# A Corpus for Large-Scale Phonetic Typology

**Elizabeth Salesky**[ʔ] **Eleanor Chodroff**[ɤ] **Tiago Pimentel**[ʒ] **Matthew Wiesner**[ʔ]
**Ryan Cotterell**[ʒ,ð] **Alan W Black**[ɬ] **Jason Eisner**[ʔ]
[ʔ]Johns Hopkins University [ɤ]University of York
[ʒ]University of Cambridge [ð]ETH Zürich [ɬ]Carnegie Mellon University
esalesky@jhu.edu eleanor.chodroff@york.ac.uk

## Abstract

A major hurdle in data-driven research on typology is having sufficient data in many languages to draw meaningful conclusions. We present VoxClamantis v1.0, the first large-scale corpus for phonetic typology, with aligned segments and estimated phoneme-level labels in 690 readings spanning 635 languages, along with acoustic-phonetic measures of vowels and sibilants. Access to such data can greatly facilitate investigation of phonetic typology at a large scale and across many languages. However, it is non-trivial and computationally intensive to obtain such alignments for hundreds of languages, many of which have few to no resources presently available. We describe the methodology to create our corpus, discuss caveats with current methods and their impact on the utility of this data, and illustrate possible research directions through a series of case studies on the 48 highest-quality readings. Our corpus and scripts are publicly available for non-commercial use at https://voxclamantisproject.github.io.

## 1 Introduction

Understanding the range and limits of cross-linguistic variation is fundamental to the scientific study of language. In speech and particularly phonetic typology, this involves exploring potentially universal tendencies that shape sound systems and govern phonetic structure. Such investigation requires access to large amounts of cross-linguistic data. Previous cross-linguistic phonetic studies have been limited to a small number of languages with available data (Disner, 1983; Cho and Ladefoged, 1999), or have relied on previously reported measures from many studies (Whalen and Levitt, 1995; Becker-Kristal, 2010; Gordon and Roettger, 2017; Chodroff et al., 2019).

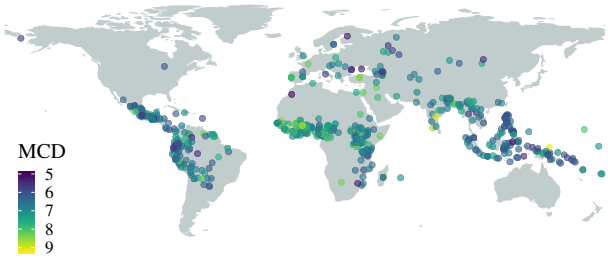



Figure 1: The 635 languages of our corpus geo-located with mean Mel Cepstral Distortion (MCD) scores.

Existing multilingual speech corpora have similar restrictions, with data too limited for many tasks (Engstrand and Cunningham-Andersson, 1988; Ladefoged and Maddieson, 2007) or approximately 20 to 30 recorded languages (Ardila et al., 2020; Harper, 2011; Schultz, 2002).

The recently developed CMU Wilderness corpus (Black, 2019) constitutes an exception to this rule with over 600 languages. This makes it the largest and most typologically diverse speech corpus to date. In addition to its coverage, the CMU Wilderness corpus is unique in two additional aspects: cleanly recorded, read speech exists for all languages in the corpus, and the same content (modulo translation) exists across all languages.

However, this massively multilingual speech corpus is challenging to work with directly. Copyright, computational restrictions, and sheer size limit its accessibility. Due to copyright restrictions, the audio cannot be directly downloaded with the sentence and phoneme alignments. A researcher would need to download original audio MP3 and text through links to bible.is, then segment these with speech-to-text sentence alignments distributed in Black (2019).[1] For phonetic research, subsequently identifying examples of specific phonetic segments in the audio is also a near-essential

[1]The stability of the links and recording IDs is also questionable. Since the release of Black (2019), many of the links have already changed, along with a few of the IDs. We have begun identifying these discrepancies, and plan to flag these in a future release.

step for extracting relevant acoustic-phonetic measurements. Carrying out this derivative step has allowed us to release a stable-access collection of token-level acoustic-phonetic measures to enable further research.

Obtaining such measurements requires several processing steps: estimating pronunciations, aligning them to the text, evaluating alignment quality, and finally, extracting phonetic measures. This work is further complicated by the fact that, for a sizable number of these languages, no linguistic resources currently exist (e.g., language-specific pronunciation lexicons). We adapt speech processing methods based on Black (2019) to accomplish these tasks, though not without noise: in §3.4, we identify three significant caveats when attempting to use our extended corpus for large-scale phonetic studies.

We release a comprehensive set of standoff markup of over 400 million labeled segments of continuous speech.[2] For each segment, we provide an estimated phoneme-level label from the X-SAMPA alphabet, the preceding and following labels, and the start position and duration in the audio. Vowels are supplemented with formant measurements, and sibilants with standard measures of spectral shape.

We present a series of targeted case studies illustrating the utility of our corpus for large-scale phonetic typology. These studies are motivated by potentially universal principles posited to govern phonetic variation: **phonetic dispersion** and **phonetic uniformity**. Our studies both replicate known results in the phonetics literature and also present novel findings. Importantly, these studies investigate current methodology as well as questions of interest to phonetic typology at a large scale.

## 2 Original Speech

The CMU Wilderness corpus (Black, 2019) consists of recorded readings of the New Testament of the Bible in many languages and dialects. Following the New Testament structure, these data are broken into 27 books, each with a variable number of chapters between 1 and 25. Bible chapters contain standardized verses (approximately sentence-level segments); however, the speech is originally split only by chapter. Each chapter has an average of 13 minutes of speech for a total of ≈20 hours of speech and text per language. These recordings are clean, read speech with a sampling rate of 16 kHz. In most languages, they are non-dramatic readings with a single speaker; in some, they are dramatic multi-speaker readings with additive music.[3] The release from Black (2019) includes several resources for processing the corpus: scripts to download the original source data from `bible.is`, 'lexicons' created using grapheme-to-phoneme (G2P) conversion, and scripts to apply their generated sentence alignments, which facilitates downstream language processing tasks, including phoneme alignment.

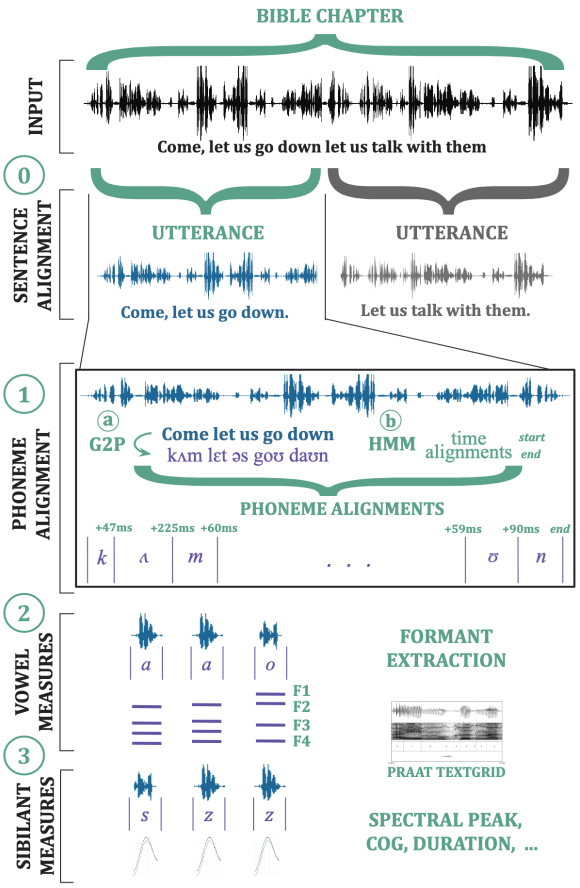



Figure 2: The extraction process for the measurements released in VoxClamantis v1.0.

## 3 The VoxClamantis v1.0 Corpus

Our VoxClamantis v1.0 corpus is derived from 690 audio readings of the New Testament of the Bible[4] in 635 languages.[5] We mark estimated speech seg-

[2]For some languages, we provide multiple versions of the markup based on different methods of predicting the pronunciation and generating time alignments (§3.1).

[3]Information about the recordings available can be found at `https://www.faithcomesbyhearing.com/mission/recordings`

[4]Nine of the readings from Black (2019) could not be aligned.

[5]We specify number of distinct languages by the number of distinct ISO 639-3 codes, which may not distinguish dialects.

ments labeled with phonemic labels, and phonetic measures for the tokens that are vowels or sibilants. The extraction process is diagrammed in Figure 2. In the sections below, we detail our procedures for extracting labeled audio segments and their phonetic measures, in both high- and low-resource languages. We then outline important caveats to keep in mind when using this corpus.

### 3.1 Extracting Phoneme Alignments

We use a multi-pronged forced alignment strategy to balance broad language coverage (§3.1.1) with utilization of existing high-quality resources (§3.1.2). We assess the quality of our approaches in §3.1.3. We release the stand-off markup for our final alignments as both text files and Praat TextGrids (Boersma and Weenink, 2019).[6]

Using scripts and estimated boundaries from Black (2019), we first download and convert the audio MP3s to waveforms, and cut the audio and text into 'sentences' (hereafter called 'utterances' as they are not necessarily sentences). This step creates shorter-length speech samples to facilitate forced alignment; utterance boundaries do not change through our processing.

To extract labeled segments, we first require pronunciations for each utterance. A pronunciation is predicted from the text alone using some grapheme-to-phoneme (G2P) method. Each word's predicted pronunciation is a sequence of categorical labels, which are 'phoneme-level' in the sense that they are usually intended to distinguish the words of the language. We then align this predicted sequence of 'phonemes' to the corresponding audio.

#### 3.1.1 All Languages

Most of our languages have neither existing pronunciation lexicons nor G2P resources. To provide coverage for all languages, we generate pronunciations using the simple 'universal' G2P system Unitran (Qian et al., 2010, as extended by Black, 2019), which deterministically expands each grapheme to a fixed sequence of phones in the Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) (Wells, 1995/2000). This naive process is error-prone for languages with opaque orthographies, as we show in §3.1.3 below and discuss further in §3.4 (Caveat **B**). Even so, it provides a starting point for exploring low-resource languages: after some manual inspection, a linguist may be able to correct the labels in a given language by a combination of manual and automatic methods.

For each reading, to align the pronunciation strings to the audio, we fit a generative acoustic model designed for this purpose: specifically, eHMM (Prahallad et al., 2006) as implemented in Festvox (Anumanchipalli et al., 2011) to run full Baum–Welch from a flat start for 15 to 30 iterations until the mean mel cepstral distortion score (see §3.1.3) converges. Baum-Welch does not change the predicted phoneme labels, but obtains a language-specific, reading-specific, contextual (triphone) acoustic model for each phoneme type in the language. We then use Viterbi alignment to identify an audio segment for each phoneme token.

#### 3.1.2 High-Resource Languages

A subset of the languages in our corpus are supported by existing pronunciation resources. Two such resources are Epitran (Mortensen et al., 2018), a G2P tool based on language-specific rules, available in both IPA and X-SAMPA, and WikiPron (Lee et al., 2020), a collection of crowd-sourced pronunciations scraped from Wiktionary. These are mapped from IPA to X-SAMPA for label consistency across our corpus. Epitran covers 29 of our languages (39 readings), while WikiPron's 'phonemic' annotations[7] provide partial coverage of 13 additional languages (18 readings). We use Epitran for languages with regular orthographies where it provides high-quality support, and WikiPron for other languages covered by WikiPron annotations. While Unitran and Epitran provide a single pronunciation for a word from the orthography, WikiPron may include multiple pronunciations. In such cases, Viterbi alignment (see below) chooses the pronunciation of each token that best fits the audio.

For most languages covered by WikiPron, most of our corpus words are out-of-vocabulary, as they do not yet have user-submitted pronunciations on Wiktionary. We train G2P models on WikiPron annotations to provide pronunciations for these words. Specifically, we use the WFST-based tool Phonetisaurus (Novak et al., 2016). Model hyperparameters are tuned on 3 WikiPron languages from SIGMORPHON 2020 (Gorman et al., 2020) (see Appendix C for details). In general, for languages that are not easily supported by Epitran-style G2P *rules*, training a G2P *model* on sufficiently many

[6] Corresponding audio will need to be downloaded from source and split by utterance using scripts from Black (2019).

[7] WikiPron annotations are available at both the phonemic and phonetic level, with a greater number of phonemic annotations, which we use here.

| ISO 639-3 | tpi | ron | azj | msa | ceb | tur | tgl | spa | ilo | rus | hau | ind | tgk | jav | kaz |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # Types | 1398 | 9746 | 18490 | 7612 | 8531 | 21545 | 9124 | 11779 | 15063 | 16523 | 4938 | 5814 | 12502 | 10690 | 20502 |
| Unitran PER | 18.4 | 21.3 | 26.9 | 30.1 | 30.1 | 31.2 | 34.4 | 34.4 | 35.0 | 37.4 | 37.6 | 38.8 | 39.8 | 49.9 | 46.8 |
| # Tokens | 291k | 169k | 125k | 157k | 190k | 125k | 185k | 168k | 169k | 130k | 201k | 170k | 159k | 177k | 142k |
| Weighted PER | 20.1 | 21.3 | 26.1 | 31.1 | 35.9 | 28.5 | 40.1 | 32.6 | 32.7 | 36.8 | 36.7 | 40.5 | 38.8 | 54.1 | 47.7 |

| ISO 639-3 | swe | kmr | som | tir | pol | hae | vie | tha | lao | ben | tel | hin | mar | tam |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # Types | 8610 | 8127 | 14375 | 22188 | 18681 | 15935 | 2757 | 23338 | 31334 | 8075 | 23477 | 7722 | 17839 | 31642 |
| Unitran PER | 46.9 | 54.3 | 54.6 | 57.8 | 67.1 | 67.3 | 73.8 | 80.3 | 89.1 | 90.0 | 90.3 | 95.7 | 97.8 | 100.5 |
| # Tokens | 165k | 176k | 156k | 121k | 141k | 164k | 211k | 26k | 36k | 173k | 124k | 191k | 159k | 139k |
| Weighted PER | 49.5 | 53.9 | 56.0 | 57.4 | 66.8 | 64.8 | 80.6 | 80.4 | 89.4 | 86.2 | 88.3 | 91.3 | 97.8 | 102.1 |

Table 1: Phoneme Error Rate (PER) for Unitran treating Epitran as ground-truth. 'Types' and 'Tokens' numbers reflect the number of unique word types and word tokens in each reading. We report PER calculated using word types for calibration with other work, as well as frequency-weighted PER reflecting occurrences in our corpus.

high-quality annotations may be more accurate.

We align the speech with the high-quality labels using a multilingual ASR model (see Wiesner et al., 2019). The model is trained in Kaldi (Povey et al., 2011) on 300 hours of data from the IARPA BABEL corpora (21 languages), a subset of Wall Street Journal (English), the Hub4 Spanish Broadcast news (Spanish), and a subset of the Voxforge corpus (Russian and French). These languages use a shared X-SAMPA phoneme label set which has high coverage of the labels of our corpus.

Our use of a pretrained multilingual model here contrasts with §3.1.1, where we had to train reading-specific acoustic models to deal with the fact that the same Unitran phoneme label may refer to quite different phonemes in different languages (see §3.4). We did not fine-tune our multilingual model to each language, as the cross-lingual ASR performance in previous work (Wiesner et al., 2019) suggests that this model is sufficient for producing phoneme-level alignments.

### 3.1.3 Quality Measures

Automatically generated phoneme-level labels and alignments inherently have some amount of noise, and this is particularly true for low-resource languages. The noise level is difficult to assess without gold-labeled corpora for either modeling or assessment. However, for the high-resource languages, we can evaluate Unitran against Epitran and WikiPron, pretending that the latter are ground truth. For example, Table 1 shows Unitran's phoneme error rates relative to Epitran. Appendix B gives several more detailed analyses with examples of individual phonemes.

Unitran pronunciations may have acceptable phoneme error rates for languages with transparent orthographies and one-to-one grapheme-to-phoneme mappings. Alas, without these conditions they prove to be highly inaccurate.

That said, evaluating Unitran labels against Epitran or WikiPron may be unfair to Unitran, since some discrepancies are arguably not errors but mere differences in annotation granularity. For example, the 'phonemic' annotations in WikiPron are sometimes surprisingly fine-grained: WikiPron frequently uses /t̪/ in Cebuano where Unitran only uses /t/, though these refer to the same phoneme. These tokens are scored as incorrect. Moreover, there can be simple systematic errors: Unitran always maps grapheme <a> to label /ɑ/, but in Tagalog, all such tokens should be /a/. Such errors can often be fixed by remapping the Unitran labels, which in these cases would reduce PER from 30.1 to 6.8 (Cebuano) and from 34.4 to 7.8 (Tagalog). Such rules are not always this straightforward and should be created on a language-specific basis; we encourage rules created for languages outside of current Epitran support to be contributed back to the Epitran project.

For those languages where we train a G2P system on WikiPron, we compute the PER of the G2P system on held-out WikiPron entries treated as ground truth. The results (Appendix C) range from excellent to mediocre.

We care less about the pronunciations themselves than about the segments that we extract by aligning these pronunciations to the audio. For high-resource languages, we can again compare the segments extracted by Unitran to the higher-quality ones extracted with better pronunciations. For each Unitran token, we evaluate its label and temporal boundaries against the high-quality token that is closest in the audio, as measured by the temporal distance between their midpoints (Appendix B).

Finally, the segmentation of speech and text into corresponding utterances is not perfect. We use the utterance alignments generated by Black (2019), in which the text and audio versions of a putative

utterance may have only partial overlap. Indeed, Black (2019) sometimes failed to align the Unitran pronunciation to the audio at all, and discarded these utterances. For each remaining utterance, he assessed the match quality using Mel Cepstral Distortion (MCD)—which is commonly used to evaluate synthesized spoken utterances (Kominek et al., 2008)—between the original audio and a resynthesized version of the audio based on the aligned pronunciation. Each segment's audio was resynthesized given the segment's phoneme label and the preceding and following phonemes, in a way that preserves its duration, using CLUSTERGEN (Black, 2006) with the same reading-specific eHMM model that we used for alignment. We distribute Black's per-utterance MCD scores with our corpus, and show the average score for each language in Appendix E. In some readings, the MCD scores are consistently poor.

### 3.2 Phonetic measures

Using the phoneme-level alignments described in §3.1, we automatically extract several standard acoustic-phonetic measures of vowels and sibilant fricatives that correlate with aspects of their articulation and abstract representation.

#### 3.2.1 Vowel measures

Standard phonetic measurements of vowels include the formant frequencies and duration information. Formants are concentrations of acoustic energy at frequencies reflecting resonance points in the vocal tract during vowel production (Ladefoged and Johnson, 2014). The lowest two formants, F1 and F2, are considered diagnostic of vowel category identity and approximate tongue body height (F1) and backness (F2) during vowel production (Figure 3). F3 correlates with finer-grained aspects of vowel production such as rhoticity (/ɹ/-coloring), lip rounding, and nasality (House and Stevens, 1956; Lindblom and Sundberg, 1971; Ladefoged et al., 1978), and F4 with high front vowel distinctions and speaker voice quality (Eek and Meister, 1994). Vowel duration can also signal vowel quality, and denotes lexical differences in many languages.

We extracted formant and duration information from each vowel using Praat (Boersma and Weenink, 2019). The first four formants (F1–F4) were measured at each quartile and decile of the vowel. Formant estimation was performed with the Burg algorithm in Praat with pre-emphasis from 50 Hz, a time window of 25 ms, a time step of 6.25 ms, a maximum of five formants permitted, and a formant ceiling of 5000 Hz, which is the recommended value for a male vocal tract (Boersma and Weenink, 2019). Note that the speakers in this corpus are predominantly male.

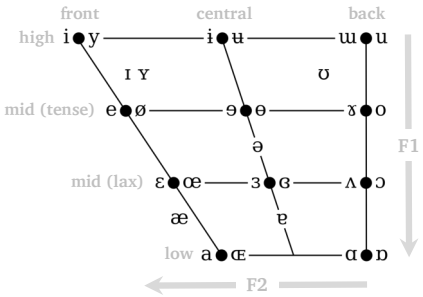



Figure 3: Vowel Chart

#### 3.2.2 Sibilant measures

Standard phonetic measurements of sibilant fricatives such as /s/, /z/, /ʃ/, and /ʒ/ include measures of spectral shape, and also segment duration. Measures of spectral shape frequently distinguish sibilant place of articulation: higher concentrations of energy generally reflect more anterior constriction locations (e.g., /s z/ are produced closer to the teeth than /ʃ ʒ/). Segment duration can also signal contrasts in voicing status (Jongman et al., 2000).

Our release contains the segment duration, spectral peak, the spectral moments of the frequency distribution (center of gravity: COG, variance, skewness, and kurtosis), as well as two measures of the mid-frequency peak determined by sibilant quality. These are the mid-frequency peak between 3000 and 7000 Hz for alveolar sibilants, and between 2000 and 6000 Hz for post-alveolar sibilants (Koenig et al., 2013; Shadle et al., 2016). The spectral information was obtained via multitaper spectral analysis (Rahim and Burr, 2017), with a time-bandwidth parameter ($nw$) of 4 and 8 tapers ($k$) over the middle 50% of the fricative (Blacklock, 2004). Measurements were made using the methods described in Forrest et al. (1988) for spectral moments and Koenig et al. (2013) for spectral peak varieties.

### 3.3 Computation times

Generating phoneme-level alignments and extracting subsequent phonetic measures takes significant time, computational resources, and domain knowledge. Our release enables the community to use this data directly without these prerequisites. Table 2 shows that the time to extract our resources,

| | Computation Time | |
|---|---|---|
| **Resource** | *Per Language* | *Total Time* |
| Utterance Alignments | 30m | 14d 13h |
| Phoneme Alignments | 3d 3h 37m | 6y 12d 16h |
| Vowel Measures | 45m | 21d 20h |
| Sibilant Measures | 20m | 9d 17h |
| | 3d 5h 0m | **6y 58d 19h** |

Table 2: Computation time to generate the full corpus.

once methods have been developed, was more than 6 CPU years, primarily for training eHMM models.

### 3.4 General caveats

We caution that our labeling and alignment of the corpus contains errors. In particular, it is difficult to responsibly draw firm linguistic conclusions from the Unitran-based segments (§3.1.1). In §5 we suggest future work to address these issues.

**A Quality of Utterance Pairs**: For some utterances, the speech does not correspond completely to the text, due to incorrect co-segmentation. In our phonetic studies, we threshold using reading-level MCD as a heuristic for overall alignment quality, and further threshold remaining readings using utterance-level MCD. We recommend others do so as well.

**B Phoneme Label Consistency and Accuracy**: Phoneme-level labels are predicted from text without the aid of audio using G2P methods. This may lead to systematic errors. In particular, Unitran relies on a 'universal' table that maps grapheme <s> (for example) to phoneme /s/ in every context and every language. This is problematic for languages that use <s> in some or all contexts to refer to other phonemes such as /ʃ/ or /s̪/, or use digraphs that contain <s>, such as <sh> for /ʃ/. Thus, the predicted label /s/ may not consistently refer to the same phoneme within a language, nor to phonetically similar phonemes across languages. Even WikiPron annotations are user-submitted and may not be internally consistent (e.g., some words use /d ʒ/ or /t/ while others use /ʤ/ or /t̻/), nor comparable across languages.

'Phoneme' inventories for Unitran and WikiPron have been implicitly chosen by whoever designed the language's orthography or its WikiPron pages; while this may reflect a reasonable folk phonology, it may not correspond to the inventory of underlying or surface phonemes that any linguist would be likely to posit.

**C Label and Alignment Assessment**: While alignment quality for languages with Epitran and WikiPron can be assessed and calibrated beyond this corpus, it cannot for those languages with only Unitran alignments; the error rate on languages without resources to evaluate PER is unknown to us. The Unitran alignments should be treated as a first-pass alignment which may still be useful for a researcher who is willing to perform quality control and correction of the alignments using automatic or manual procedures. Our automatically-generated alignment offers an initial label and placement of the boundaries that would hopefully facilitate downstream analysis.

**D Corpus Representation**: It is difficult to draw conclusions about 'average behavior' across languages. Some language families are better represented in the corpus than others, with more languages, more Bible readings per language, more hours of speech per reading, or more examples of a given phoneme of interest.[8] Additionally, the recordings by language are largely single-speaker (and predominantly male). This means that we can often draw conclusions only about a particular speaker's idiolect, rather than the population of speakers of the language. Metadata giving the exact number of different speakers per recording do not exist.

## 4 Phonetic Case Studies

We present two case studies to illustrate the utility of our resource for exploration of cross-linguistic typology. Phoneticians have posited several typological principles that may structure phonetic systems. Though previous research has provided some indication as to the direction and magnitude of expected effects, many instances of the principles have not yet been explored at scale. Our case studies investigate how well they account for cross-linguistic variation and systematicity for our phonetic measures from vowels and sibilants. Below we present the data filtering methods for our case studies, followed by an introduction to and evaluation of phonetic dispersion and uniformity.

### 4.1 Data filtering

For quality, we use only the tokens extracted using high-resource pronunciations (Epitran and WikiPron) and only in languages with mean

[8] See our corpus website for exact numbers of utterances and our phonetic measures per each language.

MCD lower than 8.0.[9] Furthermore, we only use those utterances with MCD lower than 6.0. The vowel analyses focus on F1 and F2 in ERB taken at the vowel midpoint (Zwicker and Terhardt, 1980; Glasberg and Moore, 1990).[10] The sibilant analyses focus on mid-frequency peak of /s/ and /z/, also in ERB. Vowel tokens with F1 or F2 measures beyond two standard deviations from the label- and reading-specific mean were excluded, as were tokens for which Praat failed to find a measurable F1 or F2, or whose duration exceeded 300 ms. Sibilant tokens with mid-frequency peak or duration measures beyond two standard deviations from the label- and reading-specific mean were also excluded. When comparing realizations of two labels such as /i/–/u/ or /s/–/z/, we excluded readings that did not contain at least 50 tokens of each label. We show data representation with different filtering methods in Appendix D.

After filtering, the vowel analyses included 48 readings covering 38 languages and 11 language families. The distribution of language families was 21 Indo-European, 11 Austronesian, 3 Creole/Pidgin, 3 Turkic, 2 Afro-Asiatic, 2 Tai-Kadai, 2 Uto-Aztecan, 1 Austro-Asiatic, 1 Dravidian, 1 Hmong-Mien, and 1 Uralic. Approximately 8.2 million vowel tokens remained, with a minimum of ≈31,000 vowel tokens per reading. The sibilant analysis included 22 readings covering 18 languages and 6 language families. The distribution of language families was 10 Indo-European, 6 Austronesian, 3 Turkic, 1 Afro-Asiatic, 1 Austro-Asiatic, and 1 Creole/Pidgin. The decrease in total number of readings relative to the vowel analysis primarily reflects the infrequency of /z/ cross-linguistically. Approximately 385,000 /s/ and 83,000 /z/ tokens remained, with a minimum of ≈5,200 tokens per reading.

### 4.2 Phonetic dispersion

Phonetic dispersion refers to the principle that contrasting speech sounds should be distinct from one another in phonetic space (Martinet, 1955; Jakobson, 1968; Flemming, 1995, 2004). Most studies investigating this principle have focused on its validity within vowel systems, as we do here. While languages tend to have seemingly well-dispersed vowel inventories such as {/i/, /a/, /u/} (Joos, 1948; Stevens and Keyser, 2010), the actual phonetic realization of each vowel can vary substantially (Lindau and Wood, 1977; Disner, 1983). One prediction of dispersion is that the number of vowel categories in a language should be inversely related to the degree of per-category acoustic variation (Lindblom, 1986). Subsequent findings have cast doubt on this (Livijn, 2000; Recasens and Espinosa, 2009; Vaux and Samuels, 2015), but these studies have been limited by the number and diversity of languages investigated.

To investigate this, we measured the correlation between the number of vowel categories in a language and the degree of per-category variation, as measured by the *joint* entropy of (F1, F2) conditioned on the vowel category. We model $p(\text{F1}, \text{F2} \mid V)$ using a bivariate Gaussian for each vowel type $v$. We can then compute the joint conditional entropy under this model as $\text{H}(\text{F1}, \text{F2} \mid V) = \sum_v p(v)\, \text{H}(\text{F1}, \text{F2} \mid V = v) = \sum_v p(v) \frac{1}{2} \ln \det(2\pi e \Sigma_v)$, where $\Sigma_v$ is the covariance matrix for the model of vowel $v$.

Vowel inventory sizes per reading ranged from 4 to 20 vowels, with a median of 8. Both Spearman and Pearson correlations between entropy estimate and vowel inventory size across analyzed languages were small and not significant (Spearman $\rho = 0.11$, $p = 0.44$; Pearson $r = 0.11$, $p = 0.46$), corroborating previous accounts of the relationship described in Livijn (2000) and Vaux and Samuels (2015) with a larger number of languages—a larger vowel inventory does not necessarily imply more precision in vowel category production.[11]

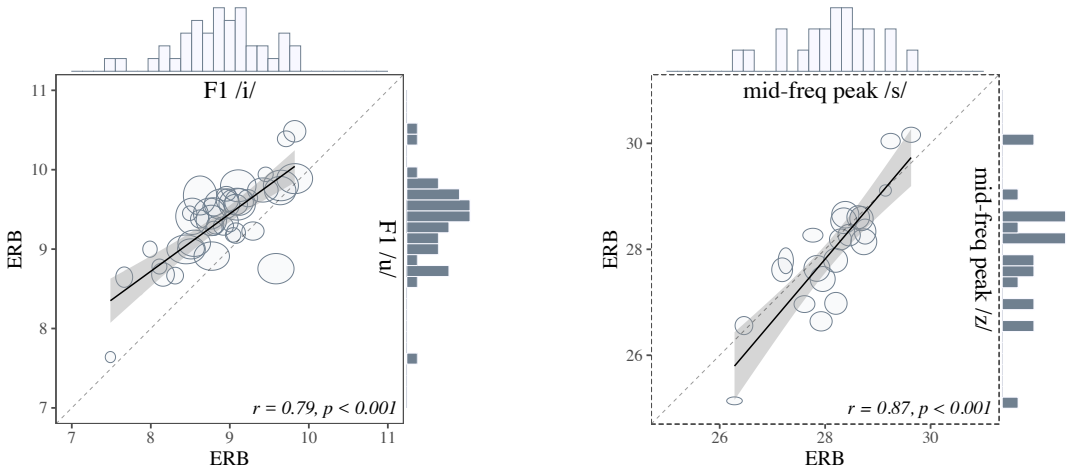
### 4.3 Phonetic uniformity

Previous work suggests that F1 is fairly uniform with respect to phonological height. Within a single language, the mean F1s of /e/ and /o/—which share a height—have been found to be correlated across speakers (*Yorkshire English*: Watt, 2000; *French*: Ménard et al., 2008; *Brazilian Portuguese*: Oushiro, 2019; *Dutch, English, French, Japanese, Portuguese, Spanish*: Schwartz and Ménard, 2019). Though it is physically possible for these vowels

[9] In the high-MCD languages, even the low-MCD utterances seem to be untrustworthy.

[10] The Equivalent Rectangular Bandwidth (ERB) scale is a psychoacoustic scale that better approximates human perception, which may serve as auditory feedback for the phonetic realization (Fletcher, 1923; Nearey, 1977; Zwicker and Terhardt, 1980; Glasberg and Moore, 1990). The precise equation comes from Glasberg and Moore (1990, Eq. 4).

[11] Since differential entropy is sensitive to parameterization, we also measured this correlation using formants in hertz, instead of in ERB, as ERB is on a logarithmic scale. This change did not the influence the pattern of results (Spearman $\rho = 0.12$, $p = 0.41$; Pearson $r = 0.13$, $p = 0.39$).




(a) **F1 of /i/–/u/ in ERB**

(b) **Mid-frequency peak of /s/–/z/ in ERB**

Figure 4: Correlations of mean F1 (ERB) between /i/ and /u/ and of mean mid-frequency peak (ERB) between /s/ and /z/. The paired segments share a relevant phonological feature specification that is approximated by the acoustic-phonetic measurement: vowel height by F1 and sibilant place by mid-frequency peak. Each reading is represented by an ellipsoid, centered on the paired means and shaped by $\frac{1}{10}$ of their respective standard deviations. The solid line reflects the best-fit linear regression line with standard error in gray shading; the dashed line shows the line of equality. Marginal histograms show the range of variation in the segment-specific means.

to differ in F1 realization, the correlations indicate a strong tendency for languages and individual speakers to yoke these two representations together.

Systematicity in the realization of sibilant place of articulation has also been observed across speakers of American English and Czech (Chodroff, 2017). Phonetic correlates of sibilant place strongly covary between /s/ and /z/, which share a [+anterior] place of articulation and are produced the alveolar ridge, and between /ʃ/ and /ʒ/, which share a [-anterior] place of articulation and are produced behind the alveolar ridge.

A principle of uniformity may account for these above findings. Uniformity here refers to a principle in which a distinctive phonological feature should have a consistent phonetic realization, within a language or speaker, across different segments with that feature (Keating, 2003; Chodroff et al., 2019). Similar principles posited in the literature include Maximal Use of Available Controls, in which a control refers to an integrated perceptual and motor phonetic target (Ménard et al., 2008), as well as a principle of gestural economy (Maddieson, 1995). Phonetic realization refers to the mapping from the abstract distinctive feature to an abstract phonetic target. We approximate this phonetic target via an acoustic-phonetic measurement, but we emphasize that the acoustic measurement is not necessarily a direct reflection of an underlying phonetic target (which could be an articulatory gesture, auditory goal, or perceptuo-motor representation of the sound). We make the simplifying assumption that the acoustic-phonetic formants (F1, F2) directly correspond to phonetic targets linked to the vowel features of height and backness.

More precisely, uniformity of a phonetic measure with respect to a phonological feature means that any two segments sharing that feature will tend to have approximately equal measurements in a given language, even when that value varies across languages. We can observe whether this is true by plotting the measures of the two segments against each other by language (e.g., Figure 4).

**Vowels.** As shown in Figure 4 and Table 3, the strongest correlations in mean F1 frequently reflected uniformity of height (e.g., high vowels /i/–/u/: $r$ = 0.79, $p < 0.001$, mid vowels /e/–/o/: $r$ = 0.62, $p < 0.01$).[12] Nevertheless, some vowel pairs that differed in height were also moderately correlated in mean F1 (e.g., /o/–/a/: $r = 0.66$, $p < 0.001$). Correlations of mean F1 were overall moderate in strength, regardless of the vowels' phonological specifications.

Correlations of mean F2 were also strongest among vowels with a uniform backness specification (e.g., back vowels /u/–/o/: $r = 0.69$, $p < 0.001$; front vowels /i/–/ɛ/: $r = 0.69$, $p < 0.05$; Table 4). The correlation between front tense vowels /i/ and /e/ was significant and in the ex-

[12] $p$-values are corrected for multiple comparisons using the Benjamini-Hochberg correction and a false discovery rate of 0.25 (Benjamini and Hochberg, 1995).

pected direction, but also slightly weaker than the homologous back vowel pair ($r = 0.41$, $p < 0.05$). Vowels differing in backness frequently had negative correlations, which could reflect influences of category crowding or language-/speaker-specific differences in peripheralization. We leave further exploration of those relationships to future study.

The moderate to strong F1 correlations among vowels with a shared height specification are consistent with expectations based on previous studies, and also with predictions of uniformity. Similarly, we find an expected correlation of F2 means for vowels with a shared height specification. The correlations of vowel pairs that were predicted to have significant correlations, but did not, tended to have small sample sizes ($< 14$ readings).

Nevertheless, the correlations are not perfect; nor are the patterns. For instance, the back vowel correlations of F2 are stronger than the front vowel correlations. While speculative, the apparent peripheralization of /i/ (as revealed in the negative F2 correlations) could have weakened the expected uniformity relation of /i/ with other front vowels. Future research should take into account additional influences of the vowel inventory composition, as well as articulatory or auditory factors for a more complete understanding of the structural forces in the phonetic realization of vowels.

**Sibilants.** The mean mid-frequency peak values for /s/ and /z/ each varied substantially across readings, and were also strongly correlated with one another ($r = 0.87$, $p < 0.001$; Figure 4).[13] This finding suggests a further influence of uniformity on the realization of place for /s/ and /z/, and the magnitude is comparable to previous correlations observed across American English and Czech speakers, in which $r$ was ≈0.90 (Chodroff, 2017).

## 5 Directions for Future Work

We hope our corpus may serve as a touchstone for further improvements in phonetic typology research and methodology. Here we suggest potential steps forward for known areas (§3.4) where this corpus could be improved:

**A Sentence alignments** were generated using Unitran, and could be improved with higher-quality G2P and verse-level text segmentation to standardize utterances across languages.

**B Consistent and comparable phoneme labels** are the ultimate goal. Concurrent work on universal phone recognition (Li et al., 2020) addresses this issue through a universal phone inventory constrained by language-specific PHOIBLE inventories (Moran and McCloy, 2019). However, free-decoding phones from speech alone is challenging. One exciting possibility is to use the orthography and audio jointly to guide semi-supervised learning of per-language pronunciation lexicons (Lu et al., 2013; Zhang et al., 2017).

**C Reliable quality assessment** for current methods remains an outstanding research question for many languages. For covered languages, using a universal label set to map additional high quality lexicons (e.g., hand-annotated lexicons) to the same label space as ours would enable direct label and alignment assessment through precision, recall, and PER.

**D Curating additional resources** beyond this corpus would improve coverage and balance, such as contributing additional Epitran modules. Additional readings exist for many languages on the original `bible.is` site and elsewhere. Annotations with speaker information are not available, but improved unsupervised speaker clustering may also support better analysis.

## 6 Conclusion

VoxClamantis v1.0 is the first large-scale corpus for phonetic typology, with extracted phonetic features for 635 typologically diverse languages. We present two case studies illustrating both the research potential and limitations of this corpus for investigation of phonetic typology at a large scale. We discuss several caveats for the use of this corpus and areas for substantial improvement. Nonetheless, we hope that directly releasing our alignments and token-level features enables greater research accessibility in this area. We hope this corpus will motivate and enable further developments in both phonetic typology and methodology for working with cross-linguistic speech corpora.

## Acknowledgments

The authors gratefully acknowledge Colin Wilson for his guidance and discussion on the topic, Florian Metze for resources, and Carlos Aguirre for helpful feedback.

[13] The magnitude of this correlation did not change when using hertz ($r = 0.86$, $p < 0.001$).

## References


Gopala Krishna Anumanchipalli, Kishore Prahallad, and Alan W. Black. 2011. Festvox: Tools for creation and analyses of large speech corpora. In *Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*.

Roy Becker-Kristal. 2010. *Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus*. Ph.D. thesis, University of California, Los Angeles.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Alan W. Black. 2006. CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. In *Proceedings of INTERSPEECH*.

Alan W. Black. 2019. CMU Wilderness Multilingual Speech Dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, Brighton, UK. IEEE.

Oliver Blacklock. 2004. *Characteristics of Variation in Production of Normal and Disordered Fricatives, Using Reduced-Variance Spectral Methods*. Ph.D. thesis, University of Southampton.

Paul Boersma and David Weenink. 2019. Praat: Doing phonetics by computer [computer program]. version 6.0.45.

Taehong Cho and Peter Ladefoged. 1999. Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, 27(2):207–229.

Eleanor Chodroff. 2017. *Structured Variation in Obstruent Production and Perception*. Ph.D. thesis, Johns Hopkins University.

Eleanor Chodroff, Alessandra Golden, and Colin Wilson. 2019. Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, 145(1):EL109–EL115.

Sandra Ferrari Disner. 1983. *Vowel Quality: The Relation between Universal and Language-specific Factors*. Ph.D. thesis, UCLA.

Gary F. Simons Eberhard, David M. and Charles D. Fennig, editors. 2020. *Ethnologue: Languages of the world*, 23 edition. SIL international. Online version: http://www.ethnologue.com.

Arvo Eek and Einar Meister. 1994. Acoustics and perception of Estonian vowel types. *Phonetic Experimental Research*, XVIII:146–158.

Olle Engstrand and Una Cunningham-Andersson. 1988. Iris - a data base for cross-linguistic phonetic research.

Edward S. Flemming. 1995. *Auditory Representations in Phonology*. Ph.D. thesis, UCLA.

Edward S. Flemming. 2004. Contrast and perceptual distinctiveness. In Bruce Hayes, R. Kirchner, and Donca Steriade, editors, *The Phonetic Bases of Phonological Markedness*, 1968, pages 232–276. University Press, Cambridge, MA.

Harvey Fletcher. 1923. Physical measurements of audition and their bearing on the theory of hearing. *Journal of the Franklin Institute*, 196(3):289–326.

Karen Forrest, Gary Weismer, Paul Milenkovic, and Ronald N. Dougall. 1988. Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, 84(1):115–123.

Brian R. Glasberg and Brian C.J. Moore. 1990. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138.

Matthew Gordon and Timo Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, 3(1).

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya D. McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the SIGMORPHON Workshop*.

Mary Harper. 2011. The IARPA Babel multilingual speech database. Accessed: 2020-05-01.

Arthur S. House and Kenneth N. Stevens. 1956. Analog studies of the nasalization of vowels. *The Journal of Speech and Hearing Disorders*, 21(2):218–232.

Roman Jakobson. 1968. *Child Language, Aphasia and Phonological Universals*. Mouton Publishers.

Allard Jongman, Ratree Wayland, and Serena Wong. 2000. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263.

Martin Joos. 1948. Acoustic phonetics. *Language*, 24(2):5–136.

Patricia A. Keating. 2003. Phonetic and other influences on voicing contrasts. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 20–23, Barcelona, Spain.

Laura Koenig, Christine H. Shadle, Jonathan L. Preston, and Christine R. Mooshammer. 2013. Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech, Language, and Hearing Research*, 56(4):1175–1189.

John Kominek, Tanja Schultz, and Alan W. Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Spoken Languages Technologies for Under-Resourced Languages*.

Peter Ladefoged, Richard Harshman, Louis Goldstein, and Lloyd Rice. 1978. Generating vocal tract shapes from formant frequencies. *The Journal of the Acoustical Society of America*, 64(4):1027–1035.

Peter Ladefoged and Keith Johnson. 2014. *A Course in Phonetics*. Nelson Education.

Peter Ladefoged and Ian Maddieson. 2007. The UCLA phonetics lab archive.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation mining with WikiPron. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA). Resources downloadable from https://github.com/kylebgorman/wikipron.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W. Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Mona Lindau and Patricia Wood. 1977. Acoustic vowel spaces. *UCLA Working Papers in Phonetics*, 38:41–48.

Björn Lindblom. 1986. Phonetic universals in vowel systems. In John J. Ohala and Jeri Jaeger, editors, *Experimental Phonology*, pages 13–44. Academic Press, Orlando.

Björn Lindblom and Johan Sundberg. 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, 50(4B):1166–1179.

Peder Livijn. 2000. Acoustic distribution of vowels in differently sized inventories–hot spots or adaptive dispersion. *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (PERILUS)*, 11.

Liang Lu, Arnab Ghoshal, and Steve Renals. 2013. Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 374–379. IEEE.

Ian Maddieson. 1995. Gestural economy. In *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, Sweden.

André Martinet. 1955. *Économie Des Changements Phonétiques: Traité de Phonologie Diachronique*, volume 10. Bibliotheca Romanica.

Lucie Ménard, Jean-Luc Schwartz, and Jérôme Aubin. 2008. Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, 50:14–28.

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

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Terrance M. Nearey. 1977. *Phonetic Feature Systems for Vowels*. Ph.D. thesis, University of Alberta. Reprinted 1978 by Indiana University Linguistics Club.

Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose. 2016. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907–938.

Livia Oushiro. 2019. Linguistic uniformity in the speech of Brazilian internal migrants in a dialect contact situation. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, pages 686–690, Melbourne, Australia. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Kishore Prahallad, Alan W. Black, and Ravishankhar Mosur. 2006. Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1. IEEE.

Ting Qian, Kristy Hollingshead, Su-youn Yoon, Kyoung-young Kim, and Richard Sproat. 2010. A Python toolkit for universal transliteration. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Karim Rahim and Wesley S. Burr. 2017. multitaper: Multitaper spectral analysis. *R package version 1.0-14*.

Daniel Recasens and Aina Espinosa. 2009. Dispersion and variability in Catalan five and six peripheral vowel systems. *Speech Communication*, 51(3):240–258.

Tanja Schultz. 2002. GlobalPhone: A multilingual speech and text database developed at Karlsruhe University. In *Seventh International Conference on Spoken Language Processing*, pages 345–348, Denver, CO.

Jean-Luc Schwartz and Lucie Ménard. 2019. Structured idiosyncrasies in vowel systems. *OSF Preprints*.

Christine H. Shadle, Wei-rong Chen, and D. H. Whalen. 2016. Stability of the main resonance frequency of fricatives despite changes in the first spectral moment. *The Journal of the Acoustical Society of America*, 140(4):3219–3220.

Kenneth N. Stevens and Samuel J. Keyser. 2010. Quantal theory, enhancement and overlap. *Journal of Phonetics*, 38(1):10–19.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, pages 901–904.

Bert Vaux and Bridget Samuels. 2015. Explaining vowel systems: Dispersion theory vs natural selection. *Linguistic Review*, 32(3):573–599.

Dominic J. L. Watt. 2000. Phonetic parallels between the close-mid vowels of Tyneside English: Are they internally or externally motivated? *Language Variation and Change*, 12(1):69–101.

John C. Wells. 1995/2000. Computer-coding the IPA: A proposed extension of SAMPA.

D .H. Whalen and Andrea G. Levitt. 1995. The universality of intrinsic F0 of vowels. *Journal of Phonetics*, 23:349–366.

Matthew Wiesner, Oliver Adams, David Yarowsky, Jan Trmal, and Sanjeev Khudanpur. 2019. Zero-shot pronunciation lexicons for cross-language acoustic model transfer. In *Proceedings of IEEE Association for Automatic Speech Recognition and Understanding (ASRU)*.

Xiaohui Zhang, Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur. 2017. Acoustic data-driven lexicon learning based on a greedy pronunciation selection framework. *arXiv preprint arXiv:1706.03747*.

Eberhard Zwicker and Ernst Terhardt. 1980. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68(5):1523–1525.

## A Pairwise Correlations between Vowel Formant Measures (§4 Case Studies)

Table 3 and Table 4 respectively show Pearson correlations of mean F1 and mean F2 in ERB between vowels that appear in at least 10 readings. As formalized in the present analysis, phonetic uniformity predicts strong correlations of mean F1 among vowels with a shared height specification, and strong correlations of mean F2 among vowels with a shared backness specification. The respective "Height" and "Backness" columns in Table 3 and Table 4 indicate whether the vowels in each pair match in their respective specifications. $p$-values are corrected for multiple comparisons using the Benjamini-Hochberg correction and a false discovery rate of 0.25 (Benjamini and Hochberg, 1995). Significance is assessed at $\alpha = 0.05$ following the correction for multiple comparisons; rows that appear in gray have correlations that are not significant according to this threshold.

| V1 | V2 | Height | # Readings | $r$ | $p$ |
|---|---|---|---|---|---|
| /i/ | /iː/ | ✓ | 12 | 0.81 | 0.006 |
| /eː/ | /oː/ | ✓ | 10 | 0.81 | 0.015 |
| /i/ | /u/ | ✓ | 40 | 0.79 | 0.000 |
| /ɛ/ | /ɔ/ | ✓ | 11 | 0.68 | 0.053 |
| /o/ | /a/ | | 37 | 0.66 | 0.000 |
| /iː/ | /oː/ | | 11 | 0.65 | 0.070 |
| /iː/ | /uː/ | ✓ | 12 | 0.64 | 0.061 |
| /e/ | /o/ | ✓ | 35 | 0.62 | 0.001 |
| /e/ | /u/ | | 36 | 0.59 | 0.001 |
| /e/ | /a/ | | 34 | 0.58 | 0.002 |
| /u/ | /ə/ | | 12 | 0.58 | 0.105 |
| /iː/ | /eː/ | | 11 | 0.58 | 0.118 |
| /i/ | /e/ | | 38 | 0.54 | 0.002 |
| /ɛ/ | /a/ | | 12 | 0.54 | 0.127 |
| /u/ | /o/ | | 38 | 0.49 | 0.007 |
| /ɛ/ | /u/ | | 14 | 0.49 | 0.135 |
| /i/ | /o/ | | 39 | 0.46 | 0.011 |
| /e/ | /ɛ/ | ✓ | 12 | 0.46 | 0.204 |
| /u/ | /a/ | | 37 | 0.42 | 0.027 |
| /iː/ | /e/ | | 11 | 0.42 | 0.288 |
| /u/ | /uː/ | ✓ | 10 | 0.41 | 0.334 |
| /iː/ | /u/ | ✓ | 11 | 0.33 | 0.430 |
| /iː/ | /a/ | | 11 | 0.28 | 0.496 |
| /i/ | /a/ | | 39 | 0.27 | 0.173 |
| /i/ | /ɛ/ | | 14 | 0.24 | 0.496 |
| /iː/ | /o/ | | 13 | 0.19 | 0.624 |
| /i/ | /ə/ | | 13 | 0.10 | 0.785 |
| /u/ | /ɔ/ | | 12 | 0.09 | 0.785 |
| /ɛ/ | /o/ | ✓ | 13 | -0.09 | 0.785 |
| /e/ | /ɔ/ | ✓ | 10 | -0.12 | 0.785 |
| /uː/ | /o/ | | 10 | -0.12 | 0.785 |
| /i/ | /ɔ/ | | 11 | -0.42 | 0.288 |
| /o/ | /ə/ | ✓ | 11 | -0.51 | 0.173 |
| /ə/ | /a/ | | 11 | -0.90 | 0.001 |

Table 3: Pearson correlations ($r$) of **mean F1** in ERB between vowel categories.

| V1 | V2 | Backness | # Readings | $r$ | $p$ |
|---|---|---|---|---|---|
| /e/ | /ɛ/ | ✓ | 12 | 0.77 | 0.019 |
| /u/ | /uː/ | ✓ | 10 | 0.77 | 0.037 |
| /i/ | /iː/ | ✓ | 12 | 0.70 | 0.038 |
| /u/ | /o/ | ✓ | 38 | 0.69 | 0.000 |
| /i/ | /ɛ/ | ✓ | 14 | 0.69 | 0.031 |
| /uː/ | /o/ | ✓ | 10 | 0.62 | 0.130 |
| /u/ | /ə/ | | 12 | 0.60 | 0.107 |
| /u/ | /ɔ/ | ✓ | 12 | 0.52 | 0.168 |
| /i/ | /e/ | ✓ | 38 | 0.41 | 0.038 |
| /ɛ/ | /a/ | | 12 | 0.32 | 0.519 |
| /o/ | /a/ | | 37 | 0.30 | 0.159 |
| /eː/ | /oː/ | | 10 | 0.27 | 0.666 |
| /e/ | /a/ | | 34 | 0.24 | 0.339 |
| /o/ | /ə/ | | 11 | 0.21 | 0.724 |
| /ə/ | /a/ | ✓ | 11 | 0.16 | 0.830 |
| /iː/ | /e/ | ✓ | 11 | 0.11 | 0.911 |
| /i/ | /a/ | | 39 | 0.06 | 0.911 |
| /iː/ | /eː/ | ✓ | 11 | 0.06 | 0.965 |
| /e/ | /o/ | | 35 | 0.01 | 0.965 |
| /u/ | /a/ | | 37 | 0.00 | 0.985 |
| /ɛ/ | /ɔ/ | | 11 | -0.03 | 0.965 |
| /iː/ | /a/ | | 11 | -0.04 | 0.965 |
| /ɛ/ | /o/ | | 13 | -0.04 | 0.965 |
| /e/ | /u/ | | 36 | -0.12 | 0.666 |
| /ɛ/ | /u/ | | 14 | -0.22 | 0.666 |
| /i/ | /ə/ | | 13 | -0.23 | 0.666 |
| /iː/ | /oː/ | | 11 | -0.42 | 0.345 |
| /i/ | /o/ | | 39 | -0.48 | 0.017 |
| /iː/ | /o/ | | 13 | -0.52 | 0.149 |
| /i/ | /u/ | | 40 | -0.55 | 0.003 |
| /i/ | /ɔ/ | | 11 | -0.63 | 0.107 |
| /e/ | /ɔ/ | | 10 | -0.65 | 0.107 |
| /iː/ | /u/ | | 11 | -0.80 | 0.019 |
| /iː/ | /uː/ | | 12 | -0.83 | 0.009 |

Table 4: Pearson correlations ($r$) of **mean F2** in ERB between vowel categories.

## B Distributions of Unitran Segment Accuracy (§3.1.3 Quality Measures)

Here we evaluate the quality of the Unitran dataset in more detail. The goal is to explore the variation in the quality of the labeled Unitran segments across different languages and phoneme labels. This evaluation includes only readings in high-resource languages, where we have not only the aligned Unitran pronunciations but also aligned high-resource pronunciations (Epitran or WikiPron) against which to evaluate them. The per-token statistics used to calculate these plots are included in the corpus release to enable closer investigation of individual phonemes than is possible here.

### B.1 Unitran Pronunciation Accuracy

First, in Figures 5 and 6, we consider whether Unitran's utterance pronunciations are accurate without looking at the audio. For each utterance, we compute the unweighted Levenshtein alignment between the Unitran pronunciation of the utterance and the high-resource pronunciation. For each reading, we then score the percentage of Unitran 'phoneme' tokens that were aligned to high-resource 'phoneme' tokens with exactly the same label.[14] We can see in Figure 6 that many labels are highly accurate in many readings while being highly inaccurate in many others. Some labels are noisy in some readings.[15]

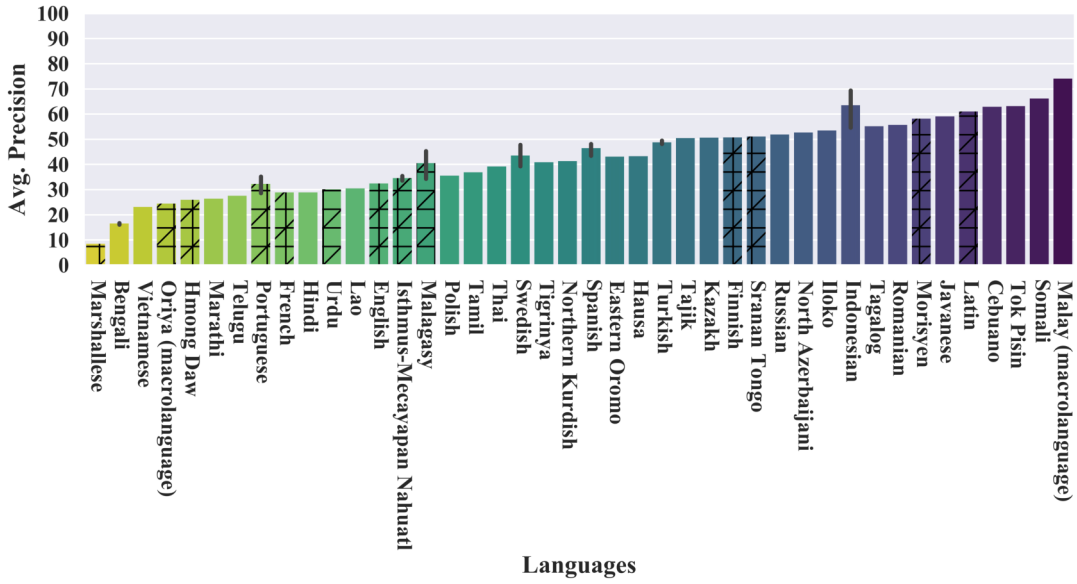



Figure 5: **Unitran pronunciation accuracy per language**, evaluated by Levenshtein alignment to WikiPron pronunciations (hatched bars) or Epitran pronunciations (plain bars). Where a language has multiple readings, error bars show the min and max across those readings.

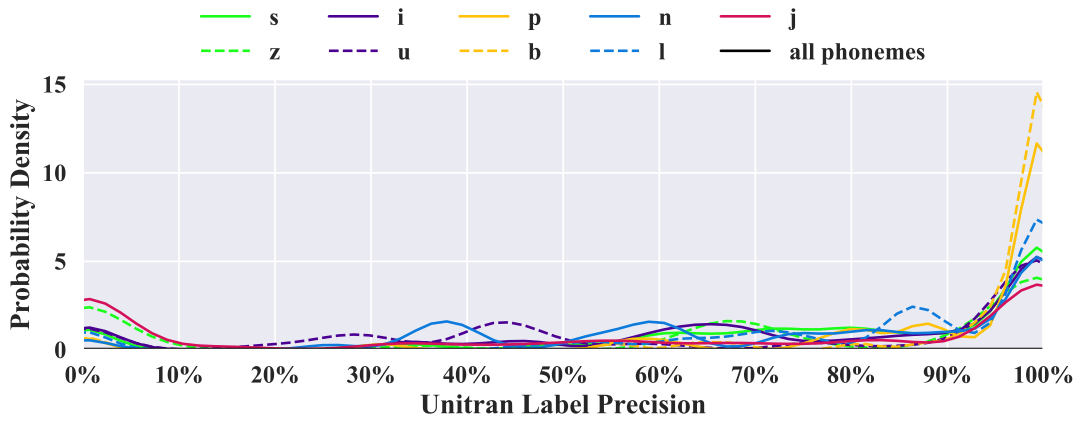



Figure 6: **Unitran pronunciation accuracy per language, for selected phonemes.** Accuracy is evaluated by Levenshtein alignment as in Figure 5. Each curve is a kernel density plot with integral 1. For the /z/ curve, the integral between 80% and 100% (for example) is the estimated probability that in a high-resource language drawn uniformly at random, the fraction of Unitran /z/ segments that align to high-resource /z/ segments falls in that range. The 'all' curve is the same, but now the uniform draw is from all pairs of (high-resource language, Unitran phoneme used in that language).

[14] By contrast, PER in Table 1 aligns at the word level rather than the utterance level, uses the number of symmetric alignment errors (insertions + deletions + substitutions) rather than the number of correct Unitran phonemes, and normalizes by the length of the high-resource 'reference' pronunciation rather than by the length of the Unitran pronunciation.

[15] Note that as §3.1.3 points out, it may be unfair to require exact match of labels, since annotation schemes vary.)

### B.2 Unitran Segment Label Accuracy

In Figures 7 and 8, we ask the same question again, but making use of the audio data. The match for each Unitran segment is now found not by Levenshtein alignment, but more usefully by choosing the high-resource segment with the closest midpoint. For each reading, we again score the percentage of Unitran 'phoneme' tokens whose aligned high-resource 'phoneme' tokens have exactly the same label. Notice that phonemes that typically had high accuracy in Figure 6, such as /p/ and /b/, now have far more variable accuracy in Figure 8, suggesting difficulty in aligning the Unitran pronunciations to the correct parts of the audio.

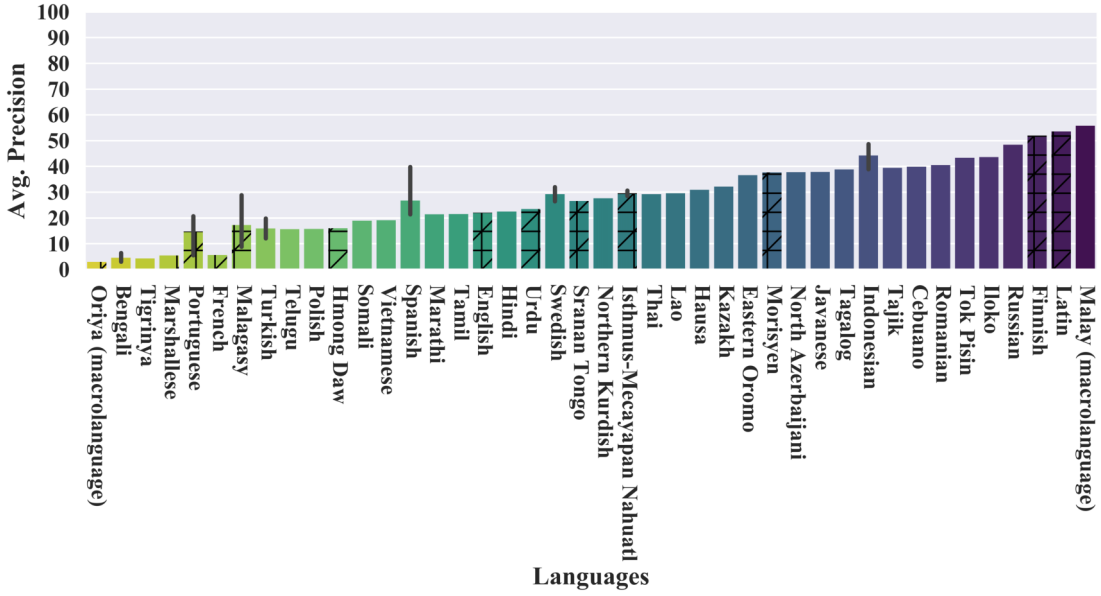



Figure 7: **Unitran pronunciation accuracy per language**, as in Figure 5 but with audio midpoint alignment in place of Levenshtein alignment.

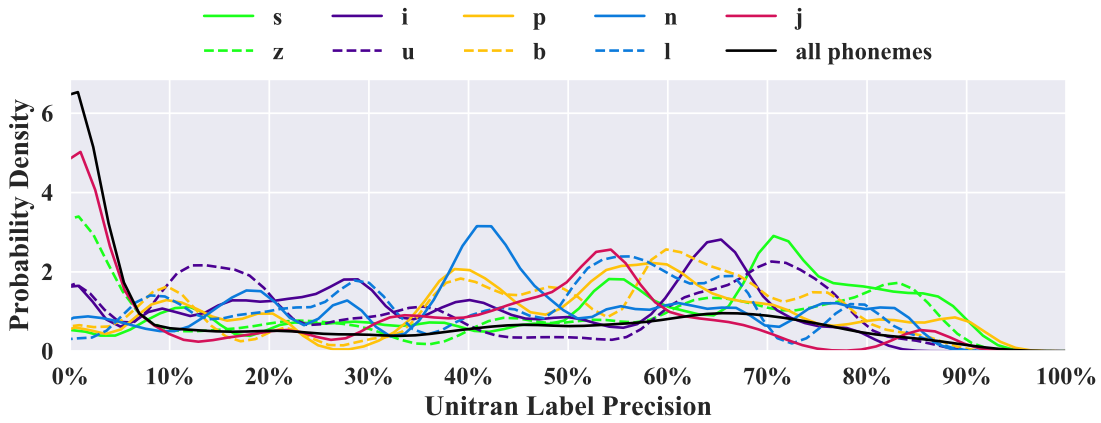



Figure 8: **Unitran pronunciation accuracy per language, for selected phonemes,** as in Figure 6 but with audio midpoint alignment in place of Levenshtein alignment.

### B.3 Unitran Segment Boundary Accuracy

Finally, in Figures 9 and 10, we measure whether Unitran segments with the "correct" label also have the "correct" time boundaries, where "correctness" is evaluated against the corresponding segments obtained using Epitran or WikiPron+G2P.

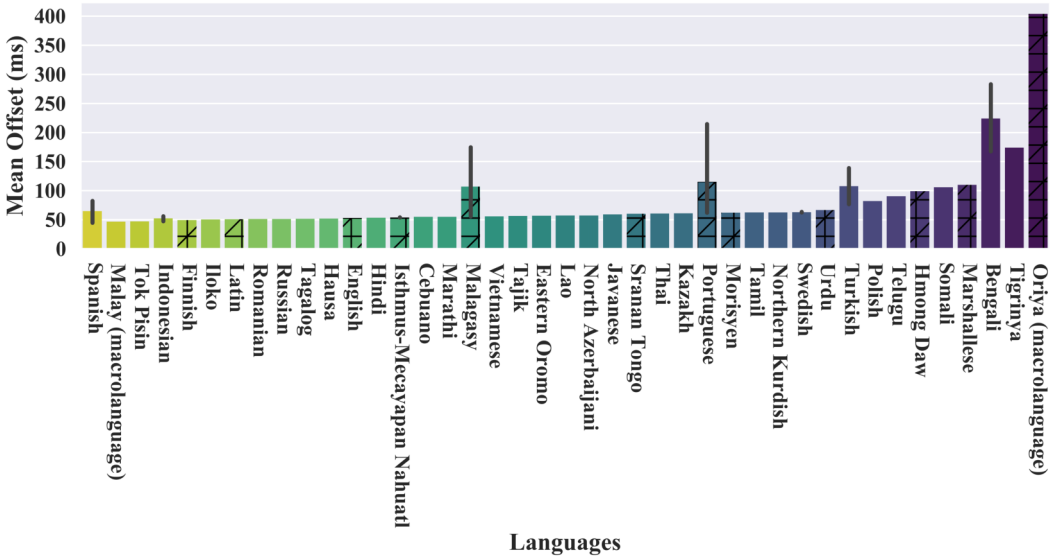



Figure 9: **Mean error per language in the temporal boundaries of Unitran segments.**. Each Unitran segment is evaluated against the WikiPron segment (hatched bars) or Epitran segment (plain bars) with the closest midpoint, as if the latter were truth. The error of a segment is the absolute offset of the left boundary plus the absolute offset of the right boundary. Only segments where the Unitran label matches the Epitran/WikiPron label are included in the average. Where a language has multiple readings, error bars show the min and max across those readings.

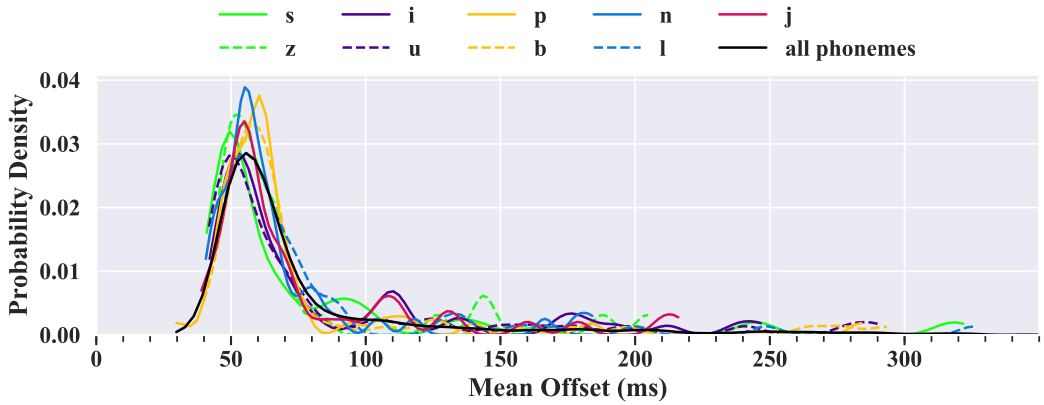



Figure 10: **Mean error per language in the temporal boundaries of Unitran segments, for selected phonemes.** Each curve is a kernel density plot with integral 1. For the /z/ curve, the integral between 50ms and 100ms (for example) is the estimated probability that in a high-resource language drawn uniformly at random, the Unitran /z/ segments whose corresponding Epitran or WikiPron segments are also labeled with /z/ have mean boundary error in that range. Small bumps toward the right correspond to individual languages where the mean error of /z/ is unusually high. The 'all' curve is the same, but now the uniform draw is from all pairs of (high-resource language, Unitran phoneme used in that language). The boundary error of a segment is evaluated as in Figure 9.

## C WikiPron Grapheme-to-Phoneme (G2P) Accuracy (§3.1.3 Quality Measures)

For each language where we used WikiPron, Table 5 shows the phoneme error rate (PER) of Phonetisaurus G2P models trained on WikiPron entries, as evaluated on held-out WikiPron entries. This is an estimate of how accurate our G2P-predicted pronunciations are on out-of-vocabulary words, insofar as those are distributed similarly to the in-vocabulary words. (It is possible, however, that out-of-vocabulary words such as Biblical names are systematically easier or harder for the G2P system to pronounce, depending on how they were transliterated.)

The same G2P configuration was used for all languages, with the hyperparameter settings shown in Table 6. (`seq1_max` and `seq2_max` describe how many tokens in the grapheme and phoneme sequences can align to each other.). These settings were tuned on SIGMORPHON 2020 Task 1 French, Hungarian, and Korean data (Gorman et al., 2020), using 20 random 80/20 splits.

| **ISO 639-3** | **fin** | **lat** | **nhx** | **srn** | **mah** | **por-po** | **mfe** | **mww** | **por-bz** | **eng** | **khm** | **mlg** | **ori** | **ban** | **urd** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Train size | 41741 | 34181 | 126 | 157 | 813 | 9633 | 203 | 227 | 10077 | 54300 | 3016 | 114 | 211 | 172 | 704 |
| **PER** | 0.8 | 2.4 | 4.1 | 4.6 | 9.6 | 10.1 | 10.7 | 10.8 | 11.4 | 14.5 | 15.5 | 15.8 | 16.1 | 19.5 | 26.7 |
| | ±0.02 | ±0.04 | ±1.02 | ±0.76 | ±0.41 | ±0.11 | ±1.2 | ±1.29 | ±0.16 | ±0.06 | ±0.38 | ±1.44 | ±1.13 | ±1.35 | ±0.60 |

Table 5: WikiPron G2P Phone Error Rate (PER) calculated treating WikiPron annotations as ground-truth. We perform 20 trials with random 80/20 splits per language, and report PER averaged across trials with 95% confidence intervals for each language.

| **Phonetisaurus Alignment Hyperparameters** | `seq1_max` | `seq2_max` | `seq1_del` | `seq2_del` | `grow` | `max EM iterations` | |
|---|---|---|---|---|---|---|---|
| | 1 | 3 | True | True | True | 11 | |
| **Graphone Language Model Hyperparameters** | `n-gram order` | `LM type` | `discounting` | `gt2min` | `gt3min` | `gt4min` | `gt5min` |
| | 5 | max-ent | Kneser-Ney | 2 | 2 | 3 | 4 |

Table 6: Table of final G2P hyperparameter settings. Alignment parameters not listed here for `phonetisaurus-align` use the default values. The language model was trained using SRILM (Stolcke, 2002) `ngram-count` using default values except for those listed above.

## D Retention Statistics (§4.1 Data Filtering)

Table 7 shows what percentage of tokens would be retained after various methods are applied to filter out questionable tokens from the readings used in §4.1. In particular, the rightmost column shows the filtering that was actually used in §4.1. We compute statistics for each reading separately; in each column we report the minimum, median, mean, and maximum statistics over the readings. The top half of the table considers vowel tokens (for the vowels in Appendix A); the bottom half considers sibilant tokens (/s/ and /z/).

On the left side of the table, we consider three filtering techniques for Unitran alignments. **Midpoint** retains only the segments whose labels are "correct" according to the midpoint-matching methods of Appendix B. **MCD** retains only those utterances with MCD $< 6$. **Outlier** removes tokens that are outliers according to the criteria described in §4.1. Finally, **AGG.** is the aggregate retention rate retention rate after all three methods are applied in order.

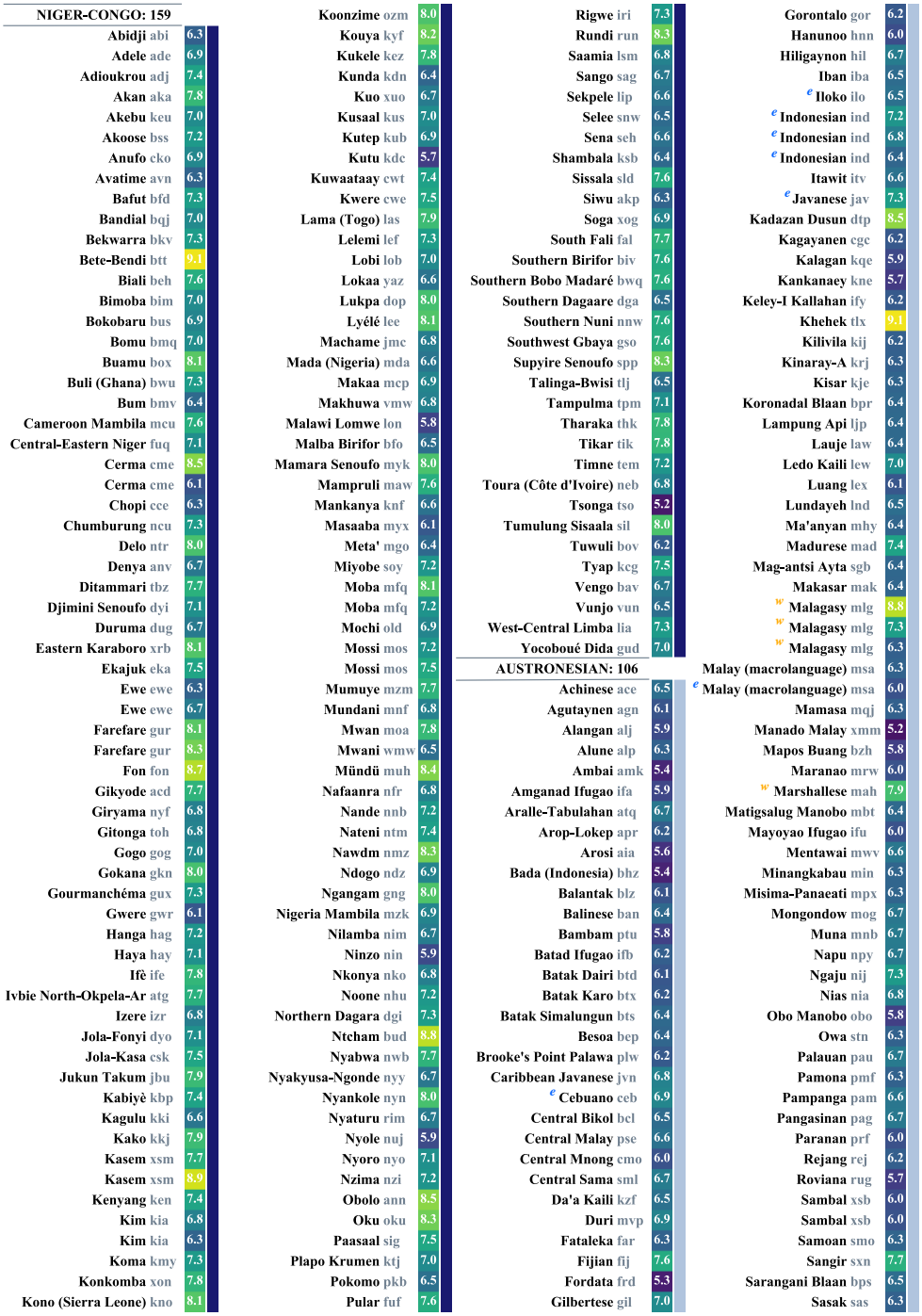
On the right side of the table, we consider the same filtering techniques for the high-resource alignments that we actually use, with the exception of **Midpoint**, as here we have no higher-quality annotation to match against.

| | | **Unitran Alignments** | | | | | **High-Resource Alignments** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **# Tokens** | **Midpoint** | **MCD** | **Outlier** | **AGG.** | **# Tokens** | **MCD** | **Outlier** | **AGG.** |
| *Vowels* | Min | 50,132 | 2% | 42% | 83% | 1% | 61,727 | 42% | 84% | 37% |
| | Median | 21,5162 | 23% | 88% | 90% | 16% | 232,059 | 88% | 90% | 79% |
| | Mean | 23,9563 | 25% | 81% | 89% | 20% | 223,815 | 81% | 90% | 73% |
| | Max | 662,813 | 65% | 100% | 93% | 60% | 468,864 | 100% | 93% | 93% |
| | # Readings | 49 | 46 | 48 | 49 | 45 | 49 | 48 | 49 | 48 |
| *Sibilants* | Min | 7,198 | 10% | 42% | 89% | 13% | 7184 | 44% | 91% | 43% |
| | Median | 28,690 | 70% | 87% | 97% | 59% | 27569 | 87% | 97% | 85% |
| | Mean | 30,025 | 63% | 80% | 95% | 56% | 27083 | 81% | 96% | 79% |
| | Max | 63,573 | 89% | 100% | 98% | 79% | 45,290 | 100% | 99% | 96% |
| | # Readings | 36 | 26 | 35 | 36 | 19 | 25 | 22 | 25 | 22 |

Table 7: Summary of quality measure retention statistics for **vowels** and **sibilants** over unique readings with reading-level MCD $< 8$ for Unitran and high-resource alignments.

## E All VoxClamantis v1.0 Languages

All 635 languages from 690 readings are presented here with their language family, ISO 639-3 code, and mean utterance alignment quality in Mel Cepstral Distortion (MCD) from Black (2019). Languages for which we release Epitran and/or WikiPron alignments in addition to Unitran alignments are marked with $^{e}$ and $^{w}$ respectively. MCD ranges from purple (*low*), blue–green (*mid*), to yellow (*high*). Lower MCD typically corresponds to better audio-text utterance alignments and higher quality speech synthesis, but judgments regarding distinctions between languages may be subjective. ISO 639-3 is not intended to provide identifiers for *dialects* or other *sub-language* variations, which may be present here where there are multiple readings for one ISO 639-3 code. We report the most up-to-date language names from the ISO 639-3 schema (Eberhard and Fennig, 2020). Language names and codes in many schema could be pejorative and outdated, but where language codes cannot be easily updated, language names can and often are.



| | |
|---|---|
| Sudest tgo | 6.1 |
| Sundanese sun | 6.9 |
| [e] Tagalog tgl | 6.5 |
| Tangoa tgp | 7.1 |
| Termanu twu | 6.1 |
| Tombonuo txa | 7.2 |
| Toraja-Sa'dan sda | 6.3 |
| Tuwali Ifugao ifk | 6.7 |
| Uma ppk | 6.7 |
| Western Bukidnon Mano mbb | 6.6 |
| Western Tawbuid twb | 6.0 |
| **AFRO-ASIATIC: 45** | |
| Bana bcw | 7.2 |
| Daasanach dsh | 6.5 |
| Daba dbq | |
| Dangaléat daa | 7.0 |
| Dawro dwr | 8.3 |
| [e] Eastern Oromo hae | 6.5 |
| Egyptian Arabic arz | 7.4 |
| Gamo gmv | 7.2 |
| Gen gej | 7.3 |
| Gofa gof | 6.5 |
| Gofa gof | 8.2 |
| Gude gde | 7.3 |
| Hamer-Banna amf | 6.5 |
| [e] Hausa hau | 5.7 |
| Hdi xed | 7.5 |
| Iraqw irk | 8.4 |
| Kabyle kab | 7.4 |
| Kafa kbr | 7.3 |
| Kambaata ktb | 6.9 |
| Kamwe hig | 7.8 |
| Kera ker | 7.3 |
| Kimré kqp | 6.7 |
| Konso kxc | 6.6 |
| Koorete kqy | 7.2 |
| Lele (Chad) lln | 6.8 |
| Male (Ethiopia) mdy | 6.6 |
| Marba mpg | 7.7 |
| Mbuko mqb | 7.9 |
| Merey meq | 8.1 |
| Mesopotamian Arabic acm | 8.3 |
| Mofu-Gudur mif | 8.0 |
| Muyang muy | 6.6 |
| Mwaghavul sur | 7.1 |
| North Mofu mfk | 7.0 |
| Parkwa pbi | 6.9 |
| Pévé lme | 7.7 |
| Sebat Bet Gurage sgw | 6.6 |
| [e] Somali som | 8.3 |
| Standard Arabic arb | 7.9 |
| Sudanese Arabic apd | 8.0 |
| Tachelhit shi | 5.0 |
| Tamasheq taq | 7.1 |
| [e] Tigrinya tir | 6.6 |
| Tumak tmc | 6.6 |
| Wandala mfi | 7.9 |
| **MAYAN: 42** | |
| Achi acr | 6.2 |
| Aguacateco agu | 5.8 |
| Chol ctu | 7.0 |
| Chortí caa | 6.4 |
| Chuj cac | 7.5 |
| Chuj cac | 6.7 |
| Huastec hus | 6.1 |
| Ixil ixl | 5.8 |
| Ixil ixl | 6.5 |
| Ixil ixl | 7.6 |
| K'iche' quc | 6.6 |
| K'iche' quc | 7.6 |
| K'iche' quc | 6.4 |
| K'iche' quc | 6.3 |
| K'iche' quc | 6.4 |
| K'iche' quc | 7.1 |
| Kaqchikel cak | 6.1 |
| Kaqchikel cak | |
| Kaqchikel cak | 5.5 |
| Kaqchikel cak | 6.8 |
| Kaqchikel cak | 7.0 |
| Kaqchikel cak | 7.9 |
| Kekchí kek | 6.5 |
| Kekchí kek | 6.3 |
| Mam mam | 6.3 |
| Mam mam | 6.7 |
| Mam mam | 7.3 |
| Mam mam | 7.1 |
| Mopán Maya mop | 7.0 |
| Popti' jac | 7.1 |
| Popti' jac | 6.3 |
| Poqomchi' poh | 6.5 |
| Poqomchi' poh | 5.3 |
| Q'anjob'al kjb | 6.8 |
| Tektiteko ttc | 6.0 |
| Tz'utujil tzj | 6.8 |
| Tzeltal tzh | 6.0 |
| Tzeltal tzh | 6.5 |
| Tzotzil tzo | 6.2 |
| Tzotzil tzo | 7.1 |
| Western Kanjobal knj | 6.8 |
| Yucateco yua | 7.0 |
| **INDO-EUROPEAN: 40** | |
| Albanian sqi | 7.0 |
| Awadhi awa | 7.4 |
| [e] Bengali ben | 8.1 |
| [e] Bengali ben | 8.1 |
| [e] Bengali ben | 8.1 |
| Caribbean Hindustani hns | 7.0 |
| Chhattisgarhi hne | 6.6 |
| Dari prs | 6.9 |
| [w] English eng | 6.9 |
| Fiji Hindi hif | 6.9 |
| French fra | 8.2 |
| French fra | 8.5 |
| [e] Hindi hin | 6.5 |
| Iranian Persian pes | 7.3 |
| [w] Latin lat | 5.8 |
| Magahi mag | 6.4 |
| Maithili mai | 7.2 |
| Malvi mup | 6.5 |
| [e] Marathi mar | 6.8 |
| [e] Northern Kurdish kmr | 7.0 |
| [w] Oriya (macrolanguage) ori | 7.6 |
| Ossetian oss | 6.3 |
| [e] Polish pol | 7.7 |
| [w] Portuguese por | 7.2 |
| [w] Portuguese por | 7.6 |
| [w] Portuguese por | 8.2 |
| [w] Portuguese por | 7.9 |
| [e] Romanian ron | 6.8 |
| [e] Russian rus | 5.6 |
| Sinte Romani rmo | 6.6 |
| [e] Spanish spa | 6.2 |
| [e] Spanish spa | 7.9 |
| [e] Spanish spa | 7.8 |
| [e] Spanish spa | 7.9 |
| [e] Spanish spa | 6.7 |
| [e] Swedish swe | 6.9 |
| [e] Swedish swe | 6.1 |
| [e] Tajik tgk | 6.8 |
| [w] Urdu urd | 6.6 |
| Vlax Romani rmy | 6.8 |
| **OTO-MANGUEAN: 27** | |
| Atatláhuca Mixtec mib | 6.2 |
| Ayutla Mixtec miy | 6.1 |
| Central Mazahua maz | 7.0 |
| Chicahuaxtla Triqui trs | 6.0 |
| Diuxi-Tilantongo Mixt xtd | 6.5 |
| Jalapa De Díaz Mazate maj | 8.3 |
| Jamiltepec Mixtec mxt | 7.4 |
| Lalana Chinantec cnl | 7.4 |
| Lealao Chinantec cle | 6.6 |
| Magdalena Peñasco Mix xtm | 5.6 |
| Mezquital Otomi ote | 6.8 |
| Nopala Chatino cya | 8.8 |
| Ozumacín Chinantec chz | 7.7 |
| Peñoles Mixtec mil | 6.7 |
| Pinotepa Nacional Mix mio | 6.0 |
| San Jerónimo Tecóatl maa | 7.7 |
| San Jerónimo Tecóatl maa | 7.9 |
| San Juan Atzingo Popo poe | 6.5 |
| San Marcos Tlacoyalco pls | 5.9 |
| San Pedro Amuzgos Amu azg | 7.2 |
| Santa María Zacatepec mza | 6.3 |
| Sochiapam Chinantec cso | 6.1 |
| Southern Puebla Mixte mit | 6.5 |
| Tepetotutla Chinantec cnt | 7.3 |
| Tezoatlán Mixtec mxb | 6.0 |
| Usila Chinantec cuc | 6.7 |
| Yosondúa Mixtec mpm | 6.7 |
| **SINO-TIBETAN: 24** | |
| Achang acn | 6.1 |
| Akeu aeu | 6.9 |
| Akha ahk | 7.0 |
| Bawm Chin bgr | 6.8 |
| Eastern Tamang taj | 6.1 |
| Falam Chin cfm | 6.7 |
| Hakka Chinese hak | 6.3 |
| Kachin kac | 6.3 |
| Khumi Chin cnk | 6.2 |
| Kulung (Nepal) kle | 6.0 |
| Lahu lhu | 6.7 |
| Lashi lsi | 7.6 |
| Lolopo ycl | 7.2 |
| Mandarin Chinese cmn | 7.7 |
| Maru mhx | 7.6 |
| Min Nan Chinese nan | 6.8 |
| Mro-Khimi Chin cmr | 7.3 |
| Newari new | 6.1 |
| Pwo Northern Karen pww | 5.5 |
| Sherpa xsr | 7.4 |
| Sunwar suz | 6.6 |
| Tedim Chin ctd | 6.6 |
| Yue Chinese yue | 8.0 |
| Zyphe Chin zyp | 7.2 |
| **QUECHUAN: 22** | |
| Ayacucho Quechua quy | 7.2 |
| Cajamarca Quechua qvc | 7.8 |
| Cañar Highland Quichu qxr | 5.6 |
| Cusco Quechua quz | 6.8 |
| Huallaga Huánuco Quec qub | 7.1 |
| Huamalíes-Dos de Mayo qvh | 6.2 |
| Huaylas Ancash Quechu qwh | 6.6 |
| Huaylla Wanca Quechua qvw | 6.7 |
| Inga inb | 6.8 |
| Lambayeque Quechua quf | 6.9 |
| Margos-Yarowilca-Laur qvm | 6.1 |
| Napo Lowland Quechua qvo | 6.4 |
| North Bolivian Quechu qul | 6.7 |
| North Junín Quechua qvn | 7.3 |
| Northern Conchucos An qxn | 5.9 |
| Northern Pastaza Quic qvz | 6.1 |
| Panao Huánuco Quechua qxh | 8.2 |
| San Martín Quechua qvs | 6.8 |
| South Bolivian Quechu quh | 6.5 |
| South Bolivian Quechu quh | 7.0 |
| Southern Pastaza Quec qup | 6.1 |
| Tena Lowland Quichua quw | 6.2 |
| **EASTERN SUDANIC: 19** | |
| Acoli ach | 6.8 |
| Adhola adh | 6.5 |
| Alur alz | 7.3 |
| Bari bfa | 5.2 |
| Datooga tcc | 6.9 |
| Kakwa keo | 6.7 |
| Karamojong kdj | 6.5 |
| Kumam kdi | 6.2 |
| Kupsabiny kpz | 6.7 |
| Lango (Uganda) laj | 7.8 |
| Luwo lwo | 8.4 |
| Mabaan mfz | 6.7 |
| Markweeta enb | 7.3 |
| Murle mur | 7.8 |
| Nuer nus | 6.9 |
| Sabaot spy | 8.1 |
| Shilluk shk | 6.9 |
| Southwestern Dinka dik | 7.5 |
| Teso teo | 7.1 |
| **TURKIC: 18** | |
| Bashkir bak | 6.0 |
| Chuvash chv | 7.3 |
| Crimean Tatar crh | 5.4 |
| Gagauz gag | 5.3 |
| Gagauz gag | 5.6 |
| Kara-Kalpak kaa | 6.7 |
| Karachay-Balkar krc | 6.9 |
| [e] Kazakh kaz | 6.8 |
| Khakas kjh | 5.4 |
| Kumyk kum | 6.5 |
| Nogai nog | 5.4 |
| [e] North Azerbaijani azj | 6.8 |
| Southern Altai alt | 7.2 |
| Tatar tat | 7.4 |
| [e] Turkish tur | 7.8 |
| [e] Turkish tur | 8.6 |
| Tuvinian tyv | 6.2 |
| Uighur uig | 6.2 |

| UTO-AZTECAN: 15 | |
|---|---|
| Central Huasteca Nahu nch | 6.3 |
| Eastern Huasteca Nahu nhe | 6.3 |
| El Nayar Cora crn | 6.9 |
| Guerrero Nahuatl ngu | 7.0 |
| Highland Puebla Nahua azz | 6.1 |
| [w] Isthmus-Mecayapan Nah nhx | 6.1 |
| [w] Isthmus-Mecayapan Nah nhx | 6.2 |
| Mayo mfy | 6.3 |
| Northern Oaxaca Nahua nhy | 5.9 |
| Northern Puebla Nahua ncj | 6.4 |
| Santa Teresa Cora cok | 5.4 |
| Sierra Negra Nahuatl nsu | 6.7 |
| Southeastern Puebla N npl | 6.4 |
| Western Huasteca Nahu nhw | 7.1 |
| Zacatlán-Ahuacatlán-T nhi | 6.4 |
| **CREOLES AND PIDGINS: 14** | |
| Belize Kriol English bzj | 6.8 |
| Bislama bis | 6.7 |
| Eastern Maroon Creole djk | 7.8 |
| Haitian hat | 7.0 |
| Islander Creole Engli icr | 6.3 |
| Jamaican Creole Engli jam | 6.3 |
| Krio kri | 7.0 |
| [w] Morisyen mfe | 7.0 |
| Nigerian Pidgin pcm | 6.9 |
| Pijin pis | 6.1 |
| Saint Lucian Creole F acf | 5.8 |
| [w] Saramaccan srm | 7.4 |
| Sranan Tongo srn | 7.6 |
| [e] Tok Pisin tpi | 6.2 |
| **CENTRAL SUDANIC: 13** | |
| Aringa luc | 6.5 |
| Avokaya avu | 6.9 |
| Bedjond bjv | 7.0 |
| Gor gqr | 6.5 |
| Gulay gvl | 7.8 |
| Jur Modo bex | 6.7 |
| Kenga kyq | 7.4 |
| Lugbara lgg | 6.8 |
| Ma'di mhi | 7.1 |
| Mbay myb | 8.4 |
| Moru mgd | 6.5 |
| Ngambay sba | 7.5 |
| Nomaande lem | 6.6 |
| **MANDE: 13** | |
| Bambara bam | 7.5 |
| Bissa bib | 7.7 |
| Boko (Benin) bqc | 7.0 |
| Busa bqp | 6.7 |
| Dyula dyu | 7.1 |
| Dyula dyu | 8.4 |
| Kuranko knk | 9.2 |
| Loko lok | 7.1 |
| Mandinka mnk | 6.8 |
| Mende (Sierra Leone) men | 7.7 |
| Northern Bobo Madaré bbo | 8.1 |
| Susu sus | 8.2 |
| Xaasongaxango kao | 8.5 |
| **TRANS-NEW GUINEA: 12** | |
| Anjam boj | 6.2 |
| Awa (Papua New Guinea awb | 7.8 |
| Ese mcq | 6.6 |
| Gwahatike dah | 7.0 |
| Huli hui | 7.7 |
| Ipili ipi | 6.8 |
| Kuman (Papua New Guin kue | 7.3 |
| Kyaka kyc | 7.7 |
| Lower Grand Valley Da dni | 6.9 |
| Lower Grand Valley Da dni | 7.1 |
| Nalca nlc | 6.7 |
| South Tairora omw | 5.8 |
| **TUCANOAN: 11** | |
| Desano des | 7.5 |
| Guanano gvc | 7.4 |
| Koreguaje coe | 6.6 |
| Macuna myy | 5.8 |
| Piratapuyo pir | 6.5 |
| Secoya sey | 6.1 |
| Siona snn | 5.9 |
| Siriano sri | 7.3 |
| Tucano tuo | 7.7 |
| Tucano tuo | 7.6 |
| Tuyuca tue | 7.6 |
| **TUPIAN: 8** | |
| Aché guq | 6.1 |
| Eastern Bolivian Guar gui | 8.2 |
| Guajajára gub | 6.4 |
| Guarayu gyr | 6.8 |
| Kayabí kyz | 6.2 |
| Paraguayan Guaraní gug | 6.6 |
| Urubú-Kaapor urb | 7.2 |
| Western Bolivian Guar gnw | 7.2 |
| **ARAWAKAN: 7** | |
| Asháninka cni | 7.8 |
| Garifuna cab | 7.1 |
| Ignaciano ign | 6.9 |
| Machiguenga mcb | 6.6 |
| Nomatsiguenga not | 6.3 |
| Parecís pab | 6.1 |
| Tereno ter | 6.5 |
| **CHIBCHAN: 7** | |
| Border Kuna kvn | 6.8 |
| Cabécar cjp | 6.2 |
| Central Tunebo tuf | 6.7 |
| Cogui kog | 6.7 |
| Ngäbere gym | 6.4 |
| San Blas Kuna cuk | 6.1 |
| Teribe tfr | 6.6 |
| **DRAVIDIAN: 5** | |
| Kannada kan | 6.6 |
| Kurukh kru | 7.0 |
| Malayalam mal | 8.7 |
| [e] Tamil tam | 6.6 |
| [e] Telugu tel | 9.1 |
| **AUSTRO-ASIATIC: 4** | |
| Eastern Bru bru | 6.3 |
| Juang jun | 7.0 |
| Khmer khm | 8.7 |
| [e] Vietnamese vie | 6.6 |
| **JIVAROAN: 4** | |
| Achuar-Shiwiar acu | 6.2 |
| Aguaruna agr | 6.4 |
| Huambisa hub | 6.0 |
| Shuar jiv | 6.9 |
| **PANOAN: 4** | |
| Cashinahua cbs | 6.4 |
| Panoan Katukína knt | |
| Sharanahua mcd | 6.0 |
| Shipibo-Conibo shp | 6.5 |
| **TAI-KADAI: 4** | |
| [e] Lao lao | 6.6 |
| Northern Thai nod | 5.8 |
| Tai Dam blt | 5.6 |
| [e] Thai tha | 6.7 |
| **TOTONACAN: 4** | |
| Coyutla Totonac toc | 6.6 |
| Highland Totonac tos | 6.4 |
| Pisaflores Tepehua tpp | 5.7 |
| Tlachichilco Tepehua tpt | 6.9 |
| **CARIBAN: 3** | |
| Akawaio ake | 8.3 |
| Galibi Carib car | 7.3 |
| Patamona pbc | 6.3 |
| **GUAHIBAN: 3** | |
| Cuiba cui | 7.3 |
| Guahibo guh | 7.6 |
| Guayabero guo | 7.6 |
| **HUITOTOAN: 3** | |
| Bora boa | 5.8 |
| Minica Huitoto hto | 7.0 |
| Murui Huitoto huu | 6.4 |
| **MIXE-ZOQUE: 3** | |
| Coatlán Mixe mco | 7.9 |
| Highland Popoluca poi | 8.0 |
| Quetzaltepec Mixe pxm | 6.1 |
| **URALIC: 3** | |
| [w] Finnish fin | 5.2 |
| Komi-Zyrian kpv | 6.7 |
| Udmurt udm | 5.9 |
| **WEST PAPUAN: 3** | |
| Galela gbi | 7.2 |
| Tabaru tby | 5.7 |
| Tobelo tlb | 6.6 |
| **AYMARAN: 2** | |
| Central Aymara ayr | 7.3 |
| Central Aymara ayr | 7.2 |
| **BARBACOAN: 2** | |
| Awa-Cuaiquer kwi | 6.8 |
| Guambiano gum | 6.4 |
| **CHOCO: 2** | |
| Epena sja | 7.3 |
| Northern Emberá emp | 7.3 |
| **MONGOLIC: 2** | |
| Halh Mongolian khk | 7.4 |
| Kalmyk xal | 6.9 |
| **NAKH-DAGHESTAN.: 2** | |
| Avaric ava | 6.0 |
| Chechen che | 7.1 |
| **TACANAN: 2** | |
| Ese Ejja ese | 6.5 |
| Tacana tna | 6.6 |
| **ALGIC: 1** | |
| Central Ojibwa ojc | 6.1 |
| **ARAUAN: 1** | |
| Paumarí pad | 5.4 |
| **BASQUE: 1** | |
| Basque eus | 6.2 |
| **CACUA-NUKAK: 1** | |
| Cacua cbv | 7.6 |
| **CAHUAPANAN: 1** | |
| Chayahuita cbt | 8.0 |
| **CHIQUITO: 1** | |
| Chiquitano cax | 7.3 |
| **COFÁN: 1** | |
| Cofán con | 6.8 |
| **DOGON: 1** | |
| Toro So Dogon dts | 8.2 |
| **EAST BIRD'S HEAD: 1** | |
| Meyah mej | 6.2 |
| **EAST BOUGAINVILLE: 1** | |
| Naasioi nas | 6.1 |
| **ESKIMO-ALEUT: 1** | |
| Central Siberian Yupi ess | 5.8 |
| **GUAICURUAN: 1** | |
| Toba tob | 6.2 |
| **HMONG-MIEN: 1** | |
| [w] Hmong Daw mww | 7.2 |
| **HUAVEAN: 1** | |
| San Mateo Del Mar Hua huv | 7.0 |
| **KHOE-KWADI: 1** | |
| Southern Samo sbd | 8.3 |
| **KOMAN: 1** | |
| Uduk udu | 6.8 |
| **KORDOFANIAN: 1** | |
| Moro mor | 6.6 |
| **LOWER SEPIK-RAMU: 1** | |
| Aruamu msy | 6.5 |
| **MACRO-GE: 1** | |
| Kayapó txu | 7.7 |
| **MASCOIAN: 1** | |
| Enxet enx | 6.0 |
| **MATACOAN: 1** | |
| Maca mca | 6.1 |
| **MISUMALPAN: 1** | |
| Mískito miq | 6.2 |
| **PÁEZAN: 1** | |
| Páez pbb | 7.2 |
| **PUINAVE: 1** | |
| Puinave pui | 5.2 |
| **SULKA: 1** | |
| Sulka sua | 6.8 |
| **TARASCAN: 1** | |
| Purepecha tsz | 8.0 |
| **TICUNA: 1** | |
| Ticuna tca | 6.6 |
| **TOL: 1** | |
| Tol jic | 7.2 |
| **TOR-ORYA: 1** | |
| Orya ury | 5.7 |
| **URARINA: 1** | |
| Urarina ura | 6.2 |
| **URU-CHIPAYA: 1** | |
| Chipaya cap | 6.5 |
| **YANOMAM: 1** | |
| Sanumá xsu | 6.6 |
| **ZAMUCOAN: 1** | |
| Chamacoco ceg | 6.5 |