Maj. label: the phone most commonly predicted by each of the three models (same order as EPR).

To provide finer-grained measurements of efficacy at recognizing a certain sound, we use the **Expected Phone Error (EPR)** metric, defined as follows. Given a phone q, we identify all its occurrences in the ground-truth transcripts $G(t)$. Assuming that q appears at position i in transcript $G(t)$, we then compute the error as the phonetic feature distance (using `panphon`; Mortensen et al., 2016) between $G(t)_i$ and the phone in $\mathcal{M}(x)$ that is aligned to $G(t)_i$. We then average the error from each occurrence of q in the dataset. We apply the Needleman-Wunsch algorithm for the alignment.

Models Vary Widely in the Phonetic Details they can Capture. In Table 6, we show three languages where one of the three models excels. For Punjabi, we see that Whisper-PPT is highly effective at identifying retroflex consonants, even distinguishing between aspirated and unaspirated stops (ʈʰ and ʈ), an important phonemic distinction in Punjabi (Jain and Cardona, 2007). By comparison, XLS-R FAIR predicts unmarked alveolar stop (t) in both cases. For Japanese, we find that both Whisper-PPT and XLS-R ND can distinguish between e and eː (another phonemic distinction) while XLS-R FAIR cannot. For English, we find

that XLS-R FAIR is significantly better at distinguishing between dental stops and fricatives, as well as high-front and near-high-front vowels (i and ɪ) than the other two models. This is unsurprising since it is the only model with English in its training dataset.

Summary. Our results indicate that the training data composition remains relevant; training models even on fairly large multilingual archives (Xu et al., 2022) does not enable fine-grained cross-linguistic generalization. Thus, acquiring language-specific training data remains important. Moreover, our results in Section 4.1 demonstrate that a systematic audit of the quality of acquired data is highly beneficial for downstream performance. Our Preference Proportion Test (Section 3.2) enables systematic and sample-efficient quality auditing for this purpose.

6 Conclusion

We present the Preference Proportion Test (PPT; Section 3.2) for efficiently and systematically auditing the quality of data from specific languages in a large multilingual archive. We apply the PPT for efficiently identifying low-quality language partitions in the recent X-IPAPack archive (Zhu et al., 2024). Our audit is effective, efficiently identifying language partitions whose complete removal brings substantial improvements in the downstream task of automatic phone recognition. Appraising the quality of multilingual archives is critical stewardship, ensuring that they are reliable, trustworthy, and representative (Kreutzer et al., 2022; Desai et al., 2024). Overall, our method contributes an important procedure for statistically-principled multilingual archive auditing.

7 Limitations

Managing Partitions that Fail The PPT. One concern with the PPT is clear: Annotating only a small fraction of samples in a language partition, yet using this as a judgment to discard the entire partition. In our study on phonetic transcript archives, this was justifiable as it became transparent that there were systematic cases of insertions, deletions, and substitutions for language varieties that failed the test (Figure 4). These errors were not straightforwardly repairable. However, we recognize that this is a highly context-specific decision. In other cases, it is possible that the systematic errors can be easily repaired with a rule-based data preprocessing script, making it possible to salvage the partition. Or the negative result may be the quality of a particularly poor-quality sample. We clarify that (failing) the PPT only indicates that the partition should be treated *differently* relative to the rest of the archive. Whether the treatment of that partition should be outright rejection (our case) or repair will vary depending on the task or dataset.

What Does it Mean to Pass the PPT? Passing the PPT does not suggest that the partition should be free from further scrutiny. Appraisal is contextually dependent (Desai et al., 2024), so a positive appraisal for one task does not suggest a positive one for *all* tasks. Moreover, the baseline for a positive appraisal can change over time. Finally, and importantly, if there is considerable variance in the quality of a language partition's samples, then the PPT would also give different results depending on the random sample that it is applied to. Still, we believe it is a net positive to apply the PPT nonetheless, since it forces practitioners and researchers to carefully and intentionally engage with the contents of the dataset, as opposed to algorithmically filtering and categorizing it (Desai et al., 2024, see Section 2).

Selecting an Effective Baseline for The PPT. Another concern is the ambiguity inherent in selecting a baseline model in order to pursue the comparative approach espoused by the PPT. We recognize that for our computational task of phonetic transcription, there was a sizeable and still growing body of literature, making it straightforward to select a performant baseline model.[11] This was important, as it served as a cognitive forcing function to ensure that the annotators (co-authors) paid close attention to whether substrings in the transcripts truly reflected the audio.

It is conceivable that for other language variety and task combinations, no suitable baseline model will be available. However, it is important to acknowledge efforts both in industry and academia towards supporting more language varieties. Hayase et al. (2024) show that language

[11] In addition to the works of Xu et al. (2022) and Taguchi et al. (2023), there is also Allophant (Glocker et al., 2023) and the hierarchical model of Li et al. (2021).

models' data mixtures are increasing the heterogeneity of their training datasets with respect to language varieties, making them more widely applicable for downstream tasks. BLOOM, a model collaboratively developed across several research organizations, was reportedly trained on 46 languages (Muennighoff et al., 2023). mT5 was reportedly trained on 101 languages (Xue et al., 2021). Even if these models are not effective outside of the box, finetuning even (relatively) small models like mt5-base significantly improves their effectiveness (Asai et al., 2024, Table 4). When it comes to speech recognition models, the OWSM model is trained on over 150 language varieties (Peng et al., 2024), while the model of Zhao et al. (2025) is trained on over 1,000 language varieties.

Another important consideration, in light of our results of the effectiveness of small-sample auditing, is to recruit language-proficient annotators. While researchers often invoke costs as they pertain to scalability for annotating massive datasets (Scheuerman et al., 2021), this argument is less forceful for annotating only dozens of samples for the purposes of auditing.

Case Study Limitations. Our study was limited to the X-IPAPack dataset, since it is among the most recently published and thus actively maintained. There are other archives which we did not audit with the PPT, including VoxClamantis (Salesky et al., 2020) and VoxCommunis (Ahn and Chodroff, 2022), comprising over 600 language varieties and 36 language varieties, respectively.

Acknowledgments

V. S.'s research is funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs program, NSERC, and a research gift from AI2. F. S. is supported by an NSERC PGS-D scholarship. J. Z. is supported in part by NSERC Discovery Program and CFI. The computing resources were provided by UBC ARC and Digital Alliance of Canada.

References

Emily P. Ahn and Eleanor Chodroff. 2022. Voxcommunis: A corpus for cross-linguistic phonetic analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5286–5294.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.100

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282. https://doi.org/10.21437/Interspeech.2022-143

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. https://doi.org/10.1145/3442188.3445922

Ruha Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge and Medford: Polity Press.

Steven Bird. 2021. Sparse transcription. *Computational Linguistics*, 46(4):713–744. https://doi.org/10.1162/coli_a_00387

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons.

Biometrika, 39(3/4):324–345. https://doi.org/10.1093/biomet/39.3-4.324

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.745

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jacob Cohen. 1992. Statistical power analysis. *Current Directions in Psychological Science*, 1(3):98–101. https://doi.org/10.1111/1467-8721.ep10768783

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE. https://doi.org/10.1109/SLT54892.2023.10023141

Meera A. Desai, Irene V. Pasquetto, Abigail Z. Jacobs, and Dallas Card. 2024. An archival perspective on pretraining data. *Patterns*, 5(4). https://doi.org/10.1016/j.patter.2024.100966, PubMed: 38645763

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.98

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1128

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo. https://doi.org/10.5281/zenodo.7385533

Caleb Everett, Damián E. Blasi, and Seán G. Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, 112(5):1322–1327. https://doi.org/10.1073/pnas.1417413112, PubMed: 25605876

Jennifer Drexler Fox, Desh Raj, Natalie Delworth, Quinn McNamara, Corey Miller, and Migüel Jetté. 2024. Updated corpora and benchmarks for long-form speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13246–13250. IEEE. https://doi.org/10.1109/ICASSP48485.2024.10446286

Kevin Glocker, Aaricia Herygers, and Munir Georges. 2023. Allophant: Cross-lingual phoneme recognition with articulatory attributes. In *Interspeech 2023*, pages 2258–2262. https://doi.org/10.21437/Interspeech.2023-772

Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, and Noah A. Smith. 2024. Data mixture inference: What do BPE tokenizers reveal about their training data? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in NLP. *arXiv preprint arXiv:2311.08391*. Version 1.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. Training compute-optimal large

language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–17. https://doi.org/10.1145/3689904.3694702

Guillaume Jacques. 2011. A panchronic study of aspirated fricatives, with new evidence from Pumi. *Lingua*, 121(9):1518–1538. https://doi.org/10.1016/j.lingua.2011.04.003

Danesh Jain and George Cardona. 2007. *The Indo-Aryan Languages*. Routledge. https://doi.org/10.4324/9780203945315

Jon Kleinberg and Eva Tardos. 2006. *Algorithm design*. Pearson Education India.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72. https://doi.org/10.1162/tacl_a_00447

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. MADLAD-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

William Lane and Steven Bird. 2021. Local word discovery for interactive transcription. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2058–2067. https://doi.org/10.18653/v1/2021.emnlp-main.157

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE. https://doi.org/10.1109/ICASSP40776.2020.9054362

Xinjian Li, Juncheng Li, Florian Metze, and Alan W. Black. 2021. Hierarchical phone recognition with compositional phonetics. In *Interspeech 2021*, pages 2461–2465. https://doi.org/10.21437/Interspeech.2021-1803

Johann-Mattis List and Robert Forkel. 2016. LingPy. A Python library for historical linguistics. *Max Planck Institute for the Science of Human History: Jena*.

Jayden L. Macklin-Cordes and Erich R. Round. 2020. Re-evaluating phoneme frequencies. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.570895, PubMed: 33329209

Mark Mazumder, Colby Banbury, Josh Meyer, Pete Warden, and Vijay Janapa Reddi. 2021. Few-shot keyword spotting in any language. In *Interspeech 2021*, pages 4214–4218. https://doi.org/10.21437/Interspeech.2021-1966

Steven Moran and Daniel McCloy. 2019. PHOIBLE 2.0. *Jena: Max Planck Institute for the Science of Human History*, 10.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.891

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453. https://doi.org/10.1016/0022-2836(70)90057-4, PubMed: 5420325

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.983

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2657–2666.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee weon Jung, and Shinji Watanabe. 2024. Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer. *Interspeech*, abs/2401.16658. https://doi.org/10.21437/Interspeech.2024-1194

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 28492–28518. PMLR.

Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W. Black, and Jason Eisner. 2020. A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.415

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2). https://doi.org/10.1145/3476058

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal automatic phonetic transcription into the international phonetic alphabet. In *Interspeech 2023*, pages 2548–2552. https://doi.org/10.21437/Interspeech.2023-2584

Jinchuan Tian, Yifan Peng, William Chen, Kwanghee Choi, Karen Livescu, and Shinji Watanabe. 2024. On the effects of heterogeneous data sources on speech-to-text foundation models. In *Interspeech 2024*, pages 3959–3963. https://doi.org/10.21437/Interspeech.2024-1938

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation

learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.80

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Interspeech 2022*, pages 2113–2117. https://doi.org/10.21437/Interspeech.2022-60

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Online. Association for Computational Linguistics. https://doi.org/10.1109/ICASSP49660.2025.10888774

Jinming Zhao, Vineel Pratap, and Michael Auli. 2025. Scaling a simple approach to zero-shot speech recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. https://doi.org/10.1109/ICASSP49660.2025.10888774

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of IPA: Towards open-vocabulary keyword spotting and forced alignment in any language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 750–772, Mexico City, Mexico. Association for Computational Linguistics.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 model for massively multilingual grapheme-to-phoneme conversion. In *Interspeech 2022*, pages 446–450. https://doi.org/10.21437/Interspeech.2022-538

A Hyperparameter Settings

See Table 7. We performed a hyperparameter sweep using the validation sets in X-IPAPack for the evaluation languages. Our hyperparameter sweep was over batch sizes of $\{8, 16, 32, 64, 128, 256\}$, precision of either FP16 or BF16, and warmup steps of $\{500, 1000\}$ and uniformly distributed learning rates of $\{1e-5, 3e-4\}$.

Hyperparam.	**Setting**
Batch size	64
Num. epochs	2
Learning rate	$1e-4$
FP16	True
Max gen. length	225
Grad. checkpointing	True
Warmup steps	500

Table 7: Hyperparameter settings for finetuning the Whisper models in Section 4.

B Annotation Interface

We show the interface for performing the annotations in Figure 5. The annotator must select from one of four options. One of the transcripts is the gold-standard transcript from X-IPAPack, while the other is generated from a baseline model (either XLS-R FAIR or XLS-R ND). We randomize whether the gold-standard or the baseline prediction is displayed first.

As long as the annotator picks either the gold-standard or the baseline model prediction, and not one of the abstention options, they are given the opportunity to select which word(s) most influenced their selection (Figure 6). We store these selections purely for documentation of the annotation, it does not influence any of the analyses in the main text.

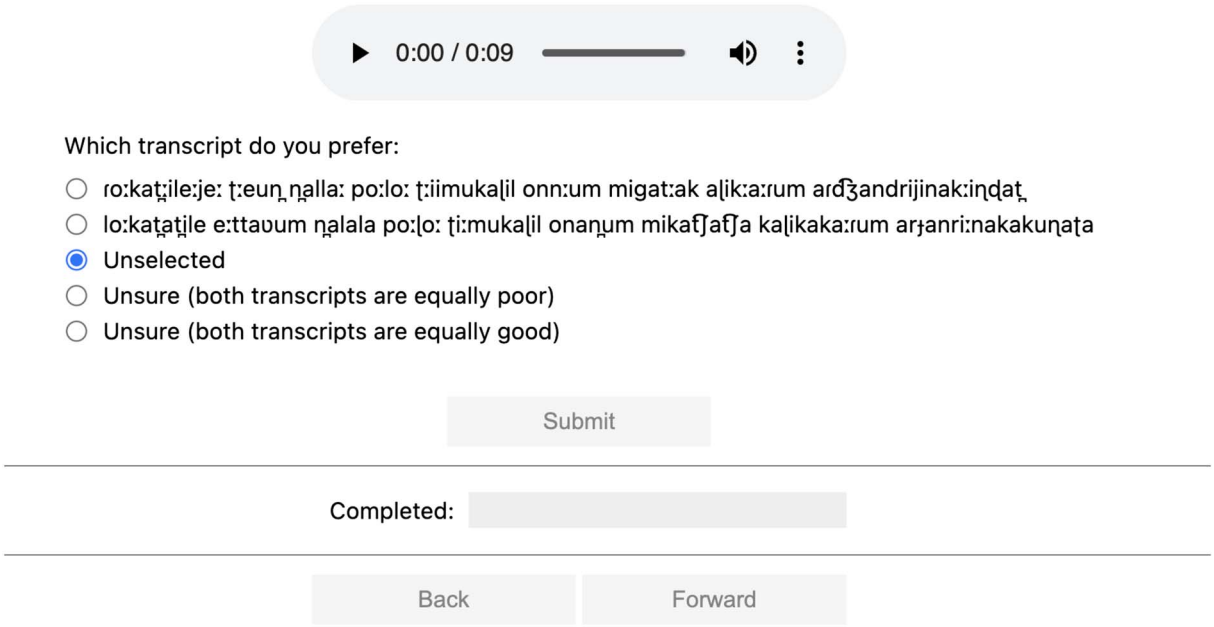

Figure 5: Interface for performing annotations.

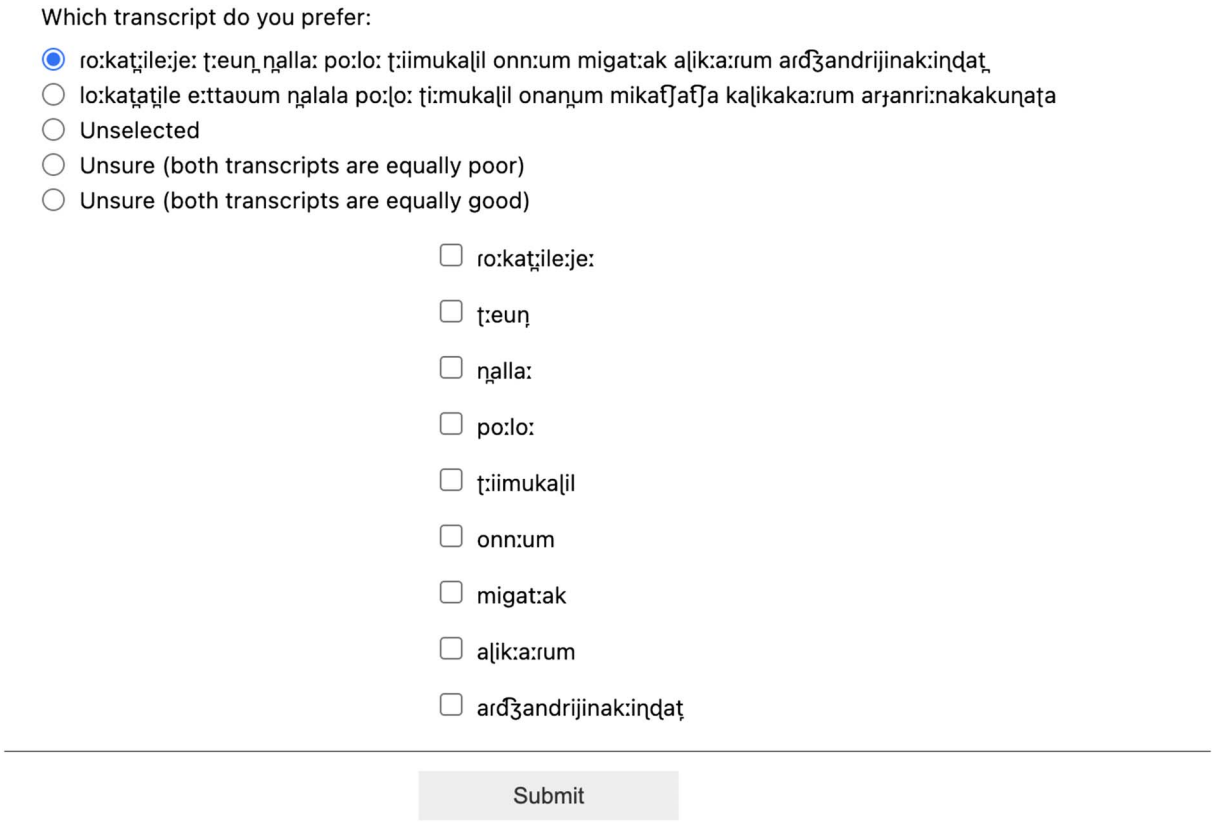

Figure 6: Selecting substrings that were influential in the decision once a decision is made.

First, we thank you for agreeing to helping us with providing your preference annotations! Your expertise in phonetic transcription is valuable and appreciated. Please read the following instructions carefully.

Options. The annotation task will comprise 50 datapoints of the form: (audio, transcript1, transcript2). You'll listen to the audio and then pick one of four options:

1. TranscriptA > TranscriptB
2. TranscriptA < TranscriptB
3. TranscriptA = TranscriptB (both are equally good)
4. TranscriptA = TranscriptB (both are equally poor)

You may replay the audio as many times as you like. You can also change the playback speed of the audio using the ''kebab menu'' (three dots).

Justification. When you select (1), you will then select which word(s) in the transcript were better represented in TranscriptA compared to TranscriptB. (Analogous for Option 2). No need to be exhaustive here, just select some of the word(s) that seemed most well represented to you (relative to the other transcript).

Avoid abstaining (options 3 and 4) if possible. Try to select option (1) or (2) when possible, only resorting to (3) or (4) when you find it impossible to pick between the two. (Ideally, no more than 10 samples should have the (3) or (4) option). When both TranscriptA and TranscriptB have problems, try to select the transcript that has fewer problems.

Transcript spacing. When comparing the two transcripts, don't use the whitespace segmentation of the transcripts in informing your decision. The spaces are automatically inserted to improve readability of the transcripts, and there may occasionally be some spacing errors in TranscriptA relative to TranscriptB (or vice-versa). For example, one of the transcripts may segment the phrase ''the car'' into ''ðək ɑr'' instead of ''ðə kɑr''. Please try to ignore these spacing discrepancies in making your determination, instead focusing on whether the phonetic segments accurately represent the speech audio. For example, if TranscriptA = ''θi kɛr'' while TranscriptB = ''ðək ɑr'', you should prefer TranscriptB since it has more accurate phonetic segments (assuming a Standard American English pronunciation), and ignore the fact that the space is inserted after the ''k'' in kɑr rather than after the vowel in ''ðə''. Note that the spacing can also result in affricates (''ʤ'') being broken up (''d ʒ''), so if you clearly hear an affricate but don't see a tie bar in the transcript, the affricate may well be represented but (inadvertently) broken up between two words.

Going back to previous annotations. The interface contains back and forward buttons. When you hit the back button to navigate to the previous sample, you'll see that your prior annotation is stored. If you want to select a different option (1–4), you can do so and hit submit. If you want to leave it as is, you can just hit the ''Forward'' button again. When you use ''Back'', the ''Submit'' button will be disabled unless you change your selection.

Display. On rare occasions, some of the transcripts can be a bit long and rendered incorrectly. Ensuring that the window is full-screen and using a larger external monitor (if possible) should mitigate this.
When you are ready, execute the next cell to begin. At the bottom of the cell, the interface for performing the annotations will appear. Your progress will be saved as you work through the annotations, so feel free to take breaks.