IPA CHILDES & G2P+: Feature-Rich Resources for Cross-Lingual Phonology and Phonemic Language Modeling

Zébulon Goriely 🍊 **Paula Buttery** 🍊🍋
🍊 Department of Computer Science & Technology, University of Cambridge, U.K.
🍋 ALTA Institute, University of Cambridge, U.K.
🍊 firstname.secondname@cl.cam.ac.uk

Abstract

In this paper, we introduce two resources: (i) G2P+, a tool for converting orthographic datasets to a consistent phonemic representation; and (ii) IPA CHILDES, a phonemic dataset of child-directed and child-produced speech across 31 languages. Prior tools for grapheme-to-phoneme conversion result in phonemic vocabularies that are inconsistent with established phonemic inventories, an issue which G2P+ addresses by leveraging the inventories in the Phoible database (Moran and McCloy, 2019). Using this tool, we augment CHILDES (MacWhinney and Snow, 1985) with phonemic transcriptions to produce IPA CHILDES. This new resource fills several gaps in existing phonemic datasets, which often lack multilingual coverage, spontaneous speech, and a focus on child-directed language. We demonstrate the utility of this dataset for phonological research by training phoneme language models on 11 languages and probing them for distinctive features, finding that the distributional properties of phonemes are sufficient to learn major class and place features cross-lingually.

🤗 phonemetransformers/ipa-childes (CC BY 4.0)
codebyzeb/g2p-plus (MIT)

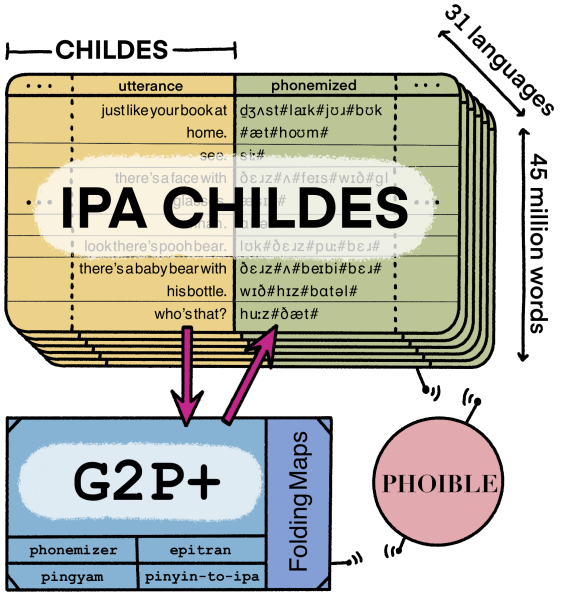


Figure 1: An overview of IPA CHILDES and G2P+, which are introduced in this paper.

1 Introduction

Phonological research can be enriched by large-scale data-oriented studies that investigate phoneme function across the globe's languages. However, while written text is plentiful and easily accessible across hundreds of languages, phonemic data is much more limited in availability. Phonemic datasets can be created by employing expert phoneticians to carefully transcribe speech, but this is a time-consuming process and completely infeasible for creating large datasets. Instead, the typical approach is to use grapheme-to-phoneme (G2P) conversion tools, which use statistical rules and pronunciation dictionaries to convert orthographic text to a phonemic representation. Open-source G2P tools have been used to create large and multilingual phonemic datasets with domains ranging from telephone conversations to legal proceedings. However, the fact that these tools are open-sourced and use a variety of statistical approaches and transcription schemes means that phonemic corpora vary considerably according to their phonemic vocabularies and level of phonetic detail, making it difficult to compare findings and incorporate other linguistic resources into analysis.

There is also a lack of phonemic data for certain domains, preventing phonological research in these areas. In particular, we note that it is difficult to find phonemic data for child-centered speech[1] and,

[1] Child-centered speech is speech occurring within a child's environment and includes child-directed and child-produced utterances.

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 502–521
July 31 - August 1, 2025

in general, spontaneous speech across several languages. The *de-facto* repository for child-centered data is the Child Language Data Exchange System (CHILDES), which currently contains over 1.4TB of transcript data in over 40 languages (MacWhinney and Snow, 1985; MacWhinney, 2019). The impact of CHILDES across clinical and linguistic research has been profound (Ratner, 2024) but the largely orthographic nature of the data has prevented phonological experimentation.[2]

We thus identify two major challenges impeding phonological research. First, the lack of consistent G2P conversion, which we address by developing `G2P+`, a tool for converting orthographic text to a phonemic representation. `G2P+` leverages existing G2P tools for conversion but carefully maps the output to established phonemic inventories in Phoible, a database of cross-linguistic phonological inventory data. Using Phoible inventories not only ensures consistency for each language regardless of the G2P backend used, but the database also contains phonological feature information, supporting fine-grained phonological analysis. Second, we address the lack of a multilingual phonemic dataset of child-centered speech by using `G2P+` to convert the majority of the CHILDES database to phonemes. The resulting dataset, IPA CHILDES, contains phonemic transcriptions of 31 languages in CHILDES, totaling 45 million words. We illustrate these resources in fig. 1.

We exemplify how to use these resources by training cross-lingual phoneme language models. Phoneme LMs have a wide variety of applications in NLP, including lyric generation (Ding et al., 2024), text-to-speech (Li et al., 2023), and low-resource language modeling (Leong and Whitenack, 2022). Developmentally plausible training corpora also provide a means of studying emergent phonology, but past work has been limited by the availability of training and evaluation resources in languages besides English. Here, after establishing the scaling conditions of phoneme LMs, we train monolingual models on the 11 largest languages in IPA CHILDES. Using the fact that `G2P+` maintains a correspondence with Phoible during conversion, we use linear probes to predict an input phoneme's phonological features from its contextual embedding. We evaluate this approach against the phoneme's feature description in Phoible and find that the probes consistently correctly predict the 'syllabic' and 'consonantal' features, indicating the broad separation of vowels and consonants across languages and demonstrating the utility of phoneme LMs for studying emergent phonology.

These experiments demonstrate the utility of our tools for phonological analysis. We release `G2P+`, IPA CHILDES, and all trained models to support future work.

2 Related Work

2.1 Phonemic Datasets

Phonemic data is required to investigate a range of linguistic phenomena. Recently, researchers have used data-driven approaches to study morphological theories of acquisition (Kirov and Cotterell, 2018), explore the role of distributional information in phonology (Mayer, 2020), calculate cross-language phonological distance (Eden, 2018) and simulate early lexicon learning (Goriely et al., 2023). Despite the benefits of phonemic data, few such datasets exist.

Written text and audio datasets are far more plentiful than phonemic datasets. Written text, being widely distributed and easy to collect through practices such as web-scraping (Bansal et al., 2022), has steered years of NLP research, ranging from the parsers trained on the Penn Treebank (Taylor et al., 2003) to the large language models trained on billion-word datasets like the Pile (Gao et al., 2020). Despite the availability of written text, it is often inappropriate for speech technology and phonological research. Instead, since tape recorders became widely available, researchers have created datasets of human speech. These now include **elicited** speech corpora such as TIMIT, (Garofolo et al., 1993), FLEURS (Conneau et al., 2023), the MSWC (Mazumder et al., 2021), GlobalPhone (Schultz, 2002) and CommonVoice (Ardila et al., 2020); **audio book** corpora such as LibriSpeech (Panayotov et al., 2015), MLS (Pratap et al., 2020) and the CMU Wilderness Corpus (Black, 2019); and **naturalistic** speech corpora such as Switchboard (Godfrey et al., 1992), the Fisher corpus (Cieri et al., 2004), the British National Corpus (Consortium, 2007), the Buckeye corpus (Pitt et al., 2007), Babel (Harper, 2011) and VoxLingua107 (Valk and Alumäe, 2021). Of these datasets, only TIMIT, MLS and Switchboard include phonemic annotations, limiting their use in phonological analysis.

[2] CHILDES does contain phonetic transcriptions for some languages as part of the PhonBank project, but only for a select few corpora and only for child-produced utterances, impeding the phonological analysis of child-*directed* speech.

Later work augmented these datasets with phonemic transcriptions. These include Audio BNC derived from the British National Corpus (Coleman et al., 2011), LibriLight derived from LibriSpeech (Kahn et al., 2020), VoxClamantis derived from the CMU Wilderness Corpus (Salesky et al., 2020), VoxCommunis derived from CommonVoice (Ahn and Chodroff, 2022) and IPAPACK derived from FLEURS and MSWC (Zhu et al., 2024).

These datasets and their phonemically-annotated successors all vary considerably according to the language coverage, number of words, domain and the presence of text-based transcriptions. We provide a summary of these properties in appendix C. Our dataset, IPA CHILDES, is the first phonemic dataset for *child-centered* speech and the first *multilingual* phonemic dataset for spontaneous speech.

2.2 Grapheme to Phoneme Conversion

Ideally, phonemic transcriptions of speech would originate from expert human annotators, but such annotation is incredibly slow. For instance, it was estimated that it would take 120 person-years to transcribe and align the 1200 hours of speech in the Audio BNC corpus (Coleman et al., 2011). Of the phonemic datasets described above, only the smallest, TIMIT, was fully transcribed by human experts, at a rate of only 100 sentences per week (Zue and Seneff, 1996; Lamel et al., 1989). Switchboard also provides human-annotated phonemic transcriptions but only for 5,000 utterances (Greenberg et al., 1996).

In practice, phonemic transcriptions are produced using G2P. In the simplest case, this involves the use of pronunciation dictionaries such as the Carnegie Mellon University (CMU) Pronouncing Dictionary[3] or the English Pronouncing Dictionary (Jones, 2011). These were used to create the phonemic transcriptions for the Buckeye Corpus, Audio BNC and Babel, but pronunciation dictionaries are limited by the items included in the dictionary and so may fail to convert part-words, interruptions or rare proper nouns, which frequently occur in spontaneous speech. More sophisticated G2P methods combine pronunciation dictionaries with statistical models. These systems have been developed for many languages using rules or finite-state transducers to generalize to unseen words (Mortensen et al., 2018; Hasegawa-Johnson et al., 2020; Bernard and Titeux, 2021). Other G2P systems have applied neural networks to automatically learn these rules and generalize to new languages (Novak et al., 2016; Zhu et al., 2022).

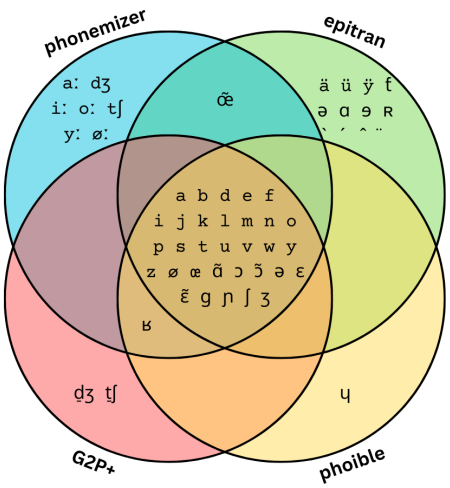


Figure 2: Venn diagram of the inventories produced by `phonemizer`, `epitran` and `G2P+` compared to Phoible inventory 2269 for French.

As G2P systems operate only from text, they may fail to capture accents and the variation found in natural speech (see appendix A for a discussion). Nevertheless, G2P systems provide a useful method for producing phonemic transcriptions at scale, and were used to produce the transcriptions for LibriSpeech, VoxClamantis and IPAPACK. The fact that transcription errors may occur is often acknowledged as a limitation, but rarely are the outputs of different G2P systems compared to each other or to established inventories. For instance, `epitran` and `phonemizer`, two popular tools described in section 3.1, produce very different inventories for French, as demonstrated in fig. 2.

In this work, we leverage existing statistical G2P tools, validate their outputs using maps to Phoible inventories, and use our resulting tool to produce phonemic transcriptions for the utterances in the CHILDES database.

2.3 Phoneme LMs and Child-Centered Data

In this work, we illustrate one use of our dataset by training small monolingual LMs on 11 languages and examining the representations they learn for individual phonemes.

Training models on such little data (here, only 500 thousand words) may be considered atypical in the modern NLP landscape, but questions of developmental plausibility have led to an increased

[3] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

interest in pretraining with limited data. For instance, the BabyLM workshop series challenges participants to train smaller models on data that is limited by both scale, 10–100 million words, and by domain, with the pre-training corpus including data from CHILDES, among other child-centered corpora (Warstadt et al., 2023; Hu et al., 2024b). Such limitations have led to the development of new architectures (Georges Gabriel Charpentier and Samuel, 2023; Charpentier and Samuel, 2024), motivated cognitively-inspired pre-training strategies (Huebner et al., 2021; Diehl Martinez et al., 2023) and allowed for gaining insights into human learning (Yedetore et al., 2023). The majority of this work has centered on English. Exceptions include Capone et al. (2024); Shen et al. (2024), who train Italian monolingual and bilingual models, respectively, Yadavalli et al. (2023) who use data from five language in CHILDES to explore second language acquisition theories (but only train an English LM) and Salhan et al. (2024), who use age-ordered data from four languages in CHILDES to explore fine-grained curricula inspired by language acquisition.

However, these BabyLMs are typically trained on orthographic text, limiting their ability to be studied at the phonological level, and generally use subword tokens, which do not generally correspond to cognitively plausible units (Beinborn and Pinter, 2023) limiting their value for psycholinguistic research (Giulianelli et al., 2024). Bunzeck et al. (2024) and Goriely et al. (2024) both establish phoneme-based training of BabyLMs (where tokens consist of individual phonemes, with word boundaries removed) but only train on English text. Here, we use IPA CHILDES to demonstrate phoneme-based training for 11 languages and leverage the fact that G2P+ maintains a correspondence to Phoible in order to probe our BabyLMs for knowledge of distinctive features.

3 G2P+

We introduce G2P+ as a tool for converting datasets from an orthographic representation to a phonemic representation. It operates either as a python library or as a command-line program; the user selects one of four backends and the language to use for conversion. Each backend supports a different set of languages as described in section 3.1. The recommended backends for each of the languages in IPA CHILDES are given in appendix B and example usage of the tool is given in appendix D.

Each line of orthographic text is converted to phonemes, represented using the International Phonetic Alphabet (IPA). Regardless of the backend selected, the representation is consistent, with phonemes separated by whitespace (for convenient tokenization) and unique delimiters used to separate words and utterances (see appendix E for details).

The output representation is also consistent in terms of the set of phonemes types produced, using *folding*, as described in section 3.2. Without folding, each backend produces a different set of phonemes (as demonstrated in fig. 2) which may not align with established phoneme inventories. Our folding maps not only ensure the output is consistent regardless of the backend chosen, but also makes it easy to leverage information in Phoible in analysis, as demonstrated in section 5.2.

3.1 G2P Backends

In order to support a wide variety of languages, we implement wrappers around four backend G2P tools:

`phonemizer`: Phonemizer (Bernard and Titeux, 2021) is a python library for G2P in various languages based on `eSpeak`[4], an open-source speech synthesizer which supports over one hundred languages and accents (Dunn and Vitolins, 2022).

`epitran`: Epitran (Mortensen et al., 2018) supports the automatic grapheme-to-phoneme conversion of text across many languages, accents and scripts, with a particular focus on low-resource languages. For the majority of the 92 languages supported,[5] it uses greedily-interpreted grapheme-to-phoneme maps augmented with context-sensitive pre-processor and post-processor rewrite rules.

`pinyin-to-ipa`: Pinyin-to-ipa (Taubert, 2024) is a python library for converting Mandarin written in pinyin to IPA using a few contextual grapheme-to-phoneme maps. The phoneme inventory is based on the phonology of Mandarin as described by (Lin, 2007) and (Duanmu, 2007) and tone markers are attached to the vowel of the syllable, rather than the

[4]For Japanese text written in Romanji, as is the case in CHILDES, we use phonemizer with the the Segments backend (Forkel et al., 2019).

[5]For English, Epitran uses the Flite Speech Sythesis System (Black and Lenzo, 2001) and for Simplified and Traditional Chinese it uses the CC-CEDict dictionary (`https://cc-cedict.org`).

end of the syllable. The tool only converts individual pinyin syllables, so our wrapper first splits the input into syllables before using the tool to convert each syllable to IPA.

pingyam: Pingyam[6] is a table storing conversion information between the various romanization systems of Cantonese (including IPA) based on data from the Open Cantonese Dictionary.[7] Our wrapper converts from the Jyutping system to IPA by first splitting the input text into syllables before using the table to convert each syllable to IPA. For consistency with `pinyin-to-ipa`, we move tone markers to the vowel of each syllable.

Although `pinyin-to-ipa` and `pingyam` only support one Chinese language each, we include them as backends because `epitran` and `phonemizer` have relatively poor G2P quality for these languages. This has prevented Chinese languages from being included in previous cross-lingual phonemic datasets (Ahn and Chodroff, 2022) and has led to them being disregarded in cross-lingual analysis (Pimentel et al., 2020). We hope that by including these backends, we address this gap. We also combine tone markers with their preceding phoneme to create a unique token (e.g., a˥ is a single token, not two). We thus treat tone markers as phonological features rather than as individual phonemes, similar to how diphthongs are unique phonemes. However, this decision is still debatable and does lead to a comparatively larger phonemic vocabulary, so we provide an option to disable this merging (see appendix D).

3.2 Phoneme inventory validation

In order to validate the set of phonemes produced by each choice of backend and language, we compare the output to the phoneme inventories for that language listed in Phoible, a database containing phoneme inventories extracted from source documents and tertiary databases for 2186 distinct languages (Moran and McCloy, 2019).

Phoible also contains typological data and phonological feature information for each phoneme, a useful resource for phonological analysis. As there are often multiple inventories in Phoible for each language, we choose the inventory that best matches the output phoneme of all backends that supports that language, according to the number of phoneme types, the number of consonants, the number of vowels and the number of diphthongs.

Once the best inventory has been found, we use a process called *folding* to align the output phoneme set with the inventory and correct errors in the output. This is achieved a manually-crafted look-up table (a *folding map*) which is applied to the output of the G2P wrapper. These maps are primarily used to solve surface-level errors, instances where the G2P tool outputs a specific Unicode string for a specific phoneme but the inventory lists a different string. For example, the `phonemizer` backend with the `ja` language code (Japanese) outputs the tied characters t͡s as one of the phonemes, but the Japanese inventory lists ts instead. These errors can be solved with a simple one-to-one mapping. These mappings will not affect the information-theoretic properties of the output but do allow the output symbols to be matched with entries in Phoible.

Besides these surface-level errors, other transcription errors can also be solved with folding maps. For example, the `epitran` backend for Serbian always outputs d ʒ as two phonemes instead of the single phoneme dʒ, which can also be solved with a single mapping. The construction of the folding maps and these additional error types are discussed further in appendix F.

3.3 Qualitative Analysis

In fig. 2, we compare the matching Phoible inventory for French to the output of `G2P+` (using `phonemizer` as a backend) and the outputs produced by `phonemizer` and `epitran` when applied to the French section of CHILDES. The outputs of `phonemizer` and `epitran` both differ considerably from the inventory and from each other whereas the `G2P+` only fails to produce a single phoneme, ɥ, and produces two additional phonemes dʒ and tʃ, which we allow as they come from loanwords such as “pizza” and “sandwich”.

4 IPA CHILDES

IPA CHILDES contains 45 million words of monolingual child-centered speech for 31 languages. The data is sorted by child age in order to support curriculum learning experiments, such as in the work of Huebner et al. (2021), and we also provide an ‘is_child’ feature to allow for filtering child or adult utterances.

In order to create the dataset, we first download all monolingual and non-SLI corpora in CHILDES.

[6] https://github.com/kfcd/pingyam

[7] https://www.kaifangcidian.com/han/yue/

CHILDES has 48 languages but only 31 are supported by a backend in G2P+ (either because the language is not supported, or because they have been transcribed using an irregular script). For languages supported by multiple backends, we produce a sample transcription using each backend and carefully examine the output. The 'best-fitting' backend (the one that produces a phonemic vocabulary closest to one of the inventories in Phoible) is selected and is the backend for which we produce a folding map, as described in section 3.2. Having selected the best backend, we use G2P+ to convert all orthographic utterances for each language to a phonemic representation, producing a CSV containing the original representation, the phonemic representation as well as additional data stored in CHILDES (such as target child age, morpheme count, part of speech information, and the IDs of each utterance, transcript, corpus and collection).

An illustration of the dataset is given in fig. 1 and a description of each language section is given in appendix B, detailing the matching Phoible inventory and CHILDES section for each language. Note that English is divided into British English (EnglishUK) and North American English (EnglishNA) to mirror the split present in CHILDES and Portuguese is also split into European and Brazilian varieties, following previous work (Caines et al., 2019; Goriely et al., 2023). For these splits, we use different `phonemizer` accents. Data is not uniformly distributed across languages. EnglishNA is the most represented, with close to 10 million words, and Farsi is the least represented, with only 43 thousand words. We discuss limitations of the dataset in appendix A.

5 Cross-Lingual Phoneme LMs

Phoneme LMs trained on developmentally plausible corpora allow for the testing of phonological representations but recent work has only explored English models trained on 10 – 100 million words (see section 2.3). Here, we establish the size requirements for models trained on data available in IPA CHILDES and then demonstrate how models trained on the 11 largest languages in our dataset can be used to explore emergent phonology.

Each of our models are auto-regressive, trained to predict phonemes in a sequence. This is similar to how standard auto-regressive models are trained, except that each token represents a single phoneme, rather than a word or subword. We refer to the suite of models as "cross-lingual" as each individual model is monolingual, only trained on data from a single language. This is in contrast to "multilingual" models that are trained on multiple languages at once.

5.1 Size Requirements of Phoneme LMs

We use the BabySLM benchmark (Lavechin et al., 2023) to evaluate syntactic and phonological knowledge. The *syntactic* score is calculated using a preference task over pairs of grammatical and ungrammatical sentences across six syntactic phenomena commonly seen in naturalistic speech. For example, models should assign ð ə g ʊ d k ɪ t i ("the good kitty") a higher likelihood than ð ə k ɪ t i g ʊ d ("the kitty good"). The *lexical* score is similarly calculated using minimal pairs of words and pseudowords, such as ɹ uː l ə ɹ z ("rulers") compared to the pseudo-word m uː k ə ɹ z ("mukers"). Lavechin et al. (2023) demonstrated that an LSTM model trained on 1.2 million words from Providence (one of the corpora in CHILDES) achieved a lexical score of 75.2 and a syntactic score of 55.1[8]. Goriely et al. (2024) later achieved lexical and syntactic scores of 87.8 and 83.9 when training a larger transformer-based model on the 100-million-word BabyLM challenge dataset (Hu et al., 2024a).

Here, we use IPA CHILDES and BabySLM to establish the scaling laws of phoneme LMs in terms of data size and model size. We subsample the EnglishNA portion of the dataset, remove word boundaries and child-produced utterances and train a suite of GPT-2 models ranging from 400 thousand to 19 million non-embedding parameters. To prevent overfitting, we train three models for each combination of model size and data size using dropouts of 0.1, 0.3 and 0.5, selecting the model with the lowest perplexity for each. Model parameters, training configurations and scripts are provided in appendix G.

The scaling graphs for the lexical and syntactic scores are given in fig. 3. For every model size, performance increases with more training data but for a particular data size the largest model is not always the best. For instance, the second smallest model is the best choice for the lexical task if only 300 thousand tokens of data are available, likely due to larger models overfitting with a sample this small (even with high dropout). It is also clear that

[8] Chance performance for both BabySLM scores is 50 and 100 indicates perfect performance

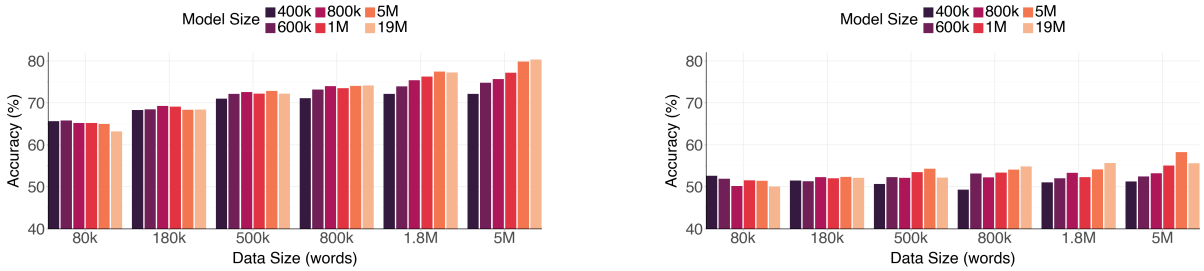


Figure 3: BabySLM lexical score (left) and syntactic score (right) achieved by a phoneme-based GPT-2 model trained on the EnglishNA portion of IPA CHILDES across model sizes and subsample sizes.

although small models with very little data seem to acquire phonological knowledge (as measured by the lexical score), much more data is required to achieve syntactic scores past 60, in line with the results of Lavechin et al. (2023) and Goriely et al. (2024). The best model parameters for each score and data size are given in appendix H.

5.2 Probing for Phonological Features

As the phonemic utterances in IPA CHILDES maintain a correspondence with Phoible, we can use the **distinctive feature** information in Phoible to probe cross-lingual phoneme LMs for phonological knowledge.

We select the 11 largest languages in the dataset and train a GPT-2 model on each, subsampling 500 thousand words[9] and using the best-fitting model for this data size according to the previous experiment (the 5-million-parameter model with a dropout of 0.3). The training configuration remains the same (see appendix G). These models allow us to compute contextual embeddings $c(x)$ for phonemes.

We then look up the distinctive features of each phoneme in each language using the matching inventories in Phoible (see table 1). We find the set of features for which, in all 11 languages, there are at least 4 phonemes that exhibit the feature and 4 that do not. For each feature f, we train a linear probe p_f to predict that feature from the contextual embeddings $c(x)$ of phonemes. Each probe is trained with an equal number of positive and negative examples and is evaluated using leave-one-group-out cross-validation (i.e for each phoneme x in the phoneme inventory V, the probe is trained on the contextual embeddings of all other phonemes

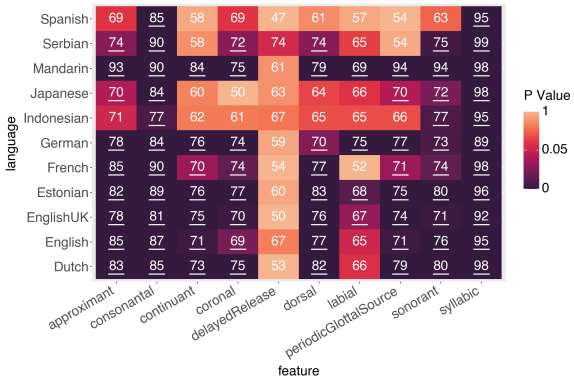


Figure 4: Accuracy of the phonological distinctive feature probe across 11 languages in IPA CHILDES and 9 distinctive features from Phoible.

$\{c(y)|y \in V \setminus \{x\}\}$, then evaluated by predicting the feature from contextual embeddings of the left-out phoneme $p_f(c(x))$, and the final score is a macro-average across all phonemes $x \in V$).

The results of each probe are provided in fig. 4. The majority of the probes achieve accuracies significantly[10] higher than chance (50%), indicating that the models learn representations that encode distinctive features. While the scores for each feature are broadly consistent across languages, some notable differences emerge. For example, nearly all feature probes achieve statistically significant results in Mandarin, whereas only two do so in Spanish. This disparity can be partly attributed to the number of unique phonemes in each language. Because we treat each combination of vowel and tone as a distinct phoneme, Mandarin has 99 phoneme types, compared to just 24 in Spanish. The smaller phoneme inventory in Spanish greatly reduces n for each probe, making it more challenging to obtain

[9] As the number of phonemes per word varies across these languages, we actually subsample 1.8 million tokens (phonemes) for each language, which is roughly 500 thousand words.

[10] Statistical significance was assessed using a binomial test, where the null hypothesis assumes a probability of success $p_0 = 0.5$ and the number of trials n is equal to the number of phonemes tested by the probe. A result was considered significant if the computed p-value was less than 0.05.

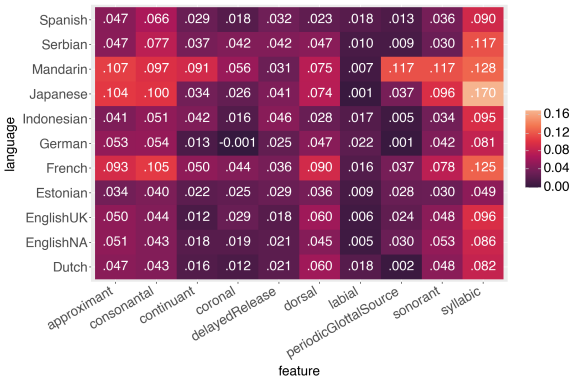


Figure 5: Average silhouette scores when using each distinctive feature to cluster contextual embeddings of the phonemes in each language.

statistically significant results.

In all 11 languages, the highest result is achieved by the probe for the 'syllabic' feature which generally[11] separates vowels from consonants. As these models only learn to predict phonemes and have no concept of how each phoneme is pronounced, the fact that this separation is learned clearly indicates that vowels and consonants provide a strong distributional signal across languages. The `consonantal` feature similarly separates vowels from consonants[12] and is learned by a probe in every language. However, not every feature can be learned by these probes. For instance, the `delayedRelease` feature, which distinguishes stops from affricates, is not learned by any probe. Our models do not encode the rate of phoneme delivery, so it is unsurprising that a feature that relates to the temporal properties of phonemes is difficult to probe.

Distributional Phoneme Clusters

To better understand why the probes capture certain phonological features, we examine whether contextual embeddings cluster according to these features. For each language, we sample 50 contextual embeddings per phoneme and label them with their associated phonological features. For each labeling, we then compute the **silhouette score** for each embedding — a metric ranging from –1 to 1, where higher values indicate that an embedding is more similar to others in its assigned cluster than to those in neighboring clusters (Rousseeuw, 1987). Averaging these scores across all embeddings allows us to compare how well different features cluster the phoneme representations, as shown in fig. 5.

The scores are all relatively close to zero, likely due to the curse of dimensionality — our embeddings have 256 dimensions, far exceeding the number of distinct phonemes in each language. Despite this, the results are consistent with the probe findings: the syllabic feature yields the highest clustering quality.

We further visualize this clustering using dendrograms, created by averaging the contextual embeddings for each phoneme and applying an incremental clustering algorithm. Figure 6 shows examples for Japanese and French, with the syllabic feature marked for each phoneme. In both cases, vowels are almost entirely separated from consonants, with one notable exception: n in Japanese. We also observe some alignment with traditional phoneme groupings (e.g., b and p), though overall the dendrograms diverge from standard phonological classifications. This suggests that the distributional behavior of phonemes in context may not neatly align with their articulatory or categorical properties.

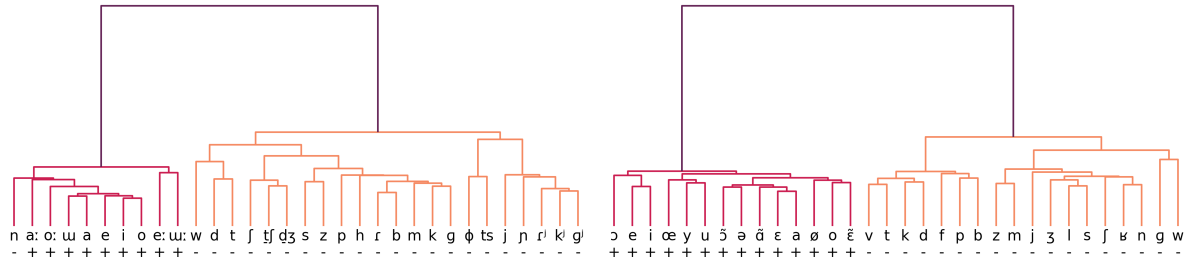
6 Discussion

IPA CHILDES addresses several limitations of past datasets, as the first large multilingual corpus of child-centered phonemic speech. In this study we demonstrate how this data can be used to train phoneme LMs, but this dataset could also support information-theoretic studies of language processing and acquisition, which have previously based their calculations on word types (Piantadosi et al., 2011; Dautriche et al., 2017a; Pimentel et al., 2020) or orthographic text (Mahowald et al., 2013; Dautriche et al., 2017b; Futrell et al., 2020), often citing a lack of phonemic data as a limiting factor. The child-centered domain of our dataset could also be beneficial for studying the 'Goldilocks' hypothesis (Kidd et al., 2014) and the properties of 'Parentese' (Ramírez-Esparza et al., 2017). We provide an example of an experiment investigating the later in appendix I, where we compute the average information of utterances directed to children aged 0–6 across 10 languages and find a general trend of increasing informative content.

Our `G2P+` tool also provides new avenues for linguistic analysis by ensuring that phonemes pro-

[11] In some languages there are also syllabic consonants, which like vowels can act as the nucleus of a syllable.

[12] This feature indicates an audible constriction of the vocal tract, separating obstruents, nasals, liquids, and trills from vowels, glides and laryngeal segments (Gussenhoven and Jacobs, 2017).



(a) Japanese (b) French

Figure 6: Similarity of the contextual embeddings for each phoneme learned by the Japanese and French phoneme LMs. Similarities are computed using Euclidean distance considering the average of 50 contextual embeddings for each phoneme and linkages are created using the incremental algorithm. The 'syllabic' distinctive feature is marked below each phoneme.

duced for each language are consistent with established inventories in Phoible. This not only addresses transcription errors, but also allows for the use of distinctive feature information provided by Phoible in analysis. We demonstrate this by training linear probes to extract distinctive features from the contextual embeddings of phonemes learned by our monolingual models. We find that certain features (e.g. `consonantal`) emerge solely from the distributional properties across all 11 languages, while others (e.g. `delayedRelease`) do not.

Our resources could also support the training of self-supervised speech models (e.g. Hsu et al., 2021). These models are trained directly on audio and lag behind phoneme or text-based models, often requiring several orders of magnitude more data to learn semantic representations (Cuervo and Marxer, 2024), but recent work has found that fine-tuning on phoneme classification can reduce this gap (Feng et al., 2023; Poli et al., 2024). Our work is closely related to recent efforts in low-resource cross-lingual language modeling — for example, the Goldfish suite of monolingual models spanning 350 languages, some trained on as little as 5MB of orthographic text (Chang et al., 2024). IPA is also a more universal representation than orthographic text, which varies considerably across languages, with multilingual IPA models proving to be effective for force-alignment (Zhu et al., 2024) and zero-shot cross-lingual NER (Sohn et al., 2024). In this study we only train monolingual models, but future work could extend this to the multilingual setting.

7 Conclusion

This work introduces `G2P+` and IPA CHILDES, two new resources for phonological research. `G2P+` improves open-source G2P tools by ensuring phonemic vocabularies align with the established inventories in the Phoible database. Using this tool, we create IPA CHILDES by converting the orthographic transcriptions in CHILDES into phonemic representations, resulting in a large corpus of child-centered spontaneous speech across 31 languages.

We demonstrate the utility of these resources for phonological analysis using phoneme LMs by extending prior work to the cross-lingual setting. Our results establish the corpus size requirements for phoneme LMs trained on developmentally plausible corpora and we show that models trained on 11 languages effectively implicitely encode distinctive features. These findings support the role of phoneme LMs in studying emergent phonology. We anticipate that `G2P+` and IPA CHILDES will enable a wide range of future studies in linguistics and NLP, particularly in phonological acquisition, cross-linguistic analysis, and speech processing.

Acknowledgments

We thank Lisa Beinborn for her advice on an an early draft of this article. We are also grateful to Pietro Lesci and Suchir Salhan for their feedback.

Our experiments were performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council. Zébulon Goriely is supported by an EPSRC DTP Studentship.

References

Emily 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, Marseille, France. European Language Resources Association.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in NMT: The effect of noise and architecture. In *International Conference on Machine Learning*, pages 1466–1482. PMLR.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430.

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.

Alan W Black and Kevin A Lenzo. 2001. Flite: a small fast run-time synthesis engine. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*.

Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarrieß. 2024. Graphemes vs. phonemes: Battling it out in character-based language models. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 54–64.

Andrew Caines, Emma Altmann-Richer, and Paula Buttery. 2019. The cross-linguistic performance of word segmentation models over time. *Journal of child language*, 46(6):1169–1201.

Luca Capone, Alice Suozzi, Gianluca E Lebani, Alessandro Lenci, et al. 2024. Babies: A benchmark for the linguistic evaluation of italian baby language models.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *arXiv preprint arXiv:2408.10441*.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. Gpt or bert: why not both? *arXiv preprint arXiv:2410.24159*.

Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. Call for papers – The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

John Coleman, Ladan Baghai-Ravary, John Pybus, and Sergio Grau. 2012. Audio BNC: the audio edition of the spoken british national corpus. *Phonetics Laboratory, University of Oxford*.

John Coleman, Mark Liberman, Greg Kochanski, Lou Burnard, and Jiahong Yuan. 2011. Mining a year of speech. *VLSP 2011: New tools and methods for very-large-scale phonetics research*, pages 16–19.

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.

BNC Consortium. 2007. British National Corpus, XML edition. Literary and Linguistic Data Service.

Santiago Cuervo and Ricard Marxer. 2024. Scaling properties of speech language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T Piantadosi. 2017a. Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T Piantadosi. 2017b. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, 41(8):2149–2169.

Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB – curriculum learning for infant-inspired model building. In

Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning, pages 84–99, Singapore. Association for Computational Linguistics.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*.

San Duanmu. 2007. *The phonology of standard Chinese*. Oxford University Press.

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. The zero resource speech challenge 2021: Spoken language modelling. In *Proc. Interspeech 2021*, pages 1574–1578.

R.H. Dunn and V. Vitolins. 2022. eSpeak NG speech synthesizer. In GitHub respository (Version 1.51).

S Elizabeth Eden. 2018. *Measuring phonological distance between languages*. Ph.D. thesis, UCL (University College London).

Siyuan Feng, Ming Tu, Rui Xia, Chuanzeng Huang, and Yuxuan Wang. 2023. Language-universal phonetic representation in multilingual speech pretraining for low-resource speech recognition. In *INTERSPEECH 2023*, Dublin, Ireland. ISCA.

Robert Forkel, Steven Moran, Johann-Mattis List, Simon J Greenhill, Lucas Ashby, Kyle Gorman, and Gereon Kaiping. 2019. cldf/segments: Unicode standard tokenization.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572, Miami, Florida, USA. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Zébulon Goriely, Andrew Caines, and Paula Buttery. 2023. Word segmentation from transcriptions of child-directed speech using lexical and sub-lexical cues. *Journal of Child Language*, pages 1–41.

Zébulon Goriely, Richard Diehl Martinez, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 37–53, Miami, FL, USA. Association for Computational Linguistics.

Steven Greenberg, Joy Hollenback, and Dan Ellis. 1996. Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In *Proc. ICSLP*, volume 96, pages 24–27.

Carlos Gussenhoven and Haike Jacobs. 2017. *Understanding phonology*. Routledge.

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

Mark Hasegawa-Johnson, Leanne Rolston, Camille Goudeseune, Gina-Anne Levow, and Katrin Kirchhoff. 2020. Grapheme-to-phoneme transduction for cross-language asr. In *Statistical Language and Speech Processing*, pages 3–19, Cham. Springer International Publishing.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Leshem Choshen, Ryan Cotterell, Alex Warstadt, and Ethan Gotlieb Wilcox, editors. 2024a. *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Miami, FL, USA.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and

Ethan Gotlieb Wilcox. 2024b. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Daniel Jones. 2011. *Cambridge English pronouncing dictionary with CD-ROM*. Cambridge University Press.

J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-Light: A benchmark for ASR with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 7669–7673. IEEE.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2017. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174.

Celeste Kidd, Steven T Piantadosi, and Richard N Aslin. 2014. The goldilocks effect in infant auditory attention. *Child development*, 85(5):1795–1804.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Lori F Lamel, Robert H Kassel, and Stephanie Seneff. 1989. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proc. SIOA 1989*, pages Vol–2.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*, pages 4588–4592, Dublin, Ireland. ISCA.

Colin Leong and Daniel Whitenack. 2022. Phone-ing it in: Towards flexible multi-modal language model training by phonetic representations of data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5306–5315, Dublin, Ireland. Association for Computational Linguistics.

Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023. Phoneme-level BERT for enhanced prosody of text-to-speech with grapheme predictions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yen-Hwei Lin. 2007. *The Sounds of Chinese*. Cambridge University Press.

Brian MacWhinney. 2019. Understanding spoken language through TalkBank. *Behavior Research Methods*, 51(4):1919–1927.

Brian MacWhinney and Catherine Snow. 1985. The Child Language Data Exchange System. *Journal of Child Language*, 12(2):271–295.

Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Connor Mayer. 2020. An algorithm for learning phonological classes from distributional similarity. *Phonology*, 37(1):91–131.

Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, Greg Diamos, Pete Warden, Josh Meyer, and Vijay Janapa Reddi. 2021. Multilingual spoken words corpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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).

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

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.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2007. Buckeye corpus of conversational speech (2nd release). *Columbus, OH: Department of Psychology, Ohio State University*, pages 265–270.

Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Maxime Poli, Emmanuel Chemla, and Emmanuel Dupoux. 2024. Improving spoken language modeling with phoneme classification: A simple fine-tuning approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5284–5292.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Proc. Interspeech 2020*, pages 2757–2761.

Nairán Ramírez-Esparza, Adrián García-Sierra, and Patricia K Kuhl. 2017. Look who's talking NOW! Parentese speech, social context, and language development across time. *Frontiers in psychology*, 8:1008.

Nan Bernstein Ratner. 2024. Augmenting clinical insights with computing: How talkbank has impacted assessment and treatment of speech and language disorders. *Language Teaching Research Quarterly*, 44:31–40.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 174–188, Miami, FL, USA. Association for Computational Linguistics.

Tanja Schultz. 2002. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Interspeech*, volume 2, pages 345–348.

Zhewen Shen, Aditya Joshi, and Ruey-Cheng Chen. 2024. Bambino-lm:(bilingual-) human-inspired continual pre-training of babylm. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–7.

Jimin Sohn, Haeji Jung, Alex Cheng, Jooeon Kang, Yilin Du, and David R Mortensen. 2024. Zero-shot cross-lingual NER using phonemic representations for low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13595–13602.

Stefan Taubert. 2024. pinyin-to-ipa.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. *Treebanks: Building and using parsed corpora*, pages 5–22.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

J. Zhu, C. Zhang, and D. Jurgens. 2022. ByT5 model for massively multilingual grapheme-to-phoneme conversion. In *Proceedings of INTERSPEECH 2022*, pages 446–450.

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.

Victor W Zue and Stephanie Seneff. 1996. Transcription and alignment of the timit database. In *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*, pages 515–525. Elsevier.

A Limitations

We consider the following limitations of our work.

Phonemes as a representation of speech: While phonemic data more closely represents how words are pronounced compared to orthographic text (the degree of this difference varies between languages), phonemes are still abstract symbolic units which do not contain many of the detailed and continuous features of speech, such as prosody. They also abstract away from phones, which are detailed realizations of phonemes, representing the physical sound produced rather than a language-specific meaningful unit. When comparing modalities that may be close to the sensory signal available to infants for developmentally plausible language modeling, some researchers consider phonemic data to be as implausible as orthographic data (Lavechin et al., 2023) and instead create language models that can be trained directly on audio (Kamper et al., 2017; Nguyen et al., 2020; Hsu et al., 2021; Dunbar et al., 2021). Nevertheless, phonemes still provide a useful unit of analysis and are necessary for certain linguistic theories and information-theoretic calculations. While phones could offer another useful representation, they are even harder to source than phonemes.

G2P conversion inaccuracies: Despite improving G2P conversion by mapping to inventories in the Phoible database, there are still limitations with `G2P+`. Firstly, our method integrates existing G2P tools, which may vary in quality between languages. When converting each language in CHILDES, we selected the most appropriate backend for each language, in particular adding two backends to support G2P for Mandarin and Cantonese, but the quality may still vary. Many of the G2P tools for certain languages convert words individually, so we do not capture vowel reduction, allophonic variation or other differences found in natural speech. We also use a single accent for each language, losing inter-speaker variability. The `phonemizer` backend supports multiple accents for certain languages (here we use a different accent for EnglishNA and EnglishUK) and future work could try to maintain accent differences during grapheme-to-phoneme conversion, but this would require speaker information or audio, as was done during the creation of Audio BNC (Coleman et al., 2012). Finally, we note that G2P methods may not produce correct transcriptions for child-produced utterances, which are often corrected by the transcriber, especially for young infants. Initially we intended to distribute IPA CHILDES without child-produced utterances (and in this study only train models with the child-directed utterances) but as they might be useful in future research, we instead note this limitation.

Phoible inventories: Although the Phoible database collects established phonemic inventories and provides distinctive feature vectors, there are still often multiple phoneme inventories for a single language. This the exact phonemic inventory for a particular language is still a matter of debate among expert phonologists. When creating folding maps we choose the 'best-fitting' inventory to map to, as detailed in table 1, but we acknowledge that these inventories may not be exact.

Phoneme LMs: We train phoneme LMs on 11 languages from IPA CHILDES but the specific architecture we use is based on our scaling experiment for the EnglishNA model. Although we do not directly compare these LMs, we note the possibility that other parameters may have better suited the non-English languages. We were only able to conduct the scaling experiments for English due to the lack of phonological benchmarks for other languages but we hope that the release of IPA CHILDES facilitates further work in multilingual phonological evaluation of phoneme LMs.

Languages: Although our dataset is multilingual, there are still limitations in terms of language coverage. The languages are predominantly European and Asian, with no languages indigenous to the Americas, Australia or Africa. English is also still the dominant language of the dataset and the Farsi section is very small, only containing 43 thousand words. In creating this dataset, we were limited by the languages available in CHILDES. The languages in CHILDES we were not able to convert were Greek, Arabic, Hebrew, Thai, Georgian, Tamil, Taiwanese, Jamaican, Sesotho, Berber, Cree and Slovenian and Russian due to missing G2P backends or unsupported orthographies.

B Breakdown of IPA CHILDES

IPA CHILDES contains transcriptions of child-centered speech for 31 languages. Details of each language section are provided in table 1.

Language	CHILDES Collection	Backend	Inventory ID	Words	Phonemes	% Child
EnglishNA	EnglishNA (49)	phonemizer	2175	9,993,744	30,986,218	36
EnglishUK	EnglishUK (16)	phonemizer	2252	7,147,541	21,589,842	39
German	German (10)	epitran	2398	5,825,166	21,442,576	44
Japanese	Japanese (11)	phonemizer	2196	2,970,674	11,985,729	44
Indonesian	EastAsian/Indonesian (1)	epitran	1690	2,347,642	9,370,983	34
French	French (15)	phonemizer	2269	2,973,318	8,203,649	40
Spanish	Spanish (18)	epitran	164	2,183,992	7,742,550	46
Mandarin	Chinese/Mandarin (16)	pinyin_to_ipa	2457	2,264,518	6,605,913	39
Dutch	DutchAfricaans/Dutch (5)	phonemizer	2405	1,475,174	4,786,803	35
Polish	Slavic/Polish (2)	phonemizer	1046	1,042,841	4,361,797	63
Serbian	Slavic/Serbian (1)	epitran	2499	1,052,337	3,841,600	29
Estonian	Other/Estonian (9)	phonemizer	2181	843,189	3,429,228	45
Welsh	Celtic/Welsh (2)	phonemizer	2406	666,350	1,939,286	69
Cantonese	Chinese/Cantonese (2)	pingyam	2309	777,997	1,864,771	34
Swedish	Scandinavian/Swedish (3)	phonemizer	1150	581,451	1,782,692	45
PortuguesePt	Romance/Portuguese (4)	phonemizer	2206	499,522	1,538,408	39
Korean	EastAsian/Korean (3)	phonemizer	423	263,030	1,345,276	37
Italian	Romance/Italian (5)	phonemizer	1145	352,861	1,309,489	39
Croatian	Slavic/Croatian (1)	epitran	1139	305,112	1,109,696	39
Catalan	Romance/Catalan (6)	phonemizer	2555	319,726	1,084,594	36
Icelandic	Scandinavian/Icelandic (2)	phonemizer	2568	279,939	1,057,235	35
Basque	Other/Basque (2)	phonemizer	2161	230,500	942,725	49
Hungarian	Other/Hungarian (3)	epitran	2191	237,062	918,002	48
Danish	Scandinavian/Danish (1)	phonemizer	2265	275,170	824,314	42
Norwegian	Scandinavian/Norwegian (2)	phonemizer	499	227,856	729,649	43
PortugueseBr	Romance/Portuguese (2)	phonemizer	2207	174,845	577,865	44
Romanian	Romanian (3)	phonemizer	2443	152,465	537,669	43
Turkish	Other/Turkish (2)	phonemizer	2217	79,404	421,129	51
Irish	Celtic/Irish (2)	phonemizer	2521	105,867	338,425	34
Quechua	Other/Quechua (2)	phonemizer	104	46,848	281,478	40
Farsi	Other/Farsi (2)	phonemizer	516	43,432	178,523	40

Table 1: A breakdown of each language available in IPA CHILDES. The bracketed number in the **CHILDES Collection column** refers to the number of corpora downloaded from that collection. The **Backend**, **Lang Code** and **Phoneme Inventory** columns refer to the `G2P+` configuration used to convert utterances for that language to phonemes and the Phoible inventory used for that language in folding. The **Words** and **Phonemes** columns refer to the number of words and tokens in each subset and **% Child** refers to the percentage of the data that is spoken by a child.

C Dataset comparison

In section 2.1 we discuss previous phonemic datasets in relation to IPA CHILDES. We provide a full comparison of these datasets in table 2.

D `G2P+` Usage

`G2P+` is a python library that can be used as an API or as a command-line tool in order to convert orthographic text to a phonemic representation. The tool allows the user to select the backend and language code to use for G2P with text provided through filepaths or standard input. Additional options include `--keep_word_boundaries` to output a dedicated `WORD_BOUNDARY` token between words and `--uncorrected` to skip the folding process and output the phonemes exactly as produced by the backend tool. Each backend also supports individual options. For instance, `--split-tones` outputs tones as individual tokens instead of merging them with the syllabic phoneme for our two Chinese language backends. See the repository's `README.txt` for further details.

E Phoneme Stream Representation

In order to ensure that phonemes are output using a consistent representation, we define the **phoneme stream representation** as follows:

- Each phoneme is represented using the International Phonetic Alphabet (IPA).
- Each phoneme is separated by a space.
- Word boundaries and utterance boundaries are represented using unique symbols.

IPA is used to represent each phoneme due to being the most widely used and comprehensive phonetic alphabet. It is important to separate phonemes by spaces because IPA symbols may be represented

Dataset	Modality	Scale (words)	Domain	Languages
The Pile (Gao et al., 2020)	Orth	100B[†]	Web-scraped written text	English only
GlobalPhone (Schultz, 2002)	Orth, Phon, Audio	5M[†]	Read speech	22
CommonVoice (Ardila et al., 2020)	Orth, Audio	30M[†]	Read speech	38
VoxCommunis (Ahn and Chodroff, 2022)	Orth, Phon, Audio	23M[†]	Read speech	40
CMU Wilderness (Black, 2019)	Orth, Audio	170M[†]	Read speech	699
VoxClamantis (Salesky et al., 2020)	Orth, Audio, Phon	152M[†]	Read speech	635
TIMIT (Garofolo et al., 1993)	Orth, Phon, Audio	40k	Read speech	English only
FLEURS (Conneau et al., 2023)	Orth, Audio	15M[†]	Read speech	102
MSWC (Mazumder et al., 2021)	Orth, Audio	20M	Read speech	102
IPAPACK (Zhu et al., 2024)	Orth, Phon	15M[†]	Read speech	115
LibriSpeech (Panayotov et al., 2015)	Orth, Audio	10M[†]	Audio books	English only
Libri-Light (Kahn et al., 2020)	Orth,[*] Phon,[*] Audio	700M[†]	Audio books	English only
MLS (Pratap et al., 2020)	Orth,[*] Phon,[*] Audio	600M[†]	Audio books	8
Switchboard (Godfrey et al., 1992)	Orth, Phon, Audio	3M[†]	Telephone conversations	English only
Fisher (Cieri et al., 2004)	Orth, Audio	12M[†]	Telephone conversations	English only
Buckeye (Pitt et al., 2005)	Orth, Phon, Audio	300k	Spontaneous speech	English only
British National Corpus (Consortium, 2007)	Orth, Audio	100M	Written & spontaneous speech	English only
Audio BNC (Coleman et al., 2012)	Orth, Phon, Audio	7M	Spontaneous speech	English only
VoxLingua107 (Valk and Alumäe, 2021)	Audio	80M	Spontaneous speech	107
Babel (Harper, 2011)	Orth, Audio	60M	Telephone conversations	25
CHILDES (MacWhinney and Snow, 1985)	Orth	59M	Child-centered speech	45
BabyLM (Choshen et al., 2024)	Orth	100M	Speech and text[**]	English only
IPA CHILDES	Orth, Phon	45M	Child-centered speech	31

Table 2: A comparative summary of the datasets discussed in section 2.1. The datasets are described in terms of their modality, scale, domain and languages. IPA CHILDES is the first multilingual phonemic dataset of spontaneous speech and the first phonemic dataset of child-centered speech.

[†] *Word counts estimated from the size in bytes or the hours of audio in the dataset, using a heuristic based on the size of Switchboard of 5 bytes per word and 12,000 words per hour.*

[*] *Libri-Light and MLS only have orthographic and phonemic transcriptions for 10 hours of audio per language..*

[**] *BabyLM contains a mix of speech and text data from a mix of adult-directed and child-directed sources, only 29% is child-directed speech.*

using multiple Unicode characters. For instance, the word "enjoy" can be transcribed in IPA as ɛndʒɔɪ which uses six characters but only contains four phonemes, since dʒ is a single consonant and ɔɪ is a diphthong. By instead representing the word as ɛ n dʒ ɔɪ, it is much easier to split the word into individual phonemes by using whitespace as a delimiter. Similarly, word boundaries and utterance boundaries are represented using the unique symbols `WORD_BOUNDARY` and `UTT_BOUNDARY`.

F Folding Maps

Folding maps are primarily used to make surface-level adjustments, but they can also be used to solve several other error types in order to create a better alignment with a Phoible inventory. These errors are detailed in table 3.

The many-to-one mappings and those that split or merge tokens may alter the number of output tokens or types. Since such a mapping will change the information-theoretic properties of the output, it is important that they are linguistically motivated and carefully implemented.

In order to construct the folding map for each backend-language pair, we run `G2P+` on orthographic text for that language and compare the output set of phonemes P_O to the phonemes in the closest inventory in Phoible P_I. We call the set of phonemes present in P_O but not P_I the "unknown phonemes" U_K where $U_K = P_O \setminus P_I$ and the set of phonemes present in P_I but not P_O the "unseen phonemes" U_S where $U_S = P_I \setminus P_O$. We then construct the folding map as follows:

1. Find pairs $(k, s) \in U_K \times U_S$ that differ according to an accent or diacritic and obviously represent the same phoneme (determined by ruling out alternatives or examining where k is produced in the output). Create a one-to-one mapping $k : s$ for each such pair, e.g. t : tʰ.

2. Find pairs $(k, s) \in U_K \times U_S$ that clearly repre-

Error type	Consequence	Example
One-to-one: The backend uses one symbol for a phoneme but the inventory lists a different symbol for that phoneme.	The one-to-one mapping does not change the number of types or tokens in the output.	`phonemizer` with language code `sv` (Swedish) outputs n but the matching inventory uses n̪.
Many-to-one: The backend produces two different phonemes that should only map to a single phoneme in the inventory.	The many-to-one mapping reduces the number of phoneme types.	`phonemizer` with language code `pt` (Portuguese) outputs both ɾ and r but the matching inventory only lists ʁ.
Consonant merging: The backend outputs two symbols for a consonant that should be written as a single phoneme.	The mapping merges the pair of consonants, reducing the number of phoneme tokens produced.	`epitran` with language code `srp-Latn` (Serbian) outputs the sequence d ʒ but these are should be written as a single phoneme dʒ.
Vowel merging: The backend outputs a pair of vowels as separate phonemes but they are typically analysed as a single diphthong.	The mapping merges the pair of vowels, reducing the number of phoneme tokens produced.	`pingyam` with language code `cantonese` outputs the sequence o u but these are should be treated as a diphthong ou.
Vowel splitting: The backend outputs a diphthong that is not listed in the inventory and should be split into individual phonemes.	The mapping splits the pair of vowels, increasing the number of phoneme tokens produced.	`phonemizer` with language code `en-us` (North American English) outputs aɪʊ as a single phoneme but this should be aɪ ʊ.
Phoneme duplication: The backend outputs duplicate phonemes to represent long vowels or consonants or because of an error.	The mapping replaces the pair of phonemes with just one, reducing the number of phoneme tokens.	`phonemizer` with language code `et` (Estonian) outputs d d but should output the long consonant dː.
Diacritic error: The backend incorrectly outputs the diacritic as a separate symbol instead of attaching it to the phoneme.	The mapping may change the number of phoneme types or tokens.	`phonemizer` with language code `ko` (Korean) outputs the diacritic for aspiration as h instead of ʰ so sequences kh and ph are mapped to kʰ and pʰ.
Orthographic error: Due to an invalid symbol in the orthographic text, the backend outputs an incorrect phoneme.	The contextual mapping changes the frequency statistics for the resulting phoneme, possibly reducing the number of phoneme types.	`epitran` with language code `hun-Latn` (Hungarian) outputs ô when the orthographic letter ő is incorrectly written as ô and so the phoneme is mapped to øː.

Table 3: A list of errors that can occur during grapheme-to-phoneme conversion that can be fixed with a folding map but that may change the information-theoretic properties of the output.

sent the same phoneme (determined as above) but may use entirely different symbols, possibly due to an alternative transcription scheme. Create a one-to-one mapping for each pair, e.g. a : æ.

3. For remaining items $k \in U_K$, determine whether these result from one of the other errors in table 3. Carefully examine instances where k is produced in the output and create a suitable mapping $k : p$ for some $p \in P_I$ to solve the error (the mapping may need to be contextual or include several characters, e.g. ɚ : ə ɹ or ʊ ɔ : w ɔ).

4. For remaining items $s \in U_S$, determine whether these result from one of the other errors in table 3. Carefully examine instances where s should be produced in the output and create a suitable mapping $k : s$ for some $k \in P_O$ to solve the error (the mapping may need to be contextual or include several characters).

5. Examine the output for cases of **phoneme duplication** and other errors that may not contain phonemes in U_K or U_S but could still be solved with the phoneme map and create suitable mappings.

The goal is for $U_K = \{\} = U_S$ or equivalently $P_I = P_O$, i.e the set of phonemes produced by the tool perfectly aligns with the phoneme inventory in Phoible. This is not always possible, often there are a few remaining phonemes in U_K and/or U_S. This can occur when no obvious mappings could be found in steps 1–4 above. For example, the `epitran` backend for German does not produce the phoneme ʒ (it is "unseen") and none of the unknown phonemes seem to be a good match. Another possibility is that the output set of phonemes P_O may not align well with any of the Phoible phoneme inventories and so the closest match may not include some of the unknown phonemes $k \in U_K$ despite being valid phonemes for that language and listed in other inventories. For example, the `epitran` backend for German produce the phonemes x and ɐ which are not listed in the matching inventory but are listed in other established inventories for German. In other cases, the unknown phonemes may come from loan words (e.g. ts for "pizza" in Portuguese). Finally, there are some cases where the output considerably disagrees with all of the Phoible inventories but is a valid phonemic analysis of the language according to other sources.

See section 3.3 for an example of using `G2P+` for French, using the `phonemizer` backend with a

folding map to approach Phoible inventory 2269.

G Implementation Details

We conduct our experiments using the `PyTorch` framework (Paszke et al., 2019) and the `Transformers` library (Wolf et al., 2020).

G.1 Hardware Details

We use a server with one NVIDIA A100 80GB PCIe GPU, 32 CPUs, and 32 GB of RAM for all experiments. Below, we report a subset of the output of the *lscpu* command:

```
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Address sizes:         46 bits physical,
                       48 bits virtual
Byte Order:            Little Endian
CPU(s):                32
On-line CPU(s) list:   0-31
Vendor ID:             GenuineIntel
Model name:            Intel(R) Xeon(R)
                       Silver 4210R CPU
                       @ 2.40GHz
CPU family:            6
Model:                 85
Thread(s) per core:    1
Core(s) per socket:    1
Socket(s):             8
Stepping:              7
BogoMIPS:              4800.11
```

G.2 Model Parameters and Training Procedure

Parameter	Value
Max Example Length	128
Learning Rate	0.001
Optimizer	AdamW
Scheduler Type	Linear
Max Steps	200k
Warm-up Steps	60k
Per Device Batch Size	32

Table 4: Hyperparameter settings for training the GPT-2 architecture. Where values are not reported, they may be assumed to be default values.

We describe training parameters in table 4 and model sizes in table 5. Following the conventions of the Pythia suite of models (Biderman et al., 2023), we report the number of non-embedding parameters. Unlike their suite, where models are named according to the number of parameters, we name our models according to the number of non-embedding parameters. This is because we use the same architecture for multiple languages, each of which has a different vocabulary size according to the number of phoneme types in that language, which alters the total number of parameters. Our 1M, 19M and 85M models are equivalent to Pythia-14M, Pythia-70M and Pythia-160M, respectively. Our training scripts are available here.

Model Size	Layers	Heads	Embd	Inner
400k	2	4	128	512
600k	3	4	128	512
800k	4	4	128	512
1M	6	4	128	512
5M	6	8	256	1024
19M	6	8	512	2048
25M	8	8	512	2048
85M	12	12	768	3072

Table 5: GPT-2 model sizes used in the size requirement experiment. Where values are not reported, they may be assumed to be default values.

Data is prepared into batches by first tokenizing the entire dataset, combining all tokens into one long vector, and then splitting the vector into chunks of 128 tokens. Only the very last example is padded, if required. At each step during training, random chunks are selected and combined into batches.

Checkpoints are taken every 20,000 steps during training. At each checkpoint, the perplexity is evaluated on the held-back evaluation set, and at the end of training the checkpoint with the lowest perplexity is returned as the best model. For the smallest models, many of the best models were from the very first checkpoint, since due to the small training dataset and small model, the model had already fit the data by this point.

In our size requirement experiment (see section 5.1), we train each model in table 5 using a dropout of 0.1, 0.3 and 0.5 on each subset size of the EnglishNA portion of IPA CHILDES.

H Best Phoneme LM Parameters Across Data Scales

Following the size experiment in section 5.1, we report the model size and dropout values that achieved the highest BabySLM scores for each subsample size of the EnglishNA portion of IPA CHILDES in table 6.

I Average Information Density of Phonemized Child-Directed Speech Increases with Age Cross-Lingually

The phonemic representation of the utterances in our dataset open up new avenues for exploring

Data Size (words)	BabySLM Lexical			BabySLM Syntactic		
	Model Size	Dropout	Score	Model Size	Dropout	Score
80k	600k	0.3	65.8	400k	0.5	52.6
180k	800k	0.3	69.3	5M	0.5	52.3
500k	5M	0.3	72.9	5M	0.3	54.3
800k	19M	0.5	74.2	19M	0.1	54.9
1.8M	5M	0.3	77.4	19M	0.1	55.6
5M	19M	0.1	80.3	5M	0.3	58.3

Table 6: Best model sizes and dropout values for the BabySLM Lexical and Syntactic scores for each subset size of the EnglishNA corpus of IPA CHILDES.

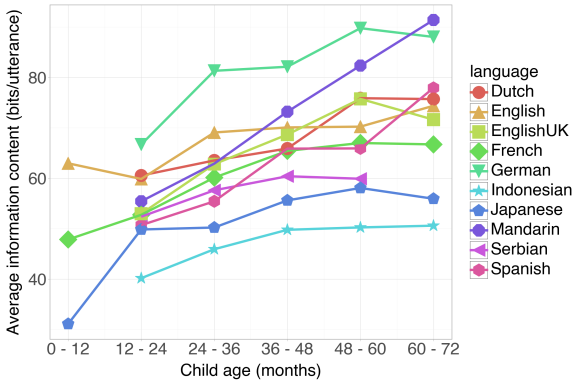


Figure 7: Average information of child-directed utterances in CHILDES

the phonotactic properties of languages and the information-theoretic properties of child-directed speech.

Here, we demonstrate one information-theoretic experiment, comparing the average information content of child-directed utterances to the age of the child being spoken to (this information is also available in CHILDES and is preserved in our dataset). We group child ages in years (0-12 months, 12-24 months, etc.) and calculate the average information content of a sample of child-directed utterances using a unigram language model. The information I_U of each utterance consisting of a sequence of phonemes $p_1, p_2, \ldots, p_n$ is given by

$$I_U = -\sum_{i=0}^{n} log_2 P(p_i),$$

where $P(p_i)$ is the probability of phoneme p_i given by its frequency in the data. We plot the average information of utterances in each age category for the largest 10 languages in the dataset in fig. 7. We find that across all 10 languages the average information of utterances increases with the age of the child, indicating that speakers of 'Parentese' may adjust the complexity of their speech according to the learner's age.