# BIPA: Brazilian Portuguese Phonetic Dataset with Dialectal Variations in IPA Standard

**Thiago Monteles de Sousa**[1] **and Lucas Rafael Gris**[1] **and Nádia Félix Felipe da Silva**[1]
[1]Federal University of Goias, Brazil
thiagomonteles@discente.ufg.br, lucas.gris@discente.ufg.br, nadia@inf.ufg.br

## Abstract

This work presents BIPA, a phonetic transcription corpus for Brazilian Portuguese that covers regional dialectal variations. The corpus was constructed through automated extraction from Wiktionary, resulting in 53,353 unique words and 350,021 transcriptions in IPA format, distributed across six dialects: general Brazilian, Rio de Janeiro, São Paulo, South Region, Northeast Region, and Center-West Region. The average density of 6.56 transcriptions per word reflects multiple regionally conditioned phonetic variations. To validate the utility of the corpus, the ByT5-small model was fine-tuned for grapheme-to-phoneme conversion, achieving a Minimum Phoneme Error Rate of 3.6% on the best epoch. BIPA addresses the scarcity of computational linguistic resources for Brazilian Portuguese, enabling applications in regional speech synthesis, automatic accent recognition, and computational sociolinguistic analysis.

## 1 Introduction

In the field of linguistics, phonemes are the smallest distinctive sound units of a language, capable of differentiating meanings when articulated in different contexts (Garay, 2016). For computational systems to process this language efficiently, precise conversion between its orthographic representation (graphemes) and the respective sound representation is essential. This fundamental task is known in computational linguistics as grapheme-to-phoneme (G2P) conversion (Unnithan and Mahapatra, 2024).

G2P models learn systematic correspondences between written character sequences and their phonetic pronunciations, allowing the inference of how unseen words should be articulated (Unnithan and Mahapatra, 2024). These technologies are widely applied in speech synthesis systems (text-to-speech, TTS) (Tan et al., 2021) and automatic speech recognition (ASR) (Kheddar et al., 2024), constituting a critical component for naturalness and the modeling of human-computer interfaces.

Although Portuguese is the eighth most spoken language in the world, with more than 260 million native speakers (Eberhard et al., 2025), the current context is a critical scarcity of high-quality computational linguistic resources for Brazilian Portuguese (Corrêa et al., 2024; Almeida et al., 2024). Compared to languages such as English, the availability of large annotated corpora, pre-trained models, and specialized datasets for Brazilian Portuguese remains significantly lower (da Rocha Junqueira et al., 2024; Cruz-Castañeda and Amadeus, 2025). Regional differences within Brazilian Portuguese itself, dialectal phonetic variations, and specific vocabulary further increase this computational challenge (Contributors, 2024; Lima et al., 2024). Recent initiatives, such as the TaRSila speech synthesis project (Contributors, 2024) and spontaneous speech data with regional annotation (Lima et al., 2024), have sought to bridge this gap.

In the context of Brazilian Portuguese phonemes, phonology is fundamentally oriented by syllabic and accentual properties, in which many phonological processes are related to or conditioned by the syllabic structure and the position of the primary stress (Souza, 2025). Several phonological phenomena affect vowels in different ways, depending on their position in the syllable (Kenstowicz and Sandalo, 2016). An example occurs in syllables preceding the tonic one (the syllable pronounced with the greatest intensity in the word), where vowels tend to be influenced by the following tonic vowel. This can be observed in words like *bonito*, where the vowel "o" in the first syllable is influenced by the proximity of the tonic "i", becoming pronounced as "u" (Battisti, 2017), and thus its phonetic form is "bu'ni.tu".

Several other rules are part of phoneme formation, such that, as character sequences unite in



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 478–487

the creation of the grapheme, such orthographic configurations condition the resulting phonetics. Thus, correctly identifying letter positions, accents, and other elements is essential for systems converting text to speech or pronunciation synthesis, ensuring natural and correct phonetic transcriptions (Casanova et al., 2023).

The IPA (International Phonetic Alphabet) standard, developed by the International Phonetic Association, constitutes a standardized system of sound representation that allows transcribing phonemes of any language through specific symbols (Association, 1999). Unlike conventional orthography, which often uses combinations of letters to represent unique sounds, IPA establishes a distinct symbol for each sound, eliminating ambiguities and offering sufficient resources to faithfully represent the phonetic richness of the language, including its regional dialectal variations (International Phonetic Association, 2015). This characteristic makes IPA particularly valuable for speech processing applications and word conversion to phonemic representation, where acoustic precision is critical.

Given this context, the main contributions of this study include: (i) The BIPA Corpus containing grapheme-phoneme pairings of Brazilian Portuguese[1]; (ii) A neural ByT5-small model (Xue et al., 2022) for the grapheme-to-phoneme conversion task trained on the BIPA Corpus[2].

In the following sections, related works to the research theme of this article, the methodology of BIPA Corpus construction, statistical analyses on dialectal coverage and phonetic variation, results obtained through the model trained on the proposed Corpus, as well as conclusions on the applicability of this resource to Brazilian Portuguese will be presented.

## 2 Related Works

Several works in the literature use IPA as a standard system for phonemic representation. Deri and Knight (2016) extracted more than 650,000 word and pronunciation pairs from Wiktionary, in addition to being combined with phonemic inventories from the Phoible base (Moran and McCloy, 2019). The authors do not detail which languages and quantity of data extracted. Lee et al. (2020) developed WikiPron, a tool that extracted 1.7 million pronunciations in 165 languages in IPA. However, its representation for Brazilian Portuguese is 9,315 grapheme-phoneme pairs. Goriely and Buttery (2025) proposed G2P+, a tool for orthographic conversion aligned with Phoible inventories in 31 languages and other sets that use IPA as a standard form, including Brazilian Portuguese, extracted from the eSpeak software[3]. The authors use phonetization rules to produce phonetic representation. Mendonça and Aluísio (2014) focuses exclusively on Brazilian Portuguese, developing a hybrid pronunciation dictionary using manual transcription rules and IPA, representing 108,389 G2P pairs focusing on the standard variety without considering regional differences.

The existence of these grapheme-to-phoneme corpora allows for the creation of translation systems. As an example, we can cite Peters et al. (2017), who developed a multilingual neural sequence-to-sequence system for hundreds of languages, achieving 11% in Phone Error Rate (PER). Řezáčková et al. (2021) performed a fine-tuning on the T5 model for G2P at the sentence level for English and Czech, obtaining a word accuracy of 99.04% and 99.89%, respectively. Dong et al. (2022) proposed GBERT (Grapheme BERT), a model pre-trained on grapheme sequences for Dutch, Serbo-Croatian, Bulgarian and Korean. Kim et al. (2022) developed a bilingual model for English and Korean based on non-autoregressive transformer (NART-CRF), reaching an inference speed 27 times greater than that of autoregressive models, with an accuracy of 87.75%.

Solving this task requires mapping words to their articulatory representations so that successful models can learn such correspondences and infer new phonemes never before observed (Unnithan and Mahapatra, 2024). Due to this, works like Zhu et al. (2022) implemented G2P models based on the ByT5 architecture for approximately 100 languages, demonstrating that ByT5, operating directly on byte-level inputs (or characters), significantly outperforms the mT5 model that works at the word level. The study aggregated pronunciation dictionaries from multiple sources, totaling 7.2 million words distributed in 99 languages in the training set. Experimental results showed that the multilingual ByT5-small model reached a PER of 8.8% and Word Error Rate (WER) of 25.9%, surpassing mT5-small which obtained a PER of

[1] https://huggingface.co/datasets/thiagomonteles/BIPA

[2] https://huggingface.co/thiagomonteles/BIPA_g2p_Multidialect_Byt5

[3] https://github.com/espeak-ng/espeak-ng

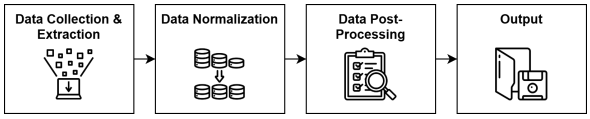


Figure 1: Steps of content extraction for the generation of the BIPA dataset.

11.9% and a WER of 37.1%.

Literature analysis demonstrates a significant gap in research with a specific focus on Brazilian Portuguese and, in particular, on computational modeling of dialectal variation. The work of Mendonça and Aluísio (2014) is the only one with an exclusive focus on Brazilian Portuguese and develops a pronunciation dictionary, but does not consider regional dialectal variations nor more robust techniques for G2P systems. Other multilingual works that incorporate Brazilian Portuguese treat it only as one among dozens or hundreds of languages, without specific treatment for phonological characteristics or geographical variation. No work was identified that develops G2P models capable of processing multiple dialects of Brazilian Portuguese, considering that Brazil possesses significant phonological diversity and that dialect-sensitive G2P systems are essential for the development of more inclusive speech technologies.

## 3 Corpus Construction

The Brazilian IPA Standard Corpus (BIPA) was built through automated data extraction processes[4] of phonetic transcriptions from Wiktionary[5], a platform that provides pronunciation annotations for Brazilian Portuguese, including dialect-specific markings following the IPA standard. These transcriptions are collaboratively contributed by volunteer editors following the platform's official pronunciation conventions[6], grounded on established phonological references such as (Barbosa and Albano, 2004), with manual dialectal annotations ensuring regional coverage across Brazilian varieties.

### 3.1 Extraction Process

As illustrated in Figure 1, the BIPA Corpus construction pipeline begins with the Wiktionary data source[7], from where phonetic transcriptions of

- (*Brazil*) IPA(key): /ze'ɾa(ʁ)/ [ze'ɾa(h)]
  - (*São Paulo*) IPA(key): /ze'ɾa(ɾ)/
  - (*Rio de Janeiro*) IPA(key): /ze'ɾa(ʁ)/ [ze'ɾa(χ)]
  - (*Southern Brazil*) IPA(key): /ze'ɾa(ɻ)/

Figure 2: Example of pronunciation present for the word "*zerar*" extracted from *Wiktionary*.

Brazilian Portuguese are extracted. In step 1 (Data Collection and Extraction), the implemented engine processes pages categorized as "Portuguese terms with IPA pronunciation", applying HTTP requests and HTML parsing to locate sections delimited by content identifications on the web page. Then, regular expressions are used to extract phonetic elements and identify dialectal labels in adjacent tags as illustrated in Figure 2.

In the second step (Data Normalization) of Figure 1, raw IPA transcriptions are subjected to standardization processes. In this phase, each Unicode symbol is validated according to the official IPA inventory, which comprises 107 basic segmental symbols (consonants and vowels) and 44 modifier diacritic marks (accentuation, nasality, among other phonetic properties), ensuring consistency in marker representation. Simultaneously, dialectal labels undergo standardization, employing the $pt-BR-UF$ scheme for regional variants. Duplicate records are identified through pair comparison (word, normalized IPA), with contextual resolution applied to homographs when multiple transcriptions coexist for the same orthographic form.

Proceeding in the sequence of Figure 1, step 3 (Data Post-Processing) applies filters to ensure the quality and linguistic coherence of the BIPA Corpus. For this purpose, the following items are excluded: identified by the presence of non-native allophone phonemes[8]; proper names, except when they present dialectally marked phonetic variation and lexicographical relevance; and abbreviations and acronyms, due to the irregularity of their phonetic realizations.

In step 4 (Output), the Corpus is serialized in *JSON Lines* (.jsonl) format. Each line of the file contains a record associated with a word, composed of three main pieces of information: (i) Dialect (regional variant or dialectal mark, such as Rio de

[4]Data extracted on 09/17/2025.

[5]en.wiktionary.org

[6]https://en.wiktionary.org/wiki/Appendix:Portuguese_pronunciation

[7]Data extracted from Wiktionary, available under CC BY-SA 4.0 license. Modifications made to data format and structure. Accessed at: https://www.wiktionary.org

[8]Considering the phoneme as the smallest distinctive unit, an allophone is the contextual variation of the same phoneme; 'non-native' designates sounds outside the standard term.

Janeiro or South Region); (ii) Word (standard orthographic form); and (iii) Pronunciation (normalized phonetic transcription in IPA symbols).

## 4 Statistical Analysis of the Corpus

### 4.1 Size and Coverage

The BIPA Corpus comprises 53,353 unique lexical entries, distributed in 350,021 total phonetic transcriptions, which contemplate the following dialectal variations for Brazilian Portuguese: Brazil (standard understanding), Rio de Janeiro, São Paulo, South Region, Northeast Region, and Center-West Region.

After the lemmatization procedure, the corpus was reduced to 46,518 unique lemmas and 294,578 corresponding transcriptions, which indicates that approximately 12.8% of the original entries correspond to morphological variants such as plural forms, verb inflections, and derivational processes of the same lexical root.

The transcription density per entry (with an average of 6.56 transcriptions per word) reflects the phonetic variability of the Corpus, where high-frequency words often exhibit multiple phonetic variations conditioned by phonetic and sociolinguistic factors. This granularity supports analyses across dialectal varieties, potentially serving as a resource for regional speech synthesis, automatic accent recognition, and other speech-related tasks (Contributors, 2024).

### 4.2 Dialectal Distribution

As presented in Table 1, the distribution of transcriptions among the six identified dialects reveals a marked asymmetry, with concentration in annotations classified as standard Brazilian Portuguese, representing 52.67% of the corpus (184,369 transcriptions). State and specific regional varieties exhibit decreasing representation: Rio de Janeiro contributes 22.51% (78,786 transcriptions), São Paulo 14.71% (51,487), and the South Region 9.98% (34,941). The Northeast and Center-West regions show extremely limited coverage, with 0.10% (363 transcriptions) and 0.02% (75 transcriptions), respectively.

This distribution partly reflects Wiktionary's documentary bias, where urban dialects associated with major population centers such as Rio de Janeiro and São Paulo tend to receive comparatively greater lexicographical coverage relative to varieties spoken in less populated regions.

| Dialect | Quantity | % |
|---|---|---|
| Brazil | 184,369 | 52.67 |
| Rio de Janeiro | 78,786 | 22.51 |
| São Paulo | 51,487 | 14.71 |
| South Region | 34,941 | 10.00 |
| Northeast Region | 363 | 0.10 |
| Center-West Region | 75 | 0.02 |

Table 1: Distribution by dialect (BIPA).

The vocabulary overlap analysis indicates that approximately 31.2% of words in the BIPA dataset exhibit two or more distinct dialectal variants, indicating a moderate level of phonological heterogeneity among Brazilian varieties. Words exhibiting three or more dialectal realizations account for approximately 8.4% of the corpus, primarily concentrated in high frequency lexical items and everyday vocabulary susceptible to sociolinguistic variation.

### 4.3 Phonetic Inventory and Dialectal Variation

Table 2 presents the ten phones exhibiting the highest degree of variation, based on their frequency across the three main dialects represented in BIPA: Rio de Janeiro (RJ), São Paulo (SP), and the South Region (Sul). Each row lists a phonemic variant illustrated with an example word from the Corpus, along with its corresponding frequencies per thousand (‰)[9] in RJ, SP, and the South Region.

| Orthographic word | Transcription | RJ | SP | Sul |
|---|---|---|---|---|
| rota | ˈ**h**ɔ.t**a** | 0.00 | 0.0 | 0.20 |
| rota | ˈ**ʁ**ɔ.t**a** | 0.00 | 0.0 | 0.20 |
| rota | ˈ**ʁ**ɔ.t**ɐ** | 0.18 | 0.0 | 0.00 |
| abidos | ˈa.bi.dos | 0.00 | 0.0 | 0.11 |
| pegas | ˈpɛ.ɣas | 0.00 | 0.0 | 0.11 |
| esfinge | esˈfĩ.ʒe | 0.00 | 0.0 | 0.11 |
| açores | aˈso.ɾes | 0.00 | 0.0 | 0.11 |
| agarta | aˈɣa ɾ.ta | 0.00 | 0.0 | 0.11 |
| rota | ˈ**χ**ɔ.t**ɐ** | 0.09 | 0.0 | 0.00 |
| rota | ˈ**h**ɔ.t**ɐ** | 0.09 | 0.0 | 0.00 |

Table 2: Top Contrastive Phonemes (freq. ‰).

The data reveal systematic patterns of consonantal and vocalic variation. The word *rota*, for example, illustrates this variability through five distinct pronunciations documented in the Corpus, which result from combining three variants of the initial rhotic ("r" sounds) (h, ʁ, χ) with two variants of

[9]The unit "per thousand" (‰) is used to facilitate comparison between dialects, indicating how many occurrences there are for every thousand records.

the final vowel (a or ɐ).

In the South of Brazil, ˈhɔ.ta and ˈʁɔ.ta predominate, both with a frequency of 0.20‰; in Rio de Janeiro, ˈʁɔ.tɐ (0.18‰), ˈχɔ.tɐ (0.09‰), and ˈhɔ.tɐ (0.09‰) are observed.

This variation [h] ∼ [ʁ] ∼ [χ] is characteristic of the Brazilian phonological system, where the same initial "r" can be pronounced in different ways (Amaral, 1920). The prevalence of [h] in the South region reflects a trend documented in regional sociolinguistic studies, where this variant constitutes the standard form in specific contexts (Brescancini and Monaretto, 2008).

The lack of substantial frequencies in the capital variety of São Paulo for several contrastive items suggests a possible underrepresentation of this variety in particular phonetic environments, or alternatively indicates that the Paulistano dialect may occupy an intermediate position in the Brazilian dialectal continuum, sharing traits with both Carioca (RJ) and Southern varieties[10] (de Oliveira and de Lurdes Zanolli, 2021).

### 4.4 Vocabulary Overlap Between Dialects

As illustrated in Figure 3, the similarity matrix quantifies the degree of lexical overlap among the four principal dialects represented in BIPA. To compute this similarity, we employed the *Jaccard* coefficient, defined by equation 1, where $A$ and $B$ denote the word sets associated with two distinct dialects:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

where $|A \cap B|$ represents the number of shared words between the two dialects and $|A \cup B|$ the size of the combined total vocabulary. The coefficient varies between 0 (no words in common) and 1 (identical vocabularies), providing a normalized measure of lexical overlap.

The analysis considers all words without lemmatization, capturing morphological and orthographic variations specific to each dialectal variety.

The analysis reveals asymmetrical patterns of vocabulary overlap. The standard dialect ("Brasil") presents low similarity with specific regional varieties ($J = 0.21$–$0.44$), reflecting the complementary nature between these sets, where transcriptions marked as "Brasil" typically correspond to words

[10] *Paulistano* and *Carioca* refer to the dialects spoken in the region of São Paulo and Rio de Janeiro, respectively.

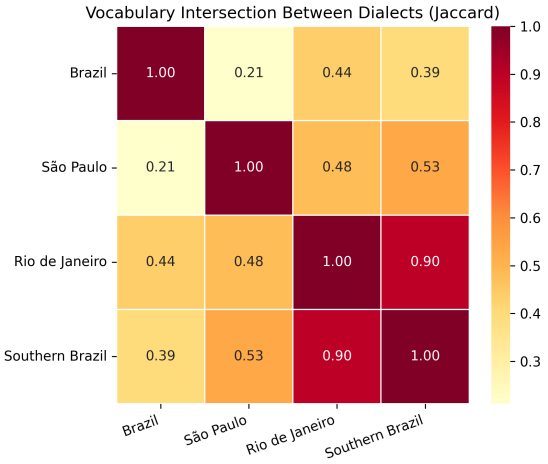


Figure 3: Vocabulary intersection between dialects using Jaccard coefficient, with all words unlemmatized. Chromatic scale: values close to 1.0 (dark red) indicate high overlap; values close to 0.0 (light yellow) indicate low overlap.

without documented dialectal variation, while regionally specified entries capture phenomena of geographically conditioned phonetic variation.

Among regional dialects, Rio de Janeiro and South Region exhibit surprisingly high overlap ($J = 0.90$), suggesting that a large part of the vocabulary with documented dialectal variation is present in both varieties, albeit with distinct phonetic realizations. This high overlap reflects the *Wiktionary* collection methodology, where words with multiple regional pronunciations receive transcriptions for various localities simultaneously.

São Paulo presents moderate overlap with Rio de Janeiro ($J = 0.48$) and the South region ($J = 0.54$), positioning itself in this context as an intermediate variety in the Brazilian dialectal space. The lower overlap of São Paulo with other varieties may indicate: (i) Lower lexicographical coverage of this variety in *Wiktionary*; (ii) greater vocabulary specificity of the Paulistano dialect; or (iii) documentation bias favoring RJ and South Region contrastive pairs to the detriment of SP.

The low similarity between "Brasil" and specific dialects ($J < 0.44$) confirms the functional partitioning of the Corpus. The generic label marks the absence of documented variation, while regional labels (RJ, SP, South Region) signal the presence of geographically distributed phonetic alternatives.

This distribution of variation present in BIPA may culminate in significant implications for regional speech synthesis systems and automatic ac-

cent identification, made possible through the use of the current set with specific sampling for the proposed target problem.

## 5 Experiments and Results

In order to illustrate the BIPA dataset's effectiveness, an experiment was implemented to develop a neural model for the grapheme-to-phoneme conversion task. The aim of this section is to verify the quality and utility of the collected data by training a model that can predict phonetic transcriptions in IPA notation from words written in standard orthography.

### 5.1 Architecture and Configuration

To demonstrate the applicability of the BIPA dataset in the grapheme-to-phoneme (G2P) conversion task, an experiment was conducted performing a complete *fine tuning* on the ByT5-small model (Xue et al., 2022). ByT5 (*Byte-level T5*) represents a transformers encoder-decoder architecture (Vaswani et al., 2017) that operates directly on UTF-8 byte sequences, dispensing with conventional tokenization and allowing modeling of phonetic relations at the character level. This enables the model to identify patterns, eliminating the necessity for conventional G2P systems to fragment the word into characters and apply conversion rules. This is a desirable attribute for the generalization of translation systems.

Figure 4 illustrates the ByT5 architecture applied to the G2P task in the BIPA context, where an input in the format "[dialect] grapheme" is encoded in UTF-8 byte sequences, processed by a heavy encoder with 3x more layers than the light decoder and decoded to generate phonetic transcriptions in IPA. With this asymmetric characteristic of ByT5, the encoder can acquire phonetic representations at the input character level with a greater degree of latent vision, whereas the more compressed decoder can generate the phonetic sequence more efficiently byte by byte.

In this experiment, the dataset contains 292,862 unique grapheme-phoneme pairs (eliminating duplicates for the trio dialect, word, and phoneme), distributed among six regional dialects of Brazilian Portuguese: Brazil (standard), Rio de Janeiro, São Paulo, South Region, Northeast Region, and Center-West Region. A multidialectal strategy was adopted, utilizing all available dialects in training.

Data were split using stratification by dialect and word length measured in characters. Because the dataset contains only unique $(\text{dialect}, \text{word}, \text{phoneme})$ tuples, we enforced a group constraint at the $(\text{dialect}, \text{word})$ level so the same $(\text{dialect}, \text{word})$ pair could not appear across different splits. Partitioning was performed in two stages with stratified group splitting: first, an approximately (10%) test set was held out, and then an approximately (10%) validation set (of the total) was split from the remaining data. The final partition sizes were (219,921) training examples, (43,699) validation examples, and (29,242) test examples.

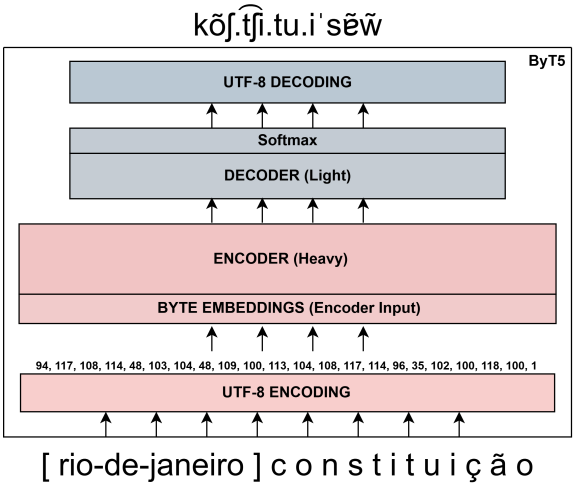


Figure 4: ByT5 architecture adapted for G2P training in the BIPA context.

### 5.2 Hyperparameters and Training

Training was configured with a learning rate of $1 \times 10^{-4}$, Adafactor optimizer (Shazeer and Stern, 2018) with linear decay and *warm-up* of 10% of iterations, weight decay of 0.05, and label smoothing factor of 0.1. 10 complete epochs were executed over the training set, with a *batch size* of 96 examples per device. Decoding during inference employed *beam search* (Graves, 2012) with width 6 and maximum output length of 35 *tokens*. Validation metrics were computed at the end of each epoch. Implementation was performed using the Transformers library (Wolf et al., 2020).

### 5.3 Evaluation Method

The basis of the adopted metric was the Phoneme Error Rate (*Phoneme Error Rate*, PER), calculated as the normalized Levenshtein distance (Levenshtein, 1966) between predicted and reference phoneme sequences, considering each IPA symbol

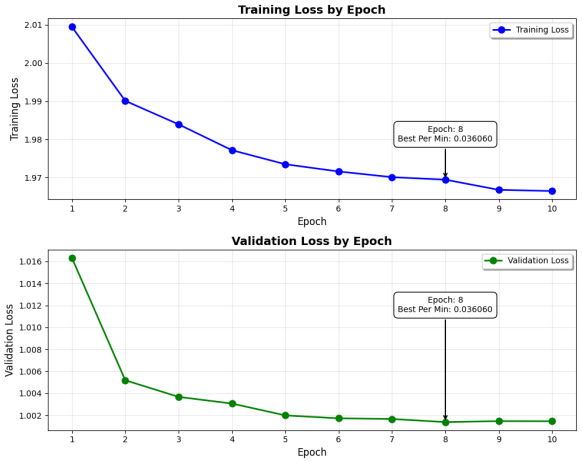


Figure 5: Evolution of training loss and validation loss by Epoch with best per min.

as an atomic unit. PER is defined by the equation:

$$PER = \frac{S + D + I}{N} \tag{2}$$

In Equation 2, $S$ represents the number of substitutions (incorrectly predicted phonemes), $D$ the number of deletions (phonemes omitted by the model), $I$ the number of insertions (erroneously added phonemes), and $N$ the total number of phonemes in the reference sequence. Lower PER values indicate greater precision in grapheme-phoneme conversion, with PER = 0% representing perfect agreement between prediction and reference.

Given the nature of the Corpus, which presents multiple phonetic variations for the same word, Minimum PER was used as a performance metric, in accordance with recent methodological approaches for handling phonemic ambiguity (Grafé et al., 2025). Thus, Minimum PER is calculated as:

$$\text{PER}_{\min}(\hat{y}, \{y_1, y_2, \ldots, y_K\}) \tag{3}$$

where $\hat{y}$ represents the phoneme sequence predicted by the model, and $\{y_1, \ldots, y_K\}$ denotes the set of valid reference pronunciations for a given word in a specific dialect. The operator $\text{PER}_{\min}$ returns the lowest PER value among all comparisons between the predicted sequence and the available pronunciation variants.

### 5.4 Results

Training was executed for 10 complete epochs and presented progressive convergence, as illustrated in Figure 5.

In the first epoch, a training loss of 2.0095 and validation loss of 1.0163 were observed, with a Per Min of 4.57%. The model demonstrated consistent convergence, with training loss steadily decreasing from 2.0095 in the first epoch to 1.9664 in the tenth epoch. The validation loss followed a similar trajectory, dropping from 1.0163 to 1.0015, reaching its minimum value of 1.0014 in epoch 8, which also registered the best Per Min of 3.6%. Starting from the sixth epoch, the model exhibited stable behavior, with validation loss remaining between 1.0015 and 1.0017 through the tenth epoch, suggesting that the learning process had reached convergence. The evolution illustrated in Figure 5 reflects successful acquisition of G2P mappings and dialectal variation patterns,

Table 3 presents model performance segmented by dialect, revealing substantial differences correlated to the sample size of each regional variety.

| Dialect | No. Examples | ↓ Min. PER (%) |
|---|---|---|
| São Paulo | 1,141 | 1.93 |
| Rio de Janeiro | 2,373 | 3.08 |
| Brazil (general) | 5,316 | 6.07 |
| South Region | 2,131 | 2.36 |
| Northeast Region | 32 | 78.12 |
| Center-West Region | 7 | 42.86 |

Table 3: Model performance (Minimum PER) by dialect in the validation set.

The results demonstrate a strong correlation between training data volume and prediction quality. Dialects with greater representation in the corpus, such as São Paulo (N=1,141, Min PER=1.93%), Rio de Janeiro (N=2,373, Min PER=3.08%), and South Region (N=2,131, Min PER=2.36%), presented superior performance, while underrepresented dialects such as Northeast Region (N=32, Min PER=78.12%) and Center-West Region (N=7, Min PER=42.86%) exhibited substantially higher error rates. This pattern is typical of low-resource settings, where limited data availability prevents adequate capture of dialect-specific phonological patterns. The category "Brazil (general)" (N=5,316) achieved intermediate performance (Min PER=6.07%), consistent with its function of representing a generic phonological variety.

To evaluate the performance of the BIPA-ByT5 model against its Multilingual ByT5 counterpart (Zhu et al., 2022), both models were assessed on the test set of the standard Brazilian dialect using IPA transcriptions. Table 4 presents the performance comparison:

The results demonstrate that the BIPA-ByT5-

| Model | ↓ Min. PER (%) |
|---|---|
| BIPA-ByT5 | 6.07 |
| CharsiuG2P Zhu et al. (2022) | 29.32 |
| BIPA-ByT5 (Normalized) | 4.81 |
| CharsiuG2P Zhu et al. (2022)(Normalized) | 6.4 |

Table 4: Comparison between BIPA-ByT5 model and Multilingual ByT5 (CharsiuG2P) for Brazilian Portuguese G2P conversion.

G2P model achieved substantially superior performance, with a minimum PER of 6.07%, representing a reduction of approximately 79.3% in phonetic error compared to the Multilingual ByT5 model (Min PER=29.32%). Additionally, a normalization experiment was conducted in which prosody marks, such as primary stress and period (*e.g.*, ˈpra.tɐ" to pratɐ"), were removed from input phonetic sequences. In this condition, both models showed substantial improvement in minimum PER, with BIPA-ByT5 reducing the error to 4.81% and CharsiuG2P to 6.40%, which suggests that part of the difficulty of the multilingual model is precisely associated with its lower capacity to explore explicit prosodic information due to the information used in training.

| **Center-West Region – *inteiro*** | |
|---|---|
| Prediction | ĩte(j).ru |
| Targets | [ĩte(j).ɾʷ] |
| Min PER ↓ | 10.00% |
| Result | Brazil Dialect |
| **South Region– *noite*** | |
| Prediction | ˈnoj.te |
| Targets | [ˈnoj.te], [ˈnoɪ̯.te] |
| Min PER ↓ | 0.00 |
| Result | Correct |
| **São Paulo – *ectomorfo*** | |
| Prediction | ˈɛ̥k.toˈmɔh.fu |
| Targets | [ˈɛ̥k.toˈmoʁ.fu], [ˈɛ̥k.toˈmoh.fu], [ˈɛ̥k.toˈmor.fu] |
| Min PER ↓ | 7.69% |
| Result | Vowel error [ɔ] → [o] |
| **Northeast Region – *mainha*** | |
| Prediction | mɐ̃ˈi.ɲa |
| Targets | [mɐ̃ˈĩ.ɐ] |
| Min PER ↓ | 14.29% |
| Result | Brazil Dialect |

Table 5: Examples of model predictions illustrating perfect hits, partial errors, and crossed dialect errors.

Table 5 illustrates four prediction scenarios of the model. A perfect hit is observed in the South Region, where *noite* reaches Min PER = 0.00%, demonstrating accurate phonological mapping for well-represented dialects. Partial errors appear in São Paulo, as in *ectomorfo* (Min PER = 7.69%), with vowel confusion between [ɔ] and [o], and in the Center-West Region, where *inteiro* (Min PER = 10.00%) reveals a cross-dialect regression to the "Brazil (general)" standard, predicting [ĩte(j).ru] instead of the expected variant [ĩte(j).ɾʷ]. The most critical case is observed in the Northeast Region, where *mainha* reaches Min PER = 14.29%, with the model again regressing to the "Brazil (general)" standard rather than producing the expected variant [mɐ̃ˈĩ.ɐ]. These results indicate that neural models prioritize majority patterns in unbalanced data scenarios, reinforcing the need for greater sample coverage for minority dialects.

## 6 Conclusion

This work addressed the challenge of building grapheme-to-phoneme (G2P) conversion systems for Brazilian Portuguese that account for regional dialectal variation, responding to the scarcity of resources for this language spoken by more than 260 million people (Eberhard et al., 2025). Through automated extraction from Wiktionary, the BIPA dataset was constructed containing 53,353 unique words and 350,021 phonetic transcriptions in IPA format, distributed among Brazilian dialects (Rio de Janeiro, São Paulo, South Region, Northeast Region, and Center-West Region). To demonstrate the effectiveness of the dataset, the ByT5-small model was fine-tuned for 10 epochs, achieving a Minimum PER of 3.6% in the best epoch. The public availability of the corpus and trained model contributes to democratizing technological resources for Brazilian Portuguese, enabling advances in digital accessibility and preservation of the language's phonetic diversity.

Future research directions include integrating the BIPA Corpus into the training of text-to-speech (TTS) and automatic speech recognition (ASR) systems, which represents a promising approach for evaluating the naturalness and intelligibility of predictions in real-world contexts. Additionally, future work should investigate model architectures optimized for inference speed and reduced size, as well as specialized architectures for few-shot scenarios to mitigate dialectal imbalance.

## Acknowledgments

The authors would like to acknowledge Brazilian Research Agencies FAPEG (Process: 202510267001606), CNPq (N 18/2024), and Center for Excellence in Artificial Intelligence (CEIA), Instituto de Informática, Universidade Federal de Goiás.

## References


Rúben Almeida, Ricardo Campos, Alípio Jorge, and Sérgio Nunes. 2024. Indexing portuguese nlp resources with pt-pump-up. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 178–181.

Amadeu Amaral. 1920. *O dialeto caipira*. Casa Editora O Livro, São Paulo, SP.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Plínio A. Barbosa and Eleonora C. Albano. 2004. Brazilian Portuguese. *Journal of the International Phonetic Association*, 34(2):227–232.

Elisa Battisti. 2017. Harmonia vocálica de altura no português brasileiro em formas nominais não derivadas: análise de um processo variável pela teoria da otimidade.

Claudia Brescancini and Valéria Neto de Oliveira Monaretto. 2008. Os róticos no sul do brasil: panorama e generalizações. *Signum: Estudos da Linguagem*, 11(2):51–66.

Edresson Casanova, Vinícius Gonçalves dos Santos, Flaviane Romani Fernandes Svartman, Marli Quadros Leite, Arnaldo Candido Junior, Ricardo Marcondes Marcacini, Solange Oliveira Rezende, and Sandra Maria Aluísio. 2023. Recursos para o processamento de fala. *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*.

TaRSila Project Contributors. 2024. Tarsila: Growing speech datasets for brazilian portuguese.

Nicholas Kluge Corrêa, Aniket Sen, and 1 others. 2024. Tucano: Advancing neural text generation for portuguese. *Patterns*. University of Bonn.

William Alberto Cruz-Castañeda and Marcellus Amadeus. 2025. Large languages models in brazilian portuguese: A chronological survey. *Journal of the Brazilian Computer Society*, 31(1):1168–1187.

Júlia da Rocha Junqueira, Émerson Lopes, Larissa Freitas, and Ulisses B Correa. 2024. A systematic analysis of multilingual and low-resource languages models: A review on brazilian portuguese.

Márcia Santos Duarte de Oliveira and Maria de Lurdes Zanolli. 2021. O /r/ retroflexo no português caipira como resultado de "interferência" da língua geral de são paulo: uma homenagem à obra de amadeu amaral. *Estudos Linguísticos*, 50(3):1159–1172.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408.

Lu Dong, Zhi-Qiang Guo, Chao-Hong Tan, Ya-Jun Hu, Yuan Jiang, and Zhen-Hua Ling. 2022. Neural grapheme-to-phoneme conversion with pre-trained grapheme models. *Preprint*, arXiv:2201.10716.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. Ethnologue: Languages of the world.

Rodrigo Garcia Garay. 2016. O fonema: linguística e história.

Zébulon Goriely and Paula Buttery. 2025. Ipa-childes g2p+: Feature-rich resources for cross-lingual phonology and phonemic language modeling. *Preprint*, arXiv:2504.03036.

Henry Grafé and 1 others. 2025. Graph connectionist temporal classification for phoneme recognition. *arXiv preprint arXiv:2509.05399*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *Representation Learning Workshop, ICML*.

International Phonetic Association. 2015. Ipa chart. 107 segmental letters, 44 diacritics, 4 prosodic marks.

Michael Kenstowicz and Filomena Sandalo. 2016. Pretonic vowel reduction in brazilian portuguese: Harmony and dispersion. *Journal of Portuguese linguistics*, 15(1).

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information fusion*, 109:102422.

Hwa-Yeon Kim, Jong-Hwan Kim, and Jae-Min Kim. 2022. Fast bilingual grapheme-to-phoneme conversion. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 289–296, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

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 modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

Rodrigo Lima, Sidney E Leal, Arnaldo Candido Junior, and Sandra M Aluísio. 2024. A large dataset of spontaneous speech with the accent spoken in são paulo for automatic speech recognition evaluation. In *Brazilian Conference on Intelligent Systems*, pages 33–47. Springer.

Gustavo Mendonça and Sandra M Aluísio. 2014. Using a hybrid approach to build a pronunciation dictionary for brazilian portuguese. In *INTERSPEECH*, pages 1278–1282.

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

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 19–26, Copenhagen, Denmark. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*.

Luann Dias Souza. 2025. Acento e peso silábico no português brasileiro.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Siddharth A Unnithan and Nihar R Mahapatra. 2024. Advancements in grapheme-to-phoneme conversion models for speech synthesis. In *International Conference on Computational Science and Computational Intelligence*, pages 139–150. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008.

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, and 3 others. 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.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. *Preprint*, arXiv:2204.03067.

Markéta Řezáčková, Jan Švec, and Daniel Tihelka. 2021. T5g2p: Using text-to-text transfer transformer for grapheme-to-phoneme conversion. In *Interspeech 2021*, pages 6–10.