Improving automated literacy assessments through a multiple output grapheme-to-phoneme approach

Rômulo C. de Mello[1,2], Patrick C. de Carvalho[1], Gustavo Silva[1], Maria Clara C. Carneiro[1], Rafaela Lopes[1], Jairo F. de Souza[1,3]

[1]LApIC Research Group – Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora – MG – Brazil

[2]Fundação CAEd – Juiz de Fora – MG – Brasil

[3]Departamento de Ciência da Computação – UFJF
Juiz de Fora – MG – Brazil

{romulo.c.mello, patrickcarvalho448, gustavosjn2013,

almeidarafaelalopes,mcosta1807}@gmail.com, jairo.souza@ufjf.br

Abstract. *Fluency assessments are essential for monitoring literacy development, but automatic systems still struggle with the phonetic diversity of Brazilian Portuguese and the specific characteristics of children's reading. We propose a rule-based grapheme-to-phoneme converter that generates multiple acceptable transcriptions per word, accounting for regional variations and child speech. Validated on children's reading data, the module reduces errors, increases accuracy from 89% to 95% on the PARC-2019 dataset, and improves performance among inconsistent readers. Flexible G2Ps make assessments fairer and more reliable.*

Resumo. *Avaliações de fluência são essenciais para monitorar a alfabetização, mas sistemas automáticos ainda enfrentam dificuldades com a diversidade fonética do português brasileiro e características da leitura infantil. Propomos um conversor grafema-fonema baseado em regras que gera múltiplas transcrições por palavra, considerando variações regionais e fala infantil. Validado com leituras infantis, o módulo reduz erros, eleva a acurácia de 89% para 95% na base PARC-2019 e melhora o desempenho entre leitores inconsistentes. G2Ps flexíveis tornam avaliações mais justas e confiáveis.*

1. Introduction

The acquisition of proficient reading represents a significant step in educational and cognitive development, broadening access to knowledge and, more broadly, to full participation in society. In this context, careful and fair assessment of reading skills in early stages is fundamental to support student development and guide effective pedagogical practices. Measuring reading fluency — which involves assessing reading speed and accuracy [Rasinski 2017], presents inherent challenges, especially in contexts of great linguistic diversity and when dealing with children, whose speech characteristics are in full development.

In Brazil, initiatives such as PARC (Reading Fluency Assessment) [Rocha et al. 2024], and in other international contexts such as NAEP (National

DOI: 10.5753/sbie.2025.12313

Assessment of Educational Progress) in the United States [White et al. 2021], exemplify the effort to monitor these skills on a large scale. To optimize the application and analysis of these assessments, automatic systems have been increasingly employed [Sorgatto et al. 2021]. Typically, these systems process audio recordings of students' reading using neural networks to generate probability matrices for the phones [Jucá et al. 2023], the smallest distinctive units of speech that accurately represent the sounds produced by the human vocal apparatus [Moore and Skidmore 2019]. Based on a reference phonetic transcription of the read text, the system can then verify the correctness of the reading. It is here that G2P (grapheme-to-phoneme or grapheme-to-phone) systems play a critical role, by providing the reference phonetic transcriptions from the written text.

The complexity increases when we consider the phonetic variation inherent in any living language. In Brazilian Portuguese, for example, the vast geographical area and rich sociocultural history have resulted in a multiplicity of regional accents and pronunciations. It is through the detailed analysis of phones that it becomes possible to distinguish these variants. A classic example of this variation is the word "teatro" in Portuguese language, which can be pronounced with the phones [te] ([te.ˈa.trʊ]) in some regions or with the phones [ʧi] ([ʧi.ˈa.trʊ]) in others. Both forms are considered correct and accepted, but an assessment system that does not recognize this diversity may incur errors in judgment.

Additionally, in the context of childhood education, a particular challenge is that variation in pronunciation is also intrinsically associated with the speaker's literacy level. Children in the process of acquiring reading are developing their phonological awareness and learning the complex mapping between graphemes and phones; as such, they may not decode words into sounds with the same consistency or "standard" expected of a proficient adult reader [Lemle 2007]. Their pronunciations, even if variable or divergent from the adult standard, can be perfectly acceptable and indicative of their developmental stage, and should not be penalized as errors. This reality poses a considerable challenge to automatic correction systems, which need to be flexible enough to recognize and validate a broad and dynamic spectrum of speech variations. The objective of this article is to present a robust technological solution that improves one key element of automatic reading fluency assessments to increase their equity and precision, especially for Brazilian children in the early years of schooling, by systematically accommodating both regional phonetic variations and those inherent to literacy development.

To achieve this objective, this work presents two main and interconnected contributions. Firstly, we propose an adaptive rule-based grapheme-to-phoneme conversion framework, explicitly designed to generate multiple phonetic transcriptions for a single orthographic input. This ability to generate multiple outputs is crucial because, unlike traditional approaches that produce a single "standard" transcription, our framework is designed to capture and represent legitimate phonetic diversity. Secondly, we introduce a detailed set of linguistic rules specifically designed for this framework, focused on the particularities of Brazilian Portuguese and, crucially, on the speech characteristics of children in the early years of elementary school. These rules were designed by experts in the domain of linguistics based on their knowledge of which phones should generally be produced as a result of reading a certain grapheme or set of graphemes.

The choice for a rule-based approach is based on its transparency, interpretability, and, fundamentally, its ease of extension and adaptation to new linguistic variants or pronunciation patterns that may be identified, an essential characteristic for a system that aims to remain current with the dynamics of the language and pedagogical needs. The ability to generate and consider multiple correct phonetic representations thus becomes indispensable in countries with vast linguistic diversity such as Brazil, with the ultimate aim of producing fairer, more accurate, and pedagogically useful assessment results.

In summary, this paper is organized as follows: Section 2 revisits the literature on grapheme-to-phoneme converters, highlighting advancements and gaps that motivate our proposal; Section 3 details the materials used and the method adopted to extend the G2P system, including the multi-output algorithm and the set of linguistic rules; Section 4 presents the experiments conducted and the results obtained on the PARC2019-Manual benchmark; finally, Section 5 discusses the pedagogical and scientific impact of the findings, and Section 6 concludes with perspectives for future work.

2. Related works

The existing methods for G2P conversion can be classified in three major groups: methods based on rules and dictionaries, statistical methods and deep learning methods [Cheng et al. 2024]. The rule and dictionary-based methods are the most traditional approach, relying on linguistic rules or dictionary bases to map graphemes to their corresponding phones. Statistical methods use machine learning and probabilistic models while deep learning methods are based on deep neural networks.

Several studies have explored these methodologies in more detail. For instance, Kubo et al. [2013] introduced the application of WFST (weighted finite state transducers) to the problem of G2P conversion and Taylor [2005] used Hidden Markov Models to accomplish this task. Both are widely used statistical models. As an example of a deep learning based method, there is the G2P developed by Rao et al. [2015], which utilizes a LSTM neural network. Despite achieving superior results [Vecchietti 2017], they are more resource intensive as they require more hardware resources and large amount of data and in order to train and run the model [Cheng et al. 2024], which is undesirable when dealing with low-resource languages. In addition, regarding deep learning methods, their lack of interpretability is a well-known issue [Zhang et al. 2021].

In the context of Brazilian Portuguese, a rule-based G2P system was introduced by Siravenha et al. [2008], with the aim of providing an efficient alternative to dictionary-based models. A more recent example of a similar approach, applied to the Polish language, was done by Kłosowski [2022]. The main advantage of the rule-based approach lies in its adaptability, which allows it to handle new words without requiring constant updates. Furthermore, this approach demands less memory, as it does not rely on storing an extensive database. In contrast, dictionary-based models require a predefined list of words that must be periodically updated to maintain lexical coverage, resulting in higher computational costs and storage requirements. The authors implemented G2P system for the São Paulo variant of Brazilian Portuguese, however, this approach is not easily extensible to other variants due to hardcoded rules.

Addressing the aforementioned limitation, Epitran [Mortensen et al. 2018] is a rule-based G2P framework that supports multiple languages and supports outputs in both IPA (International Phonetic Alphabet) and X-SAMPA (Extended Speech Assessment Methods Phonetic Alphabet), widely used standards of phonetic representation. Its aim is to provide coverage for low resource languages, that is, languages whose available data is insufficient for training a statistical or deep learning model. Their approach is flexible, allowing for new language rules to be easily included.

In the same vein, the G_i2P_i [Pine et al. 2022] also provides an extensible multilingual G2P framework, which includes already implemented rules for several indigenous languages. Furthermore, it prioritizes easy of use and facilitates contributions from individuals without expertise in programming by providing a graphical web-based interface, RESTful API and extensive documentation.

However, while transparent, adaptable and usable for professionals without knowledge in programming, the solutions presented by Mortensen et al. [2018] and Pine et al. [2022] are less suitable for the task of producing transcriptions for the assessment of fluency in children. That is because the production of the sounds might not always obey the standard rules of pronunciation and still be acceptable for one who is not yet fully literate.

3. Materials and Method

We generally followed the following steps: (i) reproduction of the Falabrasil G2P rules; (ii) creation of a gold-standard dataset with different pronunciation variants; (iii) extension of the original G2P rules to accomodate these variants and (iv) iterative refinement of the rules. This is explained in more detail in the remaining paragraphs of this section.

We assume rule-based G2P converters are particularly well suited to educational contexts: they offer transparency, low computational cost, and allow for the controlled incorporation of linguistic knowledge. As a starting point, we reproduced the converter developed by the Falabrasil group [Siravenha et al. 2008], one of the first self-contained implementations for Brazilian Portuguese. The original system employs twenty-nine regular expressions to identify the stressed vowel and 140 conversion rules, applied from the most specific to the most general, to generate a single phonetic transcription in X-SAMPA per word.

The replication was validated through regression testing on a broad vocabulary; each output matched exactly with Falabrasil, confirming the fidelity of the implementation. However, the very architecture of a single-output system revealed a drawback: regional variants, fully correct from a phonetic standpoint, were being classified as errors by automatic reading assessment systems, penalizing, for instance, typical contrasts between the dialects of the South and Northeast Brazilian regions. This limitation motivated the second phase of the project, aimed at making the converter sensitive to the multiple legitimate pronunciations found in Brazil.

To guide this expansion, linguists created a gold-standard dataset. Based on real recordings of children's reading, they cataloged the most frequent word pronunciations variants identified in states of Northeast, Southeast and South regions. From this gold-standard, we conducted an incremental development cycle:

1. Identify coverage gaps, variants present in the dataset but absent in the converter.
2. Create or adjust rules to fill each gap, documenting the phonological rationale.
3. Run automated tests at two levels: (a) *regression*, ensuring the original outputs

remain intact; (b) *coverage*, verifying that 100% of cataloged variants are represented.

Technically, all rules were reorganized into a `JSON` file, where each entry provides a detailed specification of the rule's components. These includes the rule's identifier (identified by *idx*), the graphemic group, the regular expression (described by *REGEX*), the context relative to the stressed vowel, the pointer advance in the word, a range of characters ahead to be considered, the identifier of another rule that the curent one depends on, and a list (possibly multiple) of phonetic outputs. Figure 1 illustrates a rule that determines the conditions for converting an input's graphemes <en> into the phone [ẽ], which happens in the word *quente* (hot, in Portuguese), for example. In this case, the rule 39 determines the conversion of <e> when it is followed by <m, n>, except if <m, n> are followed by <h> or any vowel. Additionally, in this context, the rule requires analyzing a three-character range within the word: <ent>.

```
{
    "idx": 39,
    "group": "e",
    "REGEX": "([mn][^haeiouáéíóúãâõê#])",
    "output": [
        "e~"
    ],
    "context": null,
    "depends_on": 21.0,
    "advance": 2,
    "range": "3"
},
```

Figure 1. Conversion rule of an input grapheme <e> into the phone [ẽ].

During conversion, the algorithm loads this set, locates the stressed syllable using the original twenty-nine rules, and scans the word by consuming graphemes; whenever a rule matches, each output is propagated, constructing a tree of all possible combinations. In the end, the system returns the unique set of generated phonetic transcriptions, allowing recognition or alignment modules to select the variant that best matches the speech signal.

The incorporation of new rules created from the gold-standard dataset was accompanied by automated tests ensuring two criteria: (i) the original Falabrasil transcriptions remain intact (regression) and (ii) all cataloged variants are covered by the expanded rule set. Successive iterations enabled the refinement of exceptions, adjustments to stressed vowel contexts, and optimization of pointer advancement to maintain efficiency.

Algorithm 1 performs a scan from left to right of the input word. At each position, it consults the rule group matching the current grapheme, checks the regular expression and tonal context, and—when a rule fires—concatenates every phone sequence in `output` onto every active path. This branching step produces a lattice of partial pronunciations that evolves as the pointer advances the number of characters specified by `advance`. If no rule applies, the raw grapheme is appended unchanged. Upon reaching the end of the word, the set $\mathcal{T}$ contains all admissible phonetic renderings, ensuring that the downstream aligner has enough alternatives to accommodate regional and developmental pronunciation differences.

Algorithm 1 Multi-output rule-based G2P

```
paths ← {ϵ}                                              ▷ partial phone sequences
i ← 0,  s ← STRESSINDEX(word)
while i < |word| do
    c ← word[i],  matched ← false
    for all rule ∈ group(c)                              ▷ priority order do
        if REGEXMATCH(rule,word[i:]) and CONTEXTOK(rule,s,i) then
            newPaths ← ∅
            for all seq ∈ paths do
                for all out ∈ rule.outputs do
                    newPaths ← newPaths ∪ {seq + out}
                end for
            end for
            paths ← newPaths
            i ← i+ rule.advance;  matched ← true;  break
        end if
    end for
    if not matched then
        paths ← {seq + c : seq ∈ paths}
        i ← i + 1
    end if
end while
return {CONCAT(seq) : seq ∈ paths}
```

For example, for the word *quente*, the conversion rules work as follows: the grapheme <q> is mapped to the phone [k]; <u> after <q> is ignored; <t> is mapped to [t] and [ʧ]; and <e> is mapped to [e], [i], [ɪ] and [ɛ]. That way, the algorithm generates every possible combination of the produced phones (respecting the order of the original graphemes). That process is illustrated in Figure 2.

4. Experiment and Results

The impact of the multi–output converter was tested on a collection of 1560 fluency-assessment recordings produced by Brazilian second-grade pupils from public schools and exhaustively annotated by expert raters (herein called PARC2019-Manual). Each audio passage was scored twice by the automatic pipeline: once with the single-output Falabrasil G2P (S1) and once with the proposed multi-output G2P (S2). All other components—acoustic model, forced alignment and profile thresholds—remained unchanged, ensuring that any difference stems exclusively from the transcription stage.

Table 1 reports four global metrics. Mean error (ME) captures the signed difference between automatic and human word counts, so negative values indicate systematic under-scoring. Mean absolute error (MAE) and its normalised version (NMAE) discard direction, while overall accuracy gauges the proportion of correctly assigned reader profiles.

The multi-output system halves the mean bias and reduces absolute error by more than one full word per passage, bringing the normalised MAE down from 20.7 % to

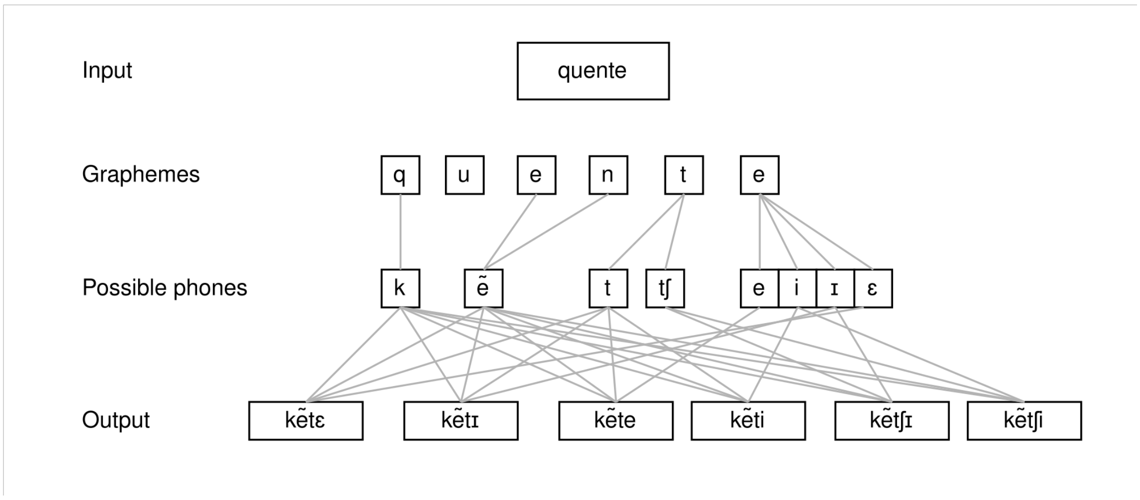

Figure 2. Illustrates the combinatorial nature of the G2P process for the word "quente".

Table 1. Global performance on PARC2019-Manual.

G2P	ME	MAE	NMAE	Accuracy
Falabrasil (S1)	-3.31	3.48	0.207	0.892
Multi-output (S2)	-1.76	2.18	0.129	0.948

12.9 %. Overall profile accuracy rises from 89.2 % to 94.8 %, an absolute gain of 5.6 percentage points.

A more granular view appears in Table 2, which breaks precision and recall by reader profile. Improvements concentrate on the *pre-reader* group—the cohort that exhibits the widest phonetic variability—while the already high scores for *beginner* readers are preserved.

Table 2. Precision and recall by profile.

	Precision		Recall	
G2P	Pre-reader	Beginner	Pre-reader	Beginner
Falabrasil (S1)	0.650	0.997	0.990	0.868
Multi-output (S2)	0.804	0.995	0.981	0.941

The multi-output converter not only increases overall accuracy, but also corrects a long-standing bias in automatic scores. The mean error is roughly halved, and the absolute deviation per pass drops by more than one word, showing that the improvement is systematic and not confined to a few outliers. Analysis of precision and recall identifies where the gains are most significant: many children who were previously below the pre-reader threshold for regional variants or developmental pronunciations are now recognized correctly, while beginning readers maintain their high accuracy and increase recall. A closer look at the tokens that change status reveals a commonality: most involve predictable alternations of vowel quality, open vs. closed sounds. Because these patterns recur across the word list, a small set of variant-aware rules produces a large reduction in mismatches. All of the gains were achieved without tuning the acoustic models, indicating that expanding the phonetic search space is a low-cost but effective path to fairer fluency scores across the spectrum of young readers.

To show in concrete terms how the multi-output rules eliminate unfair penalties, we scrutinised three recurrent items from the word list—*esporte*, *creche* and *menino* (sport, daycare and boy, in Portuguese)—and traced how each behaves before and after the rule expansion. For *esporte*, the baseline Falabrasil transcription recognised only the production with a alveolar fricative consonant and a close-mid vowel at the beginning [esˈ.pɔx.ʧi], whereas many speakers from Rio de Janeiro produced it with a postalveolar fricative and a front vowel [ijʃ.ˈpɔx.ʧi]. Once this variant was encoded as an alternative, the recogniser stopped penalising children whose pronunciation aligns with their local norm, improving score precision without affecting other dialects.

The second word, *creche*, highlights a source of variation associated with the literacy acquisition process. Fluent readers typically reduce the vowel in the final syllable, so it is not pronounced fully, as it would be if it were in the stressed syllable. In this way, the vowels [o] and [e] become [ʊ] and [ɪ], respectively, in post-tonic position—reduced vowels that resemble the pronunciation of [u] and [i]. This occurs because the emphasis in speech falls on the stressed syllable, while post-tonic syllables receive less articulatory energy. However, children in the process of learning to read may pronounce these vowels fully, even when they would normally be reduced. Thus, in a system that allows only a single acceptable output, the word *creche* would be represented solely as [ˈkrɛʃi], whereas a child in the early stages of literacy might produce [ˈkrɛʃɛ] or [ˈkrɛʃe]. Including these multiple possible outputs in the converter prevents novice readers from being unfairly penalized for a pronunciation that merely reflects a normal stage of learning.

The third word, *menino*, highlights a phonological case, rather than purely dialectal. Even fluent readers often apply vowel harmony to unstressed syllables, turning [meˈninʊ] into [miˈninʊ]. This is because of the tendency, in speech, for unstressed vowels to accompany the stressed vowel. Since the vowel <i> is a high vowel, the vowel <e> tends to rise to accompany it, also transforming into <i>. In the single-output regime every occurrence of [mi-] counted as a miscue, even though this change is a common phonological occurrence noticed in Brazilian Portuguese. By licensing the harmonised vowel as an acceptable output, the multi-output converter prevents students from being unfairly penalised for a pronunciation that reflects normal learning processes.

Together, these three cases demonstrate how missing phones can propagate into systematic under-scoring for entire regions or literacy stages, and how the enriched rule set rectifies that bias by acknowledging the full range of pronunciations children naturally produce.

The data supports the initial hypothesis: providing the aligner with a broader set of phonetic transcriptions reduces the penalization of legitimate regional variation and thus improves fluency classification. Although the present study focuses only on pre-reader and beginner profiles, the same principle may be extended to more advanced profiles, where prosodic and segmental subtleties remain relevant. In summary, the introduction of multiple phonetic outputs proves to be a low-cost, high-impact enhancement for automatic reading assessment systems in Brazilian Portuguese.

5. Discussion

The results reveal substantial progress in automated reading fluency assessments within the Brazilian context. First, the 40% reduction in mean absolute error demonstrates that expanding the phonetic search space—through rules aware of dialectal variation and literacy stages—is an effective way to align automatic scoring with human judgment. This improvement, achieved without retraining acoustic models, indicates that significant gains can arise from "lightweight" linguistic layers, which are easily transferable to other educational ASR systems.

Another key point is the impact on regional equity. The inclusion of frequent phonetic variants, rather than limiting the system to a single canonical pronunciation, eliminated the bias that previously penalized children from certain regions. This change is especially evident in the improvement of classification accuracy within the "pre-reader" profile, a group that often exhibits early and highly variable speech patterns. By embracing these legitimate variations, the system avoids misclassifying correct regional pronunciations as errors. This evidence reinforces the importance of incorporating national phonetic diversity when designing standardized assessment metrics, ensuring that such tools are fair and inclusive across different linguistic backgrounds.

From a pedagogical perspective, the increased reliability of scores enables more targeted classroom decisions. Teachers can tailor interventions based on data that now respect legitimate pronunciation variations, avoiding "false alarms" that led to misdiagnoses. Furthermore, more accurate reports provide valuable feedback for government programs that monitor literacy policies.

Finally, the strategy proposed in this work offers a replicable roadmap: (i) corpus collection with real-world variation, (ii) rule elicitation in collaboration with phonology experts, (iii) automatic regression testing, and (iv) short validation cycles. This workflow can be extended to other age groups or adapted to languages facing similar challenges, for example, Latin American Spanish, where accents and rapid linguistic changes also hinder the use of standardized models that fail to accommodate linguistic diversity.

In sum, the multi-output G2P module not only improves numerical metrics, but also enhances the social utility of reading assessments by embracing the country's linguistic richness and supporting fairer and more effective pedagogical practices.

6. Concluding remarks

The research presented here demonstrates that a rule-based, custom-designed grapheme-to-phoneme converter, enhanced to produce multiple legitimate pronunciations, can transform the accuracy and fairness of automated assessments of reading fluency in Brazilian Portuguese. By basing each new rule on a gold-standard corpus of children's speech and validating each iteration through systematic regression testing, we preserve the stability of the original Falabrasil system while opening up the phonetic search space to dialectal and developmental variations. The quantitative improvements include: systematic bias in word counts was halved, mean absolute error decreased by more than one word per passage, and profile classification accuracy increased from 89% to 95%.

Equally important is the qualitative impact. Teachers and program administrators can use this approach to receive scores that better reflect children's genuine reading abilities, rather than artifacts of regional or developmental pronunciation. The declarative architecture of the converter ensures transparency and auditability, allowing educators and linguists to inspect, debate, and refine each decision encoded in the rule base. In this way, the work bridges a persistent gap between large-scale educational technology and the nuanced linguistic realities that shape early literacy in a country as diverse as Brazil.

Several lines of inquiry can build on these results to broaden the linguistic reach and pedagogical value. First, the inventory of rules can be expanded to encompass underrepresented accents, including Amazonian, Central-Western, and immigrant varieties, so that implementations across the country truly respect the full spectrum of Brazilian Portuguese. It can also be expanded and tested for other age groups and literacy stages, as well as children in bilingual contexts or learners of Portuguese as a heritage language. There is also room to improve automatic assessment tools based on prosodic information, adding rules that model stress timing, vowel reduction, and intonational contours, features that become critical as readers move beyond the pre-reading stage. From an engineering perspective, we envision an adaptive lexicon layer that continuously harvests new pronunciation variants from classroom recordings, updating the rule base in a controlled and versioned manner while preserving full traceability. Integrating the enhanced G2P output into interactive dashboards for teachers is another interesting line for future developments, allowing educators to visualize which words or phonological patterns trigger errors and adapt interventions accordingly. Finally, by making the rule engine, test suite, and assessment scripts available as open source software, we hope to mobilize a broader community of researchers and practitioners who can replicate the development cycle for other low-resource languages, such as Latin American variants of Spanish or other languages, and thus extend the social benefits of multi-output, language-based G2P systems far beyond the Brazilian context.

References

[Cheng et al. 2024] Cheng, S., Zhu, P., Liu, J., and Wang, Z. (2024). A Survey of Grapheme-to-Phoneme Conversion Methods. *Applied Sciences*, 14(24):11790.

[Jucá et al. 2023] Jucá, B. D. O., Rocha, C. C., Mello, R. C. D., Barrére, E., and Souza, J. F. D. (2023). Identificação de silabação em áudios de leitura de crianças em anos iniciais. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação (SBIE 2023)*, pages 1363–1374, Brasil. Sociedade Brasileira de Computação - SBC.

[Kubo et al. 2013] Kubo, Y., Hori, T., and Nakamura, A. (2013). A method for structure estimation of weighted finite-state transducers and its application to grapheme-to-phoneme conversion. In *Interspeech 2013*, pages 647–651. ISCA.

[Kłosowski 2022] Kłosowski, P. (2022). A Rule-Based Grapheme-to-Phoneme Conversion System. *Applied Sciences*, 12(5):2758.

[Lemle 2007] Lemle, M. (2007). *Guia teórico do alfabetizador*, volume 104 of *Princípios*. Editora Ática, São Paulo, 17 edition.

[Moore and Skidmore 2019] Moore, R. K. and Skidmore, L. (2019). On the Use/Misuse of the Term 'Phoneme'. pages 2340–2344.

[Mortensen et al. 2018] Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[Pine et al. 2022] Pine, A., William Littell, P., Joanis, E., Huggins-Daines, D., Cox, C., Davis, F., Antonio Santos, E., Srikanth, S., Torkornoo, D., and Yu, S. (2022).

Gi22Pi Rule-based, index-preserving grapheme-to-phoneme transformations Rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

[Rao et al. 2015] Rao, K., Peng, F., Sak, H., and Beaufays, F. (2015). Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. ISSN: 2379-190X.

[Rasinski 2017] Rasinski, T. (2017). Fluency matters. *International Electronic Journal of Elementary Education*, 7(1):3–12. Section: Articles.

[Rocha et al. 2024] Rocha, C. C., Mello, R. C. D., and Souza, J. F. D. (2024). Avaliação de fluência leitora em língua portuguesa: primeira experiência com uso em larga escala de Inteligência Artificial. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, pages 3075–3084, Brasil. Sociedade Brasileira de Computação - SBC.

[Siravenha et al. 2008] Siravenha, A., Neto, N., Macedo, V., and Klautau, A. (2008). Uso de Regras Fonológicas com Determinação de Vogal Tônica para Conversão Grafema-Fone em Português Brasileiro. *7th International Information and Telecommunication Technologies Symposium.*

[Sorgatto et al. 2021] Sorgatto, D. W., Nogueira, B. M., Cáceres, E. N., and Mongelli, H. (2021). Avaliação de classificadores para relacionar características escolares a indicadores educacionais. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE 2021)*, pages 1232–1242, Brasil. Sociedade Brasileira de Computação - SBC.

[Taylor 2005] Taylor, P. (2005). Hidden Markov models for grapheme to phoneme conversion. In *Interspeech 2005*, pages 1973–1976. ISCA.

[Vecchietti 2017] Vecchietti, L. F. S. (2017). Comparison between rule-based and data-driven natural language processing algorithms for Brazilian Portuguese speech synthesis.

[White et al. 2021] White, S., Sabatini, J., Park, B.-s. J., Chen, J., Bernstein, J., and Li, M. (2021). The 2018 NAEP Oral Reading Fluency Study. Technical Report NCES 2021-025, U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Washington, DC.

[Zhang et al. 2021] Zhang, Y., Tino, P., Leonardis, A., and Tang, K. (2021). A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742.