

TINA: A Hybrid Model of Brazilian Portuguese Pronunciation Structure

Alexandre Menezes Barroso (<https://orcid.org/0009-0005-2794-767X>)¹

Preprint version, April 2026. Submitted manuscript for review.

Abstract

Automated phonetic transcription for Brazilian Portuguese is not well modeled as a flat grapheme-to-phoneme task, because usable outputs must encode syllabification, lexical stress, and dialect-sensitive realizations. This paper presents TINA, a compact hybrid system that predicts a structured, archiphonemic representation with neural components for syllables, phoneme content, and stress, and then expands that representation into surface-oriented variants through explicit phonological rules. TINA is evaluated against native-speaker pseudoword behavior, with qualitative output inspection on familiar lexical material confirming that the system produces well-formed structured transcriptions. On pseudowords, it continues to produce well-formed outputs, captures the major opposition between final and penultimate stress, and responds strongly to final-syllable structure, showing that it generalizes beyond memorized lexical mappings. Yet it also overregularizes speaker behavior: it suppresses antepenultimate stress and treats some right-edge cues too categorically. TINA is therefore an interpretable model of BP pronunciation structure rather than a state-of-the-art neural G2P architecture.

Keywords: phonetic transcription; grapheme-to-phoneme conversion; lexical stress; syllable structure; neural-symbolic phonology

¹PhD student at State University of Campinas (UNICAMP) under supervision of Dr. Maria Filomena Spatti Sandalo (<https://orcid.org/0000-0003-4595-7765>). This work forms part of my PhD research, Contact: abaroso@outlook.com.br. **How to cite this preprint:** Barroso, A. M. (2026). TINA: A Hybrid Model of Brazilian Portuguese Pronunciation Structure. Preprint. Retrieved from <https://lingbuzz.net/lingbuzz/009910>

1 Introduction

Automatic phonetic transcription for Brazilian Portuguese is not well characterized as a simple letter-to-sound conversion problem. A usable transcription must recover not only segmental material, but also syllable boundaries, lexical stress, and a range of position-sensitive realizations whose surface form may differ across dialects. In mainstream grapheme-to-phoneme work, the task has often been framed as sequence transduction from graphemes to phonemes, and recurrent neural work has also argued for treating pronunciation prediction as a word-level sequence problem rather than only as local grapheme–phoneme conversion, which is an important baseline point of departure for the present paper (Rao et al., 2015; Yao & Zweig, 2015). For that reason, a system that outputs only a flat phoneme string risks hiding exactly the parts of pronunciation that matter most for Brazilian Portuguese. What is needed is not simply a mapping from orthography to phones, but a model of pronunciation structure.

This paper presents TINA: *Transcrição Inteligente, Notação Automática*², a compact hybrid system designed around that premise. On the neural side, the model reads orthographic input and predicts a structured representation in which syllables, phoneme content, and stress are treated as related outputs rather than as independent post-processing steps, in the same broad spirit as joint pronunciation models that predict phoneme sequence, syllabification, and stress together (van Esch et al., 2016). On the symbolic side, a rule module expands the neural output into surface-oriented variants, allowing the system to preserve a relatively abstract representation when early commitment to a single realized form would be linguistically misleading. The resulting architecture is deliberately modest. Its goal is not to claim a radically new neural method, but to make Brazilian Portuguese transcription more linguistically legible.

This framing matters especially for Brazilian Portuguese because several of the language’s central regularities are organized above the level of isolated segments. Stress interacts with syllable structure, and analyses of Brazilian Portuguese have long argued that in non-verbal domains stress is sensitive to syllable weight and especially to the right-edge contrast between open and heavy final

²transl. “Intelligent Transcription, Automatic Notation”.

syllables (Wetzels, 2007); codas are sites of repair, weakening, and variability; and some segmental contrasts are better handled as underspecified classes whose realized form depends on context and dialect. A transcription system for Brazilian Portuguese therefore benefits from making these layers visible rather than flattening them away. In TINA, this leads to a division of labor in which the neural model is responsible for broad structural analysis and the rule layer is responsible for controlled surface expansion.

The paper also argues that evaluating such a system only on held-out lexical items is not enough. Lexical benchmarks can show whether the model is a competent transcriber on familiar-looking material, and that kind of exact lexical comparison remains a central evaluation tradition in grapheme-to-phoneme work, including recurrent approaches that emphasize broader contextual modeling and do not require explicit grapheme–phoneme alignment for training (Rao et al., 2015; Yao & Zweig, 2015). But they do not fully reveal what kind of generalization the model has learned. For that, pseudowords are especially informative. Recent experimental evidence for Brazilian Portuguese shows that pseudoword stress assignment is shaped by phonological similarity to real words, lexical frequency, and final-syllable structure, which makes novel forms a particularly revealing test bed for the kind of structured generalization at issue here (**Benenevides2022**). They force the system to assign stress, syllabify, and handle segmental structure in the absence of direct lexical memory. In other words, they ask whether the model behaves like a structured pronunciation system when it encounters forms it has never seen before.

That question becomes the empirical center of the paper. The central evaluation is a pseudoword benchmark aligned with native-speaker behavior, used to test whether the model extends Brazilian Portuguese regularities to novel forms in a linguistically meaningful way. Qualitative inspection of outputs on familiar lexical material confirms beforehand that the system produces well-formed structured transcriptions. This second evaluation turns out to be the more revealing one. It shows not only where the system succeeds, but also where it becomes too categorical: the model captures broad right-edge stress regularities and produces stable syllabified outputs on unfamiliar words, yet it overregularizes precisely where speaker behavior remains gradient, analogical, or

probabilistic (Benevides, 2024; Wetzels, 2007).

The contribution of the paper is therefore twofold, but modest in scope. First, it presents a hybrid transcription architecture in which syllabification, stress, abstract segmental structure, and dialect-sensitive expansion are treated as parts of the same pronunciation problem, rather than as separate downstream annotations (van Esch et al., 2016). Second, it uses pseudowords as a principled test of whether such a model encodes speaker-like generalizations rather than merely reproducing stored lexical mappings. Read this way, TINA is most interesting not as a state-of-the-art neural architecture, but as a compact and interpretable model of Brazilian Portuguese pronunciation structure.

The remainder of the paper is organized as follows. Section 2 positions the system within work on neural grapheme-to-pronunciation modeling, Brazilian Portuguese phonology–phonetics, and pseudowords as a benchmark for productive generalization. Section 3 describes the data representation, neural architecture, training objectives, and rule-based expansion layer. Section 4 defines the lexical and pseudoword evaluation tracks. Section 5 presents the results and their analysis. Section 6 discusses what kind of pronunciation knowledge the model appears to encode and where its limits become clear. Section 7 concludes the paper by summarizing its contribution and limits.

2 Background and positioning

This section situates TINA as a structured grapheme-to-pronunciation system for Brazilian Portuguese. The system is not designed as a purely segmental mapper from orthography to a flat phoneme string. Instead, it treats pronunciation as a layered object involving syllabification, lexical stress, and position-sensitive realizations that may differ across dialect profiles. That design choice is visible throughout the whole architectural design: in the organization of the training targets, in the split between syllable-, phoneme-, and stress-level prediction, and in the final rule-based expansion that derives surface-oriented variants from a more abstract neural output.

2.1 Neural G2P and structured pronunciation modeling

A conventional grapheme-to-pronunciation (G2P) system can be framed as a sequence transduction problem in which a word is converted into a single output string, and neural G2P has explicitly been treated in sequence-to-sequence terms through both encoder–decoder and alignment-aware recurrent models (Yao & Zweig, 2015). Recurrent LSTM-based G2P models also helped establish that pronunciation can be modeled with broader contextual conditioning at the word level without requiring explicit grapheme–phoneme alignment as a prerequisite for training (Rao et al., 2015). In encoder–decoder settings, attention mechanisms make it possible to condition each prediction step on the most relevant parts of the encoded input rather than relying on a single fixed representation (Luong et al., 2015). In G2P specifically, attention-based models also provide a way to avoid hand-built alignment machinery and instead learn largely monotonic grapheme-to-phoneme correspondences during decoding (Toshniwal & Livescu, 2016). Self-attentional architectures then make it possible to model broader dependencies within the encoded sequence through multi-head attention (Vaswani et al., 2017). Transformer-based designs have also been explored for G2P specifically, including direct comparisons against recurrent and convolutional alternatives (Yolchuyeva et al., 2019). Work on structured pronunciation modeling, however, makes clear that phoneme sequence alone is not always enough: syllabification and stress may need to be predicted as part of the pronunciation itself (van Esch et al., 2016). For Brazilian Portuguese, that broader view is especially important because a flat phoneme string hides distinctions that are central to the organization of pronunciation. A transcription is not only a sequence of segments. It also encodes how those segments group into syllables, where lexical prominence falls, and which portions of the output are better treated as abstract or underspecified before more specific phonetic realizations are derived. TINA is positioned around that broader view of pronunciation structure.

In practical terms, the system decomposes transcription into related prediction tasks rather than collapsing everything into one undifferentiated decoder. A shared encoder reads the input character sequence and produces contextualized representations. From there, the model predicts a syllable sequence, generates the phoneme content of each syllable, and assigns stress over the

resulting syllabic structure. This organization matters conceptually. It means that the model does not treat syllabification and stress as annotations added after the fact, but as first-class parts of the prediction problem, in the same general spirit as prior joint models that predict phoneme sequence, syllabification, and stress together (van Esch et al., 2016). It also aligns with work that treats syllable structure and boundary information as useful linguistic scaffolding for G2P rather than as incidental annotation (Lo & Nicolai, 2021). Not only that, it aligns with hierarchical approaches that argue for preserving relations among larger and smaller linguistic units instead of flattening them into a single sequence representation (Ronanki et al., 2017). More broadly, it follows the multitask idea that related outputs can be learned in parallel through a shared representation, with the auxiliary tasks contributing useful inductive bias rather than merely duplicating the main supervision (Caruana, 1997).

This structured design also affects the flow of information inside the model. Syllables are predicted before phoneme strings are completed, and stress is predicted over the syllable sequence while also using phoneme-derived information, a design that resonates with hierarchical prosody models in which phoneme, syllable, and word context interact and in which removing phoneme context hurts stress detection (Lin et al., 2020). In other words, the system explicitly assumes that pronunciation is internally articulated: syllable structure helps organize segmental prediction, and segmental content in turn helps constrain prominence assignment. In that sense, it is not simply a neural network black box, but a phonologically-driven estimator. The training procedure reinforces this interpretation by supervising not only token prediction but also syllable counts and boundary behavior, again in a way that is congenial to joint treatments of prominence and boundary-related structure (Lin et al., 2020). It also uses a decaying teacher-forcing regime in sequence prediction, gradually reducing reliance on gold previous outputs over training epochs (Bengio et al., 2015). As a result, linguistic structure is present both in the outputs and in the learning objective, in line with a multitask training philosophy in which related supervisory signals are expected to improve generalization together (Caruana, 1997).

This is the main sense in which TINA should be understood as a structured pronunciation

model rather than simply as a neural G2P baseline with a larger architecture. The design combines recurrent processing, Luong-style attention (Luong et al., 2015), and a Transformer-style contextual layer with multi-head self-attention (Vaswani et al., 2017; Yolchuyeva et al., 2019), but the main point is not architectural the novelty by itself. The more important point is that these components are assembled in order to predict pronunciation at several linguistically levels that are both meaningful and interpretable. The system therefore occupies a middle ground: it remains a neural transduction model, but one whose internal organization is shaped by syllables, stress, and controlled abstraction in the output representation rather than by phoneme-sequence prediction alone (van Esch et al., 2016).

2.2 Brazilian Portuguese phonology–phonetics targets

The linguistic motivation for TINA comes from the fact that Brazilian Portuguese pronunciation cannot be reduced to segment identity alone. At minimum, an automatic transcriber must account for lexical stress, syllable structure, coda behavior, and position-sensitive consonantal realizations. For BP, stress is not free: analyses of primary stress argue that non-verbal patterns are sensitive to syllable weight, especially the distinction between light and heavy syllables in final and prefinal position, while verbal stress is more tightly tied to morphological structure (Wetzels, 2007). In parallel, work on BP codas shows that this domain is not phonologically inert: coda behavior involves reduction, vocalization, and gradient coordination at the phonology–phonetics interface rather than a single uniformly stable segmental realization (Medeiros, 2012). At the same time, phonetic work on BP lexical stress shows that stress is cued not only by durational prominence and vowel-quality information reflected in formant patterning, but also by robust acoustic correlates such as F0 variability and spectral emphasis across speaking styles, with duration standing out as especially strong (Barbosa et al., 2013; Fontes, 2013). These phenomena are visible already in the format of the training lexicon, where transcriptions mark syllable boundaries with periods and stressed syllables with the stress mark. The target representation is therefore phonologically organized from the start. This matters not only descriptively but also for stress generalization in novel

forms, since recent experimental evidence on Brazilian Portuguese pseudoword reading suggests that segmental, syllabic, and stress structure interact in the assignment of lexical prominence (Benevides, 2024).

The computational design also makes clear that the intended output is not always a fully specified surface form. The transcriptions contain archiphonemic-like symbols such as R, S, T, D, H, and L, which stand for classes of realizations whose exact output depends on phonological context and dialect profile. This is especially important for a language like Brazilian Portuguese, where rhotics in particular sit near the phonetics–phonology boundary and show systematic positional and prosodic variation, while sibilants and coronal stops also display regular context- and dialect-sensitive alternations (Silva & Albano, 1999). By allowing the neural model to predict an abstract representation before its many dialectal variations when applicable, the system avoids forcing a single categorical realization too early in the pipeline. That abstraction, though, is not left uninterpreted. After neural inference, TINA applies a rule-based expansion layer that converts the abstract output into one or more possible variants. The rule module handles several kinds of phenomena. First, it inserts epenthetic *i* in contexts involving certain consonant codas and consonant-plus-*s* clusters. Second, it expands ambiguous symbols according to dialect-specific mappings. This kind of post-neural handling is especially plausible for BP coda-related behavior, where the surface outcome is often better understood as the result of position-sensitive coordination and reduction than as the direct output of a single categorical segment (Medeiros, 2012). In the implemented profiles, this includes different realizations for rhotics, variable behavior for S, and the contrast between palatalized and non-palatalized outputs for T and D. Third, the expansion logic treats word-final and cluster-based rhotic behavior in a special way, including cases in which a final rhotic may be weakened or absent in the generated surface variant.

This hybrid organization is important for the paper’s linguistic positioning. It acknowledges that some aspects of Brazilian Portuguese pronunciation are better modeled as broad structural predictions, while others are better derived as controlled variant expansion. The neural system is thus responsible for learning a compact phonological representation of the word, including syllable

organization and stress. The rule layer then maps that representation to dialect-sensitive surface possibilities. Such a division of labor is particularly appropriate for a language in which pronunciation regularities often interact with syllable position, repair strategies, and regional realization patterns. At the same time, the architecture design does not claim to model the full gradient richness of phonetic realization. Its outputs are categorical strings, even when those strings are deliberately underspecified. The resulting system should therefore be read as a phonology-oriented transcription model with promising phonetic reach, not as a full model of continuous speech realization. That modesty is important for the rest of the paper: the interesting question is not whether TINA captures every detail of Brazilian Portuguese phonetics, but whether it captures enough structural regularity to behave in a linguistically meaningful way.

2.3 Pseudowords as a linguistic benchmark

This perspective has a direct consequence for evaluation. A model like TINA should not be judged only by how well it reproduces transcriptions for held-out lexical items. Lexicon-based evaluation remains necessary, but it mainly tests whether the system can generalize within the distribution of known words. For a structured pronunciation model, the more revealing question is what happens when the input is novel. Pseudowords are therefore not a peripheral add-on to the project: they are one of the clearest ways to test whether the system has learned productive regularities rather than merely storing lexical patterns.

Pseudowords are especially informative in the present case because they can be evaluated at exactly the levels that TINA was built to represent. For Brazilian Portuguese, this is not an abstract methodological preference: recent experimental work shows that native speakers' stress assignment in pseudoword reading is shaped by phonological similarity to real words, the frequency of the lexical base, and the shape of the final syllable, which makes pseudowords a particularly useful way to test whether a model reproduces productive speaker-like regularities rather than only memorized lexical outputs; importantly, that final-syllable effect also converges with long-standing analyses of BP stress that treat syllable weight as a real structural factor (Benevides, 2024; Wetzels,

2007). The first level is stress placement: does the model assign prominence in ways that resemble Brazilian Portuguese speakers' intuitions about new words? The second level is syllabification: does it parse unfamiliar material into plausible syllable structures? The third level is segmental and repair behavior: when a novel form creates pressure on coda structure or consonant clustering, does the system respond with the kinds of adjustments that are characteristic of the language, including the kinds of coda-sensitive weakening and restructuring that BP phonology has long treated as central rather than peripheral (Medeiros, 2012)? Because TINA predicts syllables and stress explicitly and then expands abstract outputs into variant forms, pseudowords allow inspection of all three levels in a unified way.

This benchmark is also a good test of the hybrid design itself. If the neural component has learned useful generalizations about syllable structure and stress, and if the rule layer is encoding the right kinds of repair and dialectal variants, then the combined system should extend those patterns to forms it has never seen before. If it fails, the failures are linguistically interpretable: they can reveal overregularized stress assignment, implausible syllable parses, or rule expansions that are too categorical or too permissive. In that sense, pseudowords do more than provide another test set. They expose what kind of knowledge the model has actually internalized and, more importantly, some of the more opaque phonological aspects of BP.

For this reason, the most important contribution of this paper lies in the pseudoword results rather than in lexical output inspection or the offer of TINA as a tool for scholars in need of automated BP phonetic transcriptions. In practice, that means comparing TINA's behavior on novel forms with the kinds of stress patterns reported for Brazilian Portuguese speakers in pseudoword reading experiments, such as done by Benevides, 2022, 2024. TINA, therefore, is most interesting when it is treated as a model of structured Brazilian Portuguese pronunciation and then confronted with novel forms that require genuine generalization. Accordingly, the paper uses lexical output inspection as a baseline check and treats pseudowords as the more linguistically revealing benchmark.

3 TINA

This section describes the implemented TINA neural network architecture. The system is designed as a structured grapheme-to-pronunciation pipeline for Brazilian Portuguese in which orthographic input is first mapped to a relatively abstract, syllabified, stress-marked representation, and only then expanded into more surface-oriented variants. In that sense, the method is hybrid in two complementary ways: internally, because the neural model predicts several interdependent levels of structure rather than a single flat string; and externally, because a rule layer converts the neural output into dialect-sensitive realizations.

3.1 Data and representations

The training lexicon is stored as tab-separated orthographic words paired with phonological transcriptions. The training lexicon used contains 95,927 such pairs. The source side is a character sequence taken directly from the orthographic word. The target side is a transcription wrapped in slashes, with periods marking syllable boundaries and the symbol ' marking the stressed syllable, following IPA conventions. Thus, the target representation already encodes information that many transcription systems would leave implicit. In that respect, TINA treats pronunciation as something richer than a flat phoneme string, in the same general space of tasks where phoneme sequence, syllabification, and stress are modeled together (van Esch et al., 2016). A form such as /a.'ba.Dɪ/ is not merely a string of segments: it also specifies how those segments are grouped into syllables and where lexical prominence falls. This is also close to recent work that reintroduces syllable-sensitive structure into neural G2P rather than treating pronunciation as a completely undifferentiated output string (Lo & Nicolai, 2021).

The dataset wrapper converts each lexical entry into several aligned supervision channels. For a word such as *abade* /a.'ba.Dɪ/, the code derives (i) a syllable sequence a + ba + Dɪ, (ii) a binary stress sequence aligned to those syllables, and (iii) a phoneme sequence for each syllable. In addition, the expected number of syllables is computed directly from the transcription by counting

the periods and adding one. This makes syllable count available as an explicit supervisory signal rather than only as a by-product of decoding.

Table 1: Representational levels used by TINA. The important point is that the model is supervised not only on segmental content but also on syllable structure, stress, and sequence length.

Level	Basic unit	Special symbols	Implementation details
Orthographic input	Character	<pad_char>	Lexicon character vocab; padded/truncated to 30 positions.
Syllable target	Whole syllable token	<bow>, <eow>, <pad_syl>	Atomic syllable vocab; teacher sequence with boundary markers, max 10 positions.
Stress target	Binary label	0 or 1	Per-syllable binary label; stressed = 1, all others = 0.
Phoneme target	Phoneme symbols	<bop>, <eop>, <pad_ph>	Per-syllable phoneme sequence; padded/truncated to 12 positions.
Count target	Number of syllables in the word	—	Derived automatically from the number of syllable separators in the gold transcription.

This organization leads to three distinct vocabularies on the neural side: a character vocabulary for the input, a syllable vocabulary for the first decoder, and a phoneme vocabulary for the second decoder. In the current lexicon, these correspond to 38 source characters, 2,751 syllable types, and 45 phoneme symbols once the relevant padding and boundary tokens are included. The choice to predict syllables as whole units is especially important. It means that the first decoder is

not simply inserting syllable boundaries into a phoneme stream; it is predicting syllabic objects in their own right, which then serve as structural anchors for the rest of the model.

The target notation is also intentionally mixed in its degree of phonological specification. Many entries contain ordinary IPA-like segment symbols, but the lexicon also includes uppercase archiphonemic symbols such as D, H, L, R, S, and T. These symbols encode classes of realizations rather than a single fixed surface output. For instance, forms such as /a.ba.'Di.a/, /a.baw.Ho.'a.dʊ/, /a.ba.fa.'doR/, /a.'ba.saS/, and /a.ba.e.Ti.'tu.ba/ make clear that the training representation is already partially abstract. This is crucial for the architecture as a whole. The neural model is not forced to decide every dialect-sensitive or context-sensitive realization inside the decoder itself; instead, it can output a structured phonological representation whose remaining ambiguities are resolved by the subsequent rule layer. In other words, in terms of BP phonetic transcriptions, the neural model output itself is variant- or dialect-agnostic at the inference step.

3.2 Neural architecture: brief overview

TINA is implemented as a shared encoder followed by three coordinated decoders. The information flow is deliberately staged. Orthographic characters are encoded once; a syllable decoder then predicts a sequence of syllabic units; a phoneme decoder generates the segmental content of each predicted syllable; and a stress decoder assigns prominence over the resulting syllable sequence while also consulting phoneme-derived information. This kind of staged organization is also conceptually close to hierarchical sequence models that aim to preserve relationships among larger and smaller linguistic units instead of flattening everything to a single level (Ronanki et al., 2017). The result is not a set of independent subtasks but a single architecture in which each level constrains the next, in line with a multitask learning view where related outputs share representations and provide useful inductive bias to one another (Caruana, 1997). As such, we can affirm it is a single phonologically-driven model that treats segmental content, syllabification, and stress as jointly predicted structure rather than as separate downstream annotations.

The encoder begins with trainable character embeddings of dimension 128. These embed-

dings pass through a one-dimensional convolution with kernel size 3, batch normalization, and a ReLU nonlinearity. The convolutional output is added back to the original embeddings through a residual connection, allowing local graphemic patterns to be encoded without discarding the original character representation. The sequence then enters a two-layer bidirectional GRU with hidden size 256 and dropout 0.1. After recurrent encoding, the output is layer-normalized, projected back to the working hidden dimension, and passed through a Transformer encoder layer with four attention heads (Vaswani et al., 2017). Including this self-attentional stage is also consistent with later G2P work showing that Transformer-based models can perform competitively relative to recurrent and convolutional alternatives in grapheme-to-phoneme conversion (Yolchuyeva et al., 2019). The encoder therefore combines three kinds of context aggregation: local pattern extraction in the convolutional front end, sequential context in the bidirectional recurrent layer, and broader self-attentional recontextualization in the Transformer layer (Vaswani et al., 2017; Yolchuyeva et al., 2019). The encoder produces both contextualized time-step representations and a compressed hidden state that initializes downstream decoding. In addition, it contains a small regression head that predicts the number of syllables in the word from the mean of the encoded sequence. This count prediction has two functions. During training it provides an auxiliary supervisory signal tied to word-level structure, giving the shared encoder explicit pressure to encode a global property of the word alongside the sequence-decoding tasks (Caruana, 1997). During inference, when gold syllable counts are unavailable, the predicted count is rounded and used to guide the syllable decoder’s stopping behavior.

The first decoder predicts syllabic structure. It is a GRU decoder with attention over the encoder outputs and with <bow> and <eow> markers delimiting the syllable sequence. In that respect, the architecture does not treat syllabification as an afterthought, but as an explicit intermediate structure inside the pronunciation pipeline, which is also the general direction advocated in recent linguistically informed G2P work (Lo & Nicolai, 2021). In this respect, the decoder belongs to the same broad G2P lineage in which attention is used to learn soft, largely monotonic input–output correspondences without requiring an external aligner (Toshniwal & Livescu, 2016).

In the architecture this mechanism is explicitly implemented as Luong-style attention: the current decoder state consults the encoder states to build a context vector that conditions the next prediction (Luong et al., 2015). At each step, the decoder consumes the embedding of the current syllable token together with this attention-derived context vector from the encoder. The resulting hidden state is transformed into a shared syllable representation. From that shared representation, the decoder branches in two directions. One branch predicts the next syllable token from the syllable vocabulary. The other branch predicts a boundary-related signal that is later used both for auxiliary supervision and for a second estimate of syllable count. The crucial output of this first decoder is therefore not merely a boundary sequence but a sequence of hidden syllable states, each of which stands for one predicted syllabic unit in the word.

The second decoder operates once per predicted syllable. For each syllable position, the corresponding syllable hidden state is used as the initial state of a phoneme GRU decoder. If necessary, that state is repeated across decoder layers so that it matches the two-layer phoneme decoder. In other words, segmental realization is explicitly conditioned on a previously predicted syllabic scaffold rather than being decoded as one undifferentiated phone string from the outset, again following the broader intuition that supra-segmental and segmental structure can be modeled more effectively when their relations are preserved (Ronanki et al., 2017). Decoding proceeds autoregressively from <bop> until <eop> or a maximum number of steps is reached. Like the syllable decoder, the phoneme decoder attends back to the encoder rather than relying only on the initial state, again following the same attention-based, alignment-light G2P logic and the same Luong-style decoder-to-encoder consultation pattern (Luong et al., 2015; Toshniwal & Livescu, 2016). It also applies a residual refinement step before projecting to phoneme logits. Importantly, the phoneme decoder returns not only token predictions but also a refined hidden representation for each phoneme step. These phoneme-level hidden states are average-pooled within each syllable to create a syllable-level phoneme feature vector.

The stress decoder is the final neural component and is explicitly conditioned on both syllable structure and phoneme content. This design is close in spirit to hierarchical stress-related modeling

that combines phoneme-, syllable-, and word-level context and benefits from retaining phoneme information in the prediction pathway (Lin et al., 2020). It takes the sequence of syllable hidden states, adds positional embeddings, and processes the resulting sequence with a bidirectional GRU. A Transformer layer is then applied over the syllable sequence, after which the contextualized syllable representation is projected back to the hidden dimension (Vaswani et al., 2017; Yolchuyeva et al., 2019). The pooled phoneme feature for each syllable is projected into the same space and added to this syllable representation. Finally, for each syllable position, the decoder attends once more to the encoder outputs and predicts a binary stress label. Stress assignment is thus not reduced to a purely local choice on the syllable string alone. It is informed by the syllable sequence, by the segmental makeup of each syllable, and by the original orthographic encoding.

One practical consequence of this design is that the system can reconstruct a fully structured transcription during inference even without gold targets. Another consequence is that its aim becomes not merely to achieve success on BP phonetic transcriptions from orthographic words, but that it does so in a manner that is phonologically plausible architecturally. The encoder first predicts a word-level count estimate, the syllable decoder builds a syllable sequence, the phoneme decoder fills in each syllable internally, and the stress decoder marks which syllable receives prominence. The inference script then prefixes ' to syllables whose stress classifier selects the stressed class, removes empty or padded outputs, and joins the surviving syllables with periods. At that stage the system has produced a base transcription that is syllabified and stress-marked, but not necessarily fully surface-specified.

3.3 Training objectives

The multi-level architecture is matched by a multi-objective training regime. Rather than optimizing only one terminal phoneme sequence, TINA combines losses defined at the syllable, phoneme, stress, and count levels, plus an additional boundary-related auxiliary term (Lin et al., 2020). This is another place where the architecture's design converges with linguistically motivated G2P work that treats syllable boundaries as useful structure in their own right rather than as incidental

formatting (Lo & Nicolai, 2021). This matters methodologically because it makes linguistic structure part of the optimization target itself rather than something inferred only indirectly from the final output, and because the auxiliary tasks are intended to improve the shared representation rather than merely decorate the final prediction (Caruana, 1997). That general choice is also consistent with prior work showing that one model can jointly predict pronunciation, syllabification, and stress rather than relegating the latter two to a post-processing stage (van Esch et al., 2016). The total training objective, thus, is a weighted sum of five components:

$$\mathcal{L} = 0.30 \mathcal{L}_{\text{syl}} + 0.30 \mathcal{L}_{\text{stress}} + 0.20 \mathcal{L}_{\text{phon}} + 0.20 \mathcal{L}_{\text{align}} + 0.10 \mathcal{L}_{\text{boundary}}.$$

The syllable loss is a cross-entropy objective over the syllable vocabulary with label smoothing set to 0.1. Because the syllable decoder predicts whole syllable tokens, this loss encourages the model to learn recurrently useful syllabic units rather than only local boundary decisions. The phoneme loss is another cross-entropy term with the same amount of label smoothing, applied to the phoneme decoder outputs after they have been aligned with the number of actually decoded syllable steps. The stress loss is a focal loss with $\gamma = 2.0$ and class weights that upweight the stressed class, which is sensible because only one syllable in a word normally bears primary stress. Stress is computed only on valid syllable positions, excluding padding and end-of-word symbols.

Count supervision enters in two places. The encoder’s syllable-count head predicts a scalar estimate of the number of syllables in the word. In parallel, the boundary branch of the syllable decoder produces a second count estimate by summing the probability of boundary events across decoder steps and adding one implicit initial syllable. These two estimates are compared with the gold syllable count by mean squared error and averaged into a single alignment loss. This auxiliary term encourages the decoder not only to predict the right syllables, but also to predict them in roughly the right number.

Table 2: Training objectives in TINA. The loss design mirrors the structure of the output representation itself.

Loss	Supervision Source	Criterion	Function in the Model
$\mathcal{L}_{\text{syl}}$	Gold syllable-token sequence	Cross-entropy with label smoothing (0.1)	Trains the syllable decoder to recover structured syllabic units.
$\mathcal{L}_{\text{stress}}$	Binary stress labels per syllable	Focal loss ($\gamma = 2.0$) with class weighting	Trains prominence assignment while compensating for class imbalance.
$\mathcal{L}_{\text{phon}}$	Gold phoneme sequence inside each syllable	Cross-entropy with label smoothing (0.1)	Trains the phoneme decoder to realize the internal segmental content of each syllable.
$\mathcal{L}_{\text{align}}$	Gold syllable count	Mean squared error	Regularizes decoder length through encoder-based and boundary-based count prediction.
$\mathcal{L}_{\text{boundary}}$	Boundary or termination labels derived from the teacher sequence	Cross-entropy	Provides an auxiliary boundary signal to the syllable decoder.

Training uses ADAM (Adaptive Moment Estimation) with learning rate 0.001, a batch size of 16, dropout 0.1 throughout the model, cosine annealing over a nominal 100-epoch horizon, and gradient clipping with maximum norm 1.0. Teacher forcing is used in both the syllable and phoneme decoders, but its probability is exponentially decayed across epochs according to $e^{-\text{epoch}/30}$. Early in training, the model therefore relies more heavily on gold previous outputs; later, it is pushed to cope

with its own predictions, following the same broad motivation as scheduled-sampling style training procedures that reduce the mismatch between fully guided training and autoregressive inference, as demonstrated by Bengio et al., 2015. Checkpoints are written every epoch, and the inference script reloads the saved model weights on demand before transcription.

3.4 Rule-based phonological expansion

The neural model does not attempt to emit a fully surface-specific transcription in one step. Instead, inference is divided into two phases. First, the neural stack produces a syllabified and stress-marked base transcription. Second, a variant-generation function expands that transcription into a set of surface-oriented variants, working as a rule engine for surface forms. The hybrid character of TINA is clearest at this point. While the neural component learns broad pronunciation structure, the variant-generation component translates that structure into a controlled space of realizational possibilities, in a design that is conceptually close to earlier systems combining data-driven mapping with explicit phonological structure (Meng et al., 1994). This division is also especially apt for Brazilian Portuguese, where coda behavior is a major locus of gradient and position-sensitive realization at the phonetics–phonology interface (Medeiros, 2012). This rule engine builds a set of variants in stages. The first stage inserts the canonical BP epenthetic /i/ after certain consonants in coda position, including word-final codas and codas immediately before a syllable boundary. The second stage handles consonant-plus-/s/ clusters, again allowing epenthetic /i/ in both onset-like and coda-like environments. These rules are applied combinatorially, so a single base form may yield several candidates if more than one position licenses epenthesis.

A further stage handles reduced-vowel and coronal alternations. When the transcription contains the abstract markers I or U, the rule engine generates corresponding vowel alternatives: /i/~/ɪ/ and /u/~/ʊ/ respectively. If those markers are preceded by t or d, the same step may also generate palatalized variants such as tʃ and dʒ. Although these markers are more abstract than the main uppercase symbols found in the lexicon, their presence in the rule module shows that the post-neural layer is intended to cover both dialectal variation and context-sensitive vowel/consonant

alternation.

The last stage applies dialect profiles. Each profile specifies how ambiguous symbols are to be realized, and all occurrences of the same symbol inside a word are replaced consistently. The three profiles implemented in the system design are Northern or Northeastern dialects (*Nortista/Nordestino*), Southern or Southeastern dialects (*Sulista/Sudestino*), and the *Carioca* dialect. They differ most clearly in the treatment of rhotics, sibilants, and coronal stops, especially in the rhotic domain, where Brazilian Portuguese variation is well known to cut across the phonetics–phonology boundary (Silva & Albano, 1999). It is important to note that this is not a claim that these regions are homogeneous in terms of dialectal pronunciation, but within this architecture they function as broad heuristic profiles for regional surface forms, since they capture recurrent phonetic tendencies despite substantial internal variation (Callou, 2009; Cardoso, 1986).

Table 3: Core dialect-sensitive mappings in the rule layer. Additional mappings such as $L \rightarrow \mathcal{K}$ are also defined, but the main contrasts relevant for the present paper involve rhotics, sibilants, and coronal stop alternations.

Symbol	Interpretation in the Base Output	Northern or Northeastern	Southern or Southeastern	Carioca
H	Dorsal/uvular rhotic class, initial rhotics	ʁ, h	ʁ, h	ʁ, h
T	Context-sensitive coronal stop class	t	tʃ	tʃ
D	Voiced counterpart of T	d	ɖ	ɖ
S	Sibilant class, especially salient in coda behavior	ʃ, s	s	s, ʃ

Continued on next page

Table 3 (*continued*)

Symbol	Interpretation in the Base Output	Northern or Northeastern	Southern or Southeastern	Carioca
R	Rhotic class, including strong and final rhotics	ʁ, h, ʒ	ɾ, ɽ	ʁ, ʒ, h

The implementation also includes special handling for rhotics in two environments. A word-final R may be deleted in the surface variant, and R in RS clusters may also be dropped unless it is realized as a tap or retroflex rhotic, following known BP phonotactics (Alencar, 2007; Callou et al., 2015). After all replacements are applied, the resulting strings are cleaned up by removing empty segments and duplicated syllable separators. The final output of the rule engine is therefore not one privileged pronunciation but a deduplicated list of variants. Furthermore, a modest vowel harmony module is implemented, as will be demonstrated in the following section, in Table 5.

This separation between neural prediction and rule-based expansion is one of the most important design choice in TINA. It lets the neural network focus on learning stable regularities in orthography-to-structure mapping while preserving room for controlled phonological and dialectal variation downstream. For Brazilian Portuguese, this is an especially natural division of labor, since coda-related phenomena are precisely one of the domains in which gradient realization, weakening, and vocalization become especially salient (Medeiros, 2012). Stress placement, syllable count, and broad segmental organization are modeled directly inside the neural architecture, while epenthesis, rhotic behavior, and dialect-sensitive alternations are handled explicitly after the fact. The result is not a purely end-to-end transcriber and not a purely symbolic phonological analyzer, but a compact hybrid system in which each component is responsible for a different layer of the pronunciation problem.

4 Experimental design

This section defines the evaluation used in the paper. The central test is a pseudoword benchmark that asks whether TINA generalizes to novel items in ways that resemble Brazilian Portuguese speaker behavior. Before that benchmark, a qualitative inspection of outputs on familiar lexical material establishes that the system produces well-formed structured transcriptions. Both are evaluated in a way that respects the internal organization of the model described in Section 3: pronunciation is treated not as a flat string alone, but as a structured object involving syllabification, stress placement, and a later rule-based expansion stage, in the same general spirit as work that models phoneme sequence, syllabification, and stress together, or that injects explicit syllable-boundary information into G2P rather than treating syllable structure as a separate afterthought (Lo & Nicolai, 2021; van Esch et al., 2016).

Table 4: The evaluation design used in the paper. The lexical inspection establishes output well-formedness on familiar material; the pseudoword benchmark probes productive generalization to unfamiliar forms.

Evaluation Track	Input Material	Primary Outputs Inspected	Question Asked
Lexicon-based evaluation	Sample lexical items drawn from the training lexicon for illustrative inspection	Structural well-formedness of base transcriptions: syllabification, stress marking, and archiphonemic output	Does TINA function as a competent structured transcriber on ordinary lexical data?

Continued on next page

Table 4 (*continued*)

Evaluation Track	Input Material	Primary Outputs Inspected	Question Asked
Pseudoword evaluation	Orthographically written pseudowords without accent marks; direct speaker comparison restricted to the relevant human-study subsets	Stress placement, syllabification, and segmental/repair behavior in the base and expanded outputs	Does TINA generalize to novel forms in ways that resemble Brazilian Portuguese speakers?

4.1 Lexicon-based evaluation

The lexical track is not a formal held-out benchmark. No separate test split was evaluated against gold references, and no phoneme error rate or exact-word accuracy is reported. Instead, a sample of items from the training lexicon is used to illustrate that the system produces structurally coherent outputs: syllable boundaries are placed plausibly, stressed syllables are marked, and archiphonemic symbols appear in the expected positions. This kind of qualitative inspection is sufficient to establish that the model does not produce malformed strings on familiar material before the paper turns to the more demanding pseudoword test. The primary evidence for transcription competence therefore comes from the structure of the outputs themselves rather than from a quantitative comparison to held-out gold transcriptions. Our aim here is make sure the transcriptions are coherent with BP phonotactics and dialectal variants in general. Below is a table with a sample of transcriptions realized by TINA:

Table 5: Transcription outputs by TINA.

Input	Output(s)
martins	/ˈmah.tĩs/, /ˈmah.tĩʃ/, /ˈmah.tʃĩs/, /ˈmah.tʃĩʃ/, /ˈmaɾ.tʃĩs/, /ˈmaɾ.tʃĩʃ/, /ˈmaɐ̯.tĩs/, /ˈmaɐ̯.tĩʃ/, /ˈmaɐ̯.tʃĩs/, /ˈmaɐ̯.tʃĩʃ/, /ˈmaχ.tĩs/, /ˈmaχ.tĩʃ/, /ˈmaχ.tʃĩs/, /ˈmaχ.tʃĩʃ/
carambola	/ka.rã.ˈbɔ.la/
carteira	/kah.ˈtej.ra/, /kaɾ.ˈtej.ra/, /kaɾ.ˈtej.ra/, /kaɐ̯.ˈtej.ra/, /kaχ.ˈtej.ra/
cantinho	/kã.ˈtĩ.ɲʊ/, /kã.ˈtʃĩ.ɲʊ/
polido	/po.ˈli.dʊ/, /po.ˈʎi.dʊ/, /pu.ˈli.dʊ/, /pu.ˈʎi.dʊ/
polidinho	/po.li.ˈdĩ.ɲʊ/, /po.ʎi.ˈdʒĩ.ɲʊ/
teixeira	/tej.ˈfej.ra/
fonética	/fo.ˈnɛ.ti.ka/, /fo.ˈnɛ.tʃi.ka/, /fɔ.ˈnɛ.ti.ka/, /fɔ.ˈnɛ.tʃi.ka/
seguro	/se.ˈgu.rʊ/, /si.ˈgu.rʊ/
perigo	/pe.ˈri.gʊ/, /pi.ˈri.gʊ/
menino	/me.ˈnĩ.ɲʊ/, /mi.ˈnĩ.ɲʊ/
peruca	/pe.ˈru.ka/, /pi.ˈru.ka/
pepino	/pe.ˈpĩ.ɲʊ/, /pi.ˈpĩ.ɲʊ/
tio	/ˈti.ʊ/, /ˈtʃi.ʊ/

Continued on next page

Table 5 (*continued*)

Input	Output(s)
pai	/'pa.j/
mãe	/'mã.j/

Table 5 is a sample showcase of its performance. Transcriptions are well-formed structurally, stress is assigned correctly, and the rule engine produces several alternations in the appropriate regions: /ʌ/~/l/, /d/~/dʒ/ and /t/~/tʃ/, in addition to implementing coda variants when applicable. Furthermore, we can observe the operation of vowel harmony in cases such as /pe.'pĩ.nũ/~/pi.'pĩ.nũ/ or /se.'gu.rũ/~/si.'gu.rũ/. Given these outputs, we can proceed to the pseudoword evaluation against native speakers.

4.2 Pseudoword evaluation against native speakers

The central empirical probe uses the lexical-processing pseudoword benchmark reported in Benevides’s journal article and documented in fuller methodological detail in the corresponding dissertation (Benevides, 2022, 2024). On the speaker side, this benchmark is grounded in an orthographic reading task in which the stimuli are presented without accent marks; the broader dissertation also motivates that choice by comparing orthographic and image-based elicitation and finding no meaningful advantage for the latter for the present purpose (Benevides, 2022, 2024). This makes the comparison especially suitable for TINA, since the model also receives only orthographic input and must infer stress from the form itself rather than from explicit diacritics. It also keeps the benchmark close to actual spoken stress assignment rather than to metalinguistic labeling alone: in Brazilian Portuguese, lexical stress perception is supported by acoustic cues such as duration and vowel-quality/formant pattern, and duration, F0 variability, and spectral emphasis remain informative stress cues across speaking styles, with phrase reading patterning closely with more spontaneous speech (Barbosa et al., 2013; Fontes, 2013).

The stimuli are primarily trisyllabic and are built so that all three major stress patterns remain possible. In the lexical-processing experiment summarized in Benevides (2024), they are restricted to the templates CV-CV-CV and CV-CV-CVC, which makes final-syllable structure an explicit part of the design. This choice is also linguistically well motivated for Brazilian Portuguese, since analyses that treat non-verbal stress as sensitive to syllable weight make exactly the open-versus-closed contrast at the right edge central to the system and argue that newly created words reveal the same tendency (Wetzels, 2007). It is also phonologically meaningful beyond stress alone, because BP codas are a privileged site of weakening, vocalization, and gradient realization, so a final CVC pseudoword probes a structurally loaded position rather than a trivial segmental variant (Medeiros, 2012). The benchmark also crosses phonological similarity to a real base word with the frequency of that base word, yielding four experimental groups: similar high-frequency, dissimilar high-frequency, similar low-frequency, and dissimilar low-frequency (Benevides, 2022, 2024). Because the materials were subjected to an independent similarity-validation stage before analysis, the usable speaker-side sets are smaller and less balanced than the total stimulus inventory, and this asymmetry is part of the benchmark rather than something to be normalized away in the present paper (Benevides, 2022, 2024).

For the present paper, the pseudoword benchmark is used in two related ways. First, the full pseudoword list is passed through TINA so that the model’s default behavior on unfamiliar forms can be characterized exhaustively. Second, whenever the paper directly compares model behavior with speaker behavior, the comparison is restricted to the validated and retained speaker-side subsets for which the experimental conditions are explicitly defined. This distinction matters because the model was run over the complete 372-item stimulus set, while the human benchmark reported in Benevides (2024) is analyzed over a validated 309-item subset read by 34 native speakers, with 113 production-level exclusions yielding 10,393 usable tokens (Benevides, 2024). Different totals therefore correspond to different stages of the experimental pipeline rather than to a contradiction in the benchmark itself. The actual comparison is structured at three levels. The first and most important is stress placement. Here the question is not simply whether TINA reproduces the intended

stress of a base word, but whether it shows the same kinds of conditioning effects that speakers show when reading novel forms: overall stress-pattern preferences, sensitivity to target stress category, sensitivity to the similarity/frequency grouping used in the lexical-processing experiment, and sensitivity to the open-versus-closed character of the final syllable (Benevides, 2022, 2024; Wetzels, 2007). The second level is syllabification. Because TINA predicts syllables explicitly, the paper can ask whether a pseudoword receives a plausible syllable parse and whether deviations in stress are connected to deviations in syllable structure. That kind of inspection is also consistent with syllable-aware G2P work that treats onset/coda-sensitive boundary structure as useful linguistic information rather than redundant markup (Lo & Nicolai, 2021). The third level is segmental and repair behavior. Since the neural output is partly archiphonemic and the rule layer expands that output into more surface-oriented variants, novel forms also allow inspection of how the system handles coda pressure, epenthesis, and other repair-like behavior that would be invisible in a stress-only evaluation. This is especially relevant in BP, where coda realization is not phonetically inert but tied to recurrent weakening and vocalization patterns, including nasal-coda behavior, that make the mapping from abstract representation to surface form particularly informative on unfamiliar items (Medeiros, 2012).

This design also respects the hybrid architecture of the system. The immediate output used for the main pseudoword comparison is the archiphonemic transcription generated by the neural model, because that is where syllable boundaries and stress are predicted directly. The rule module is then treated as a second-stage probe that shows how those same novel forms are expanded into variant surface realizations. This division of labor is especially useful for pseudowords: it lets the paper distinguish between errors of structural prediction and errors or overextensions introduced by the rule layer. All 372 pseudowords received a transcription and each item produced a unique base output. The results are organized around overall stress-pattern distribution, production by target stress category, the effect of final-syllable structure, similarity and frequency sensitivity, and group-specific stress-match behavior. The speaker-side interpretation of these dimensions follows the published lexical-processing studies in treating stress assignment in pseudowords as the result of

interacting factors rather than of a single categorical rule, while the specific attention to final-syllable weight connects naturally to the broader BP stress literature (Benevides, 2022, 2024; Wetzels, 2007).

5 Results and analysis

This section reports the evaluation results. A brief qualitative inspection of lexical outputs confirms that the system produces well-formed structured transcriptions on familiar material. The substantive evaluation is the pseudoword benchmark, where the model must generalize beyond memorized lexical entries and where the hybrid design becomes analytically useful.

5.1 Lexical outputs: structural well-formedness

A sample of items from the training lexicon confirms that TINA produces well-formed structured transcriptions on familiar material. Syllable boundaries are placed at plausible positions, the stressed syllable is consistently marked, archiphonemic symbols appear where expected, and outputs do not contain malformed or empty strings. No quantitative comparison against held-out gold transcriptions is reported here: the lexical inspection serves only to establish that the system is a functioning structured transcriber before the paper asks the more revealing question of how it behaves on novel forms.

5.2 Pseudowords: where TINA behaves like speakers

The pseudoword batch is operationally stable. All 372 pseudowords received a base transcription and no runtime errors were recorded. The outputs also remain structurally well formed across the batch, with three-syllable analyses clearly dominating and only a very small residue collapsing to two syllables. Before any comparison to speaker behavior is made, these facts already show that the model does not lose control of syllable count or output formatting when faced with novel forms.

A methodological clarification is necessary before the comparison. In Benevides's Experiment 1, 372 pseudowords were originally constructed and read by 34 speakers, yielding 12,648 raw productions; the thesis then tracks validation status explicitly, distinguishing validated, quasi-validated, and non-validated stimuli, and the later article condenses the same experiment into the benchmark comparison most directly relevant here (Benevides, 2022, 2024). The 309-item human benchmark used in the the speaker comparison corresponds to the validated plus quasi-validated portion of that design, whereas TINA was run over the complete set of 372 pseudowords. The denominator mismatch should therefore be read as a benchmarking decision on the human side: speakers are compared on the validated-plus-quasi-validated subset, while the model was run over the complete designed stimulus set.

A second scope note concerns the level of comparison. What is evaluated here is categorical stress location, not its phonetic implementation. That distinction matters in Brazilian Portuguese, where the perceptual recovery of lexical stress depends on acoustic cues such as duration and formant pattern, and where duration, F0 variability, and spectral emphasis remain robust correlates of stress across speaking styles (Barbosa et al., 2013; Fontes, 2013). Agreement between TINA and the speaker benchmark should therefore be read as agreement about where prominence falls, not about how that prominence would be acoustically realized in speech. The same caution extends to segmental and coda implementation, since Brazilian Portuguese rhotics have been shown to vary in ways that are prosodically conditioned and that may be either categorical or gradient at the phonetics–phonology boundary, and BP codas themselves have been analyzed as gradient coordination patterns rather than purely symbolic slots, especially in the nasal case (Medeiros, 2012; Silva & Albano, 1999).

Table 6: Stimulus composition of Benevides’s Experiment 1, adapted from Table 5.3 in Benevides, 2022. The 309-item human benchmark used in the prepared comparison corresponds to the validated plus quasi-validated stimuli, whereas the repository-side TINA batch contains 372 pseudowords.

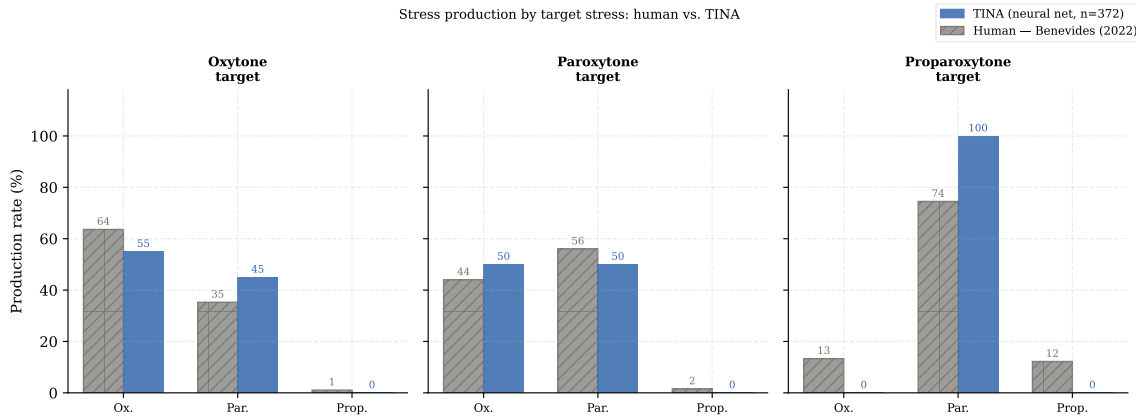
Group	Total	Validated	Quasi-validated	Non-validated	Kept in 309-item benchmark
G1	90	61	6	23	67
G2	90	75	11	4	86
G3	96	59	6	31	65
G4	96	74	17	5	91
Total	372	269	40	63	309

At the coarsest level of stress assignment, TINA behaves surprisingly well. Across the 372 pseudowords, it produces 183 final-stressed outputs (49.2%) and 189 penultimate-stressed outputs (50.8%), with no antepenultimate-stressed outputs at all. In Benevides’s speaker experiment, the corresponding aggregate distribution is 44.0% final-stressed, 52.3% penultimate-stressed, and 3.7% antepenultimate-stressed over 10,393 valid productions (Benevides, 2022, 2024). Table 7 places the two distributions side by side. The important point is not that the model reproduces the human distribution exactly, but that it captures a real structural tension in Brazilian Portuguese pseudoword stress. It does not simply collapse to a penultimate default or to a final-stress default. Instead, it distributes stress across the two main patterns in a way that is broadly plausible for novel forms, even though it suppresses the small but non-negligible antepenultimate stress residue documented in the human data detected by Benevides, 2022.

Table 7: Overall stress-pattern distribution in the speaker-comparison materials and in TINA’s 372 pseudoword outputs.

Stress pattern	Speaker comparison		TINA	
	n	%	n	%
Final-stressed	4,569	44.0	183	49.2
Penultimate-stressed	5,436	52.3	189	50.8
Antepenultimate-stressed	388	3.7	0	0.0
Total	10,393	100.0	372	100.0

Figure 1: Produced stress broken down by target-stress category in the prepared speaker comparison derived from (Benevides, 2022). The figure makes visible where TINA remains broadly aligned with speaker tendencies and where it collapses the comparison space, especially for antepenultimate-stress targets.



The broader comparison looks more uneven once the materials are organized by the target stress category of the source forms. Figure 1 shows that the model still tracks the final-versus-penultimate opposition in a directionally plausible way, but it compresses the antepenultimate space much more aggressively than the speaker data do. This does not contradict the overall near-even balance in Table 7; rather, it clarifies that the aggregate agreement is achieved by getting the main two-way opposition mostly right while eliminating a weaker but still meaningful third option. That weaker option matters in Benevides’s analysis because antepenultimate stress remains available to speakers on novel forms and is therefore treated as part of the active phonological grammar rather

than as a purely exceptional lexical residue (Benevides, 2022, 2024).

The strongest speaker-like behavior appears when the outputs are broken down by final-syllable structure. This is also the part of the comparison that connects most directly to formal analyses of Brazilian Portuguese stress: for non-verbs, weight-sensitive accounts place special importance on the contrast between open and heavy final syllables (Wetzels, 2007). In Benevides’s data, open final syllables overwhelmingly favor paroxytone production, whereas closed finals strongly favor oxytone production; the 2024 article makes this contrast especially clear by foregrounding final-syllable structure as a robust determinant alongside similarity and lexical frequency (Benevides, 2022, 2024). This is also the clearest pattern in TINA. The model is therefore not assigning stress arbitrarily; it is strongly sensitive to syllable structure at the right edge. What it appears to encode here, however, is the abstract structural contrast more than the full phonetics of coda realization, since BP codas have also been described as gradient coordination and vocalization patterns at the phonetics–phonology interface rather than as mere symbolic slot filling (Medeiros, 2012). More generally, this is the kind of result that becomes visible when syllable boundaries are kept explicit in the modeling pipeline rather than flattened away, which is also the perspective taken in recent G2P work that uses syllabification as useful linguistic structure rather than as post hoc annotation (Lo & Nicolai, 2021). The same pattern can be seen directly in the transcriptions. Open-final pseudowords such as *tarana* /ta.ˈrã.na/, *cadada* /ka.ˈda.da/, and *lumero* /lu.ˈmɛ.rʊ/ receive penultimate stress, whereas closed-final items such as *lapital* /la.pi.ˈtaw/, *siretor* /si.re.ˈtoR/, and *bocivel* /bo.si.ˈvew/ are pushed to final stress.

Table 8 gives the full stress distribution by final-syllable structure on both the speaker and model sides. The human benchmark is not fully categorical: even with CVC finals, speakers still produce 18.2% paroxytones and 2.4% proparoxytones, while with CV finals they still produce a small oxytone and paroxytone residue (Benevides, 2022). In other words, the model’s best generalization is not a lexical one but a structural one. It has learned that the right edge matters, and that final closure is a strong cue for stress assignment on unfamiliar words. Read against Wetzels’s weight-based analysis, this is the point at which TINA looks most linguistically well-anchored—but

also the point at which it starts to implement a human tendency too categorically (Wetzels, 2007).

Table 8: Stress distribution by final-syllable structure in the speaker benchmark from Benevides, 2022, Table 5.12 and in the full 372-item TINA batch. Speaker-side denominators are 5,355 tokens for open finals and 5,038 for closed finals; TINA denominators are 216 open-final and 156 closed-final pseudowords.

Final syllable	Speaker benchmark			TINA		
	Final	Penult.	Antepenult.	Final	Penult.	Antepenult.
CV-CV-CV (open)	10.6	84.4	5.0	12.5	87.5	0.0
CV-CV-CVC (closed)	79.4	18.2	2.4	100.0	0.0	0.0
Overall	44.0	52.3	3.7	49.2	50.8	0.0

This information is particularly important because it dispels any notion that the model is simply randomly assigning stress to the final and penultimate syllables at a 50% rate, as if no more precise than a coin toss. Here we can confirm that, although proparoxytone stress is not being modeled correctly, the remaining stress distributions clearly follow the openness (or lack thereof) of the final syllable, very much aligned with how native BP speakers produce pseudowords: 10% vs. 12.5% (open), and 79.4% vs. 100% (closed), due to the failure of modeling proparoxytone stress.

The hybrid character of the system also becomes visible on pseudowords once rule expansion is applied to the base outputs. Across the 372-item batch, most outputs remain single-variant after the rule layer, but a substantial minority generate multiple surface forms, in some cases reaching as many as 12 variants. The extra variants concentrate in outputs containing archiphonemic rhotics and coronal stops. For instance, *jomite* /ʒo.'mi.tɪ/ expands into /ʒo.'mi.tɪ/ and /ʒo.'mi.tʃɪ/, while *siretor* /si.re.'toR/ expands into several rhotic realizations, including /si.re.'toɾ/, /si.re.'toβ/, and /si.re.'to/. This does not prove speaker equivalence by itself, but it does show that TINA handles many novel forms in the intended two-stage way: a stable base representation first, then controlled surface diversification. That division of labor is linguistically useful precisely because part of BP coda behavior is more naturally treated as an abstract-to-surface mapping problem than as direct symbol copying, and because hybrid letter-to-sound traditions have long benefited from keeping phonological structure explicit rather than flattening everything into a single undifferentiated

mapping (Medeiros, 2012; Meng et al., 1994).

5.3 Error analysis and limits

The same results also show where the model is too rigid. The clearest failure is the total absence of antepenultimate-stressed outputs. Over the entire 372-item pseudoword batch, TINA produces 0 antepenultimate-stressed forms. This is not a rare accident but a systematic ceiling. Even items patterned on familiar proparoxytone words are regularized away from antepenultimate stress: *lumero* /lu.'mɛ.rʊ/, *neculo* /ne.'ku.lʊ/, *setodo* /se.'tɔ.dʊ/, and *metode* /me.'tɔ.dɪ/ all end up as paroxytones. The model therefore captures the broad final-versus-penultimate competition but not the weaker antepenultimate option that speakers still preserve on novel forms. In Benevides’s Experiment 1, antepenultimate-stressed productions account for 3.7% of the aggregate speaker data, and the rate remains much higher in the similar groups when the base word is itself antepenultimate-stressed (Benevides, 2022, 2024). In the present model, even when the comparison is restricted to antepenultimate-target items, the model never produces antepenultimate stress. The most reasonable explanation for this is, actually, quite straightforward. It is nothing more than a dataset problem. In Portuguese orthographic convention, all antepenultimate-stressed forms *must* receive a diacritic (acute and circumflex accents). This means that, as the model was trained on Portuguese words following orthographic conventions, all proparoxytones received a diacritic at the antepenultimate syllable. Since no pseudowords received diacritics, no pseudowords would be considered proparoxytones as per formal orthographic convention of written Portuguese. This shows us that the *direction* of learning and speech acquisition matters, as native speakers can improvise outside formal grammar conventions, as spoken language is learned before writing. G2P neural network models, though, are usually – if not always – trained under written forms first and cannot improvise as easily, because they follow the complete opposite path, in this case, from formal orthographic conventions to speech. This is a design problem that is not easily solved, as the introduction of non-marked proparoxytones in BP would most likely lead to a higher error rate in oxytone and paroxytone forms as ambiguity rises.

A second limitation is overregularization of closed finals. Table 8 already suggest that coda-sensitive stress is one of the model’s strengths, but it is also one of its slight excesses. Seen against a weight-sensitive account of Brazilian Portuguese non-verbal stress, TINA behaves like an exaggerated version of the same right-edge bias. Medeiros’s account of BP codas makes the same excess visible from another angle: the model treats final codas as clean categorical triggers, whereas the language’s coda behavior is not exhaustively captured by a simple symbolic closed-versus-open dichotomy (Medeiros, 2012). The same flattening appears in the group-wise findings. In Benevides’s analysis, human productions vary with similarity and with the interaction between similarity and lexical frequency, especially when the target pattern is oxytone or proparoxytone (Benevides, 2022, 2024). TINA, by contrast, is largely invariant across those contrasts. The model behaves much more like a structural default system than an analogical one: once syllable shape is fixed, the output is mostly fixed as well. In other words, TINA learns broad phonological regularities, but it does not appear to encode the finer neighbourhood-sensitive pressures that make speaker behavior gradient rather than categorical. On the other hand, the pseudoword benchmark experiment we are comparing TINA to has provided *words in context*, *i.e.*, pseudowords in a sentence (Benevides, 2022). TINA, however, transcribes it as a single input, a pseudoword completely decontextualized. So although results are directly comparable in terms of numerical values and data, we have to be careful in regards to the interpretation of such comparisons. It is difficult to deny that the *conditions* the native speakers were exposed to in terms of lexical content before and after the target pseudowords probably did affect how they would be perceived semantically and, thus, produced in speech.

Figure 2: Correct stress-match rate by experimental group and target category in the the speaker comparison derived from (Benevides, 2022). Human performance shifts substantially across the similarity and frequency groups, while TINA remains much flatter across conditions.

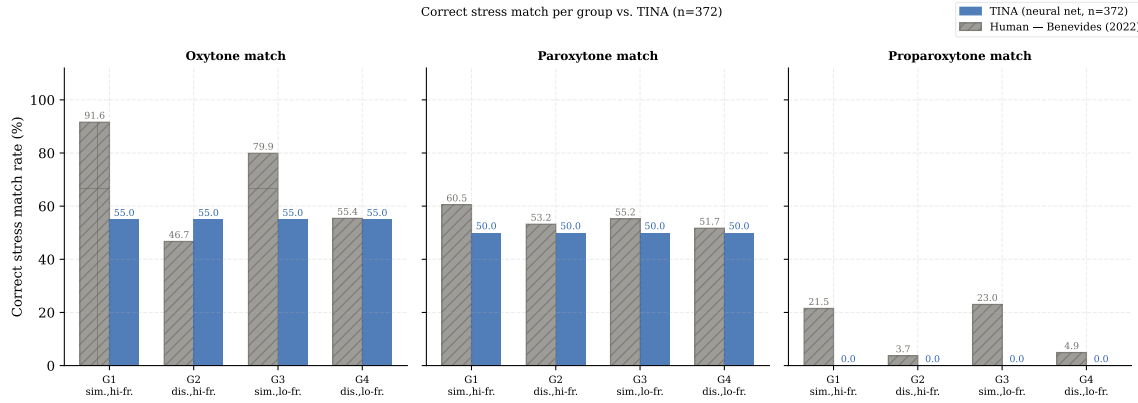


Figure 2 broadens that point across all three target categories. The largest gaps are not random errors distributed evenly across the space. They concentrate in precisely those conditions where the human comparison is most gradient: final-stress targets vary markedly by group, and antepenultimate targets remain possible for speakers but unavailable to the model. This is also where Benevides argues that *similarity* and *lexical access* are interacting with syllable structure rather than being reducible to it (Benevides, 2022, 2024). The neural side therefore looks structurally coherent but distributionally too flat.

Figure 3: Condition-by-condition divergence between the speaker benchmark derived from (Benevides, 2022) and TINA. The strongest mismatches cluster around residual non-final stress in closed-final words and around the proparoxytone responses that speakers still preserve but the model suppresses.

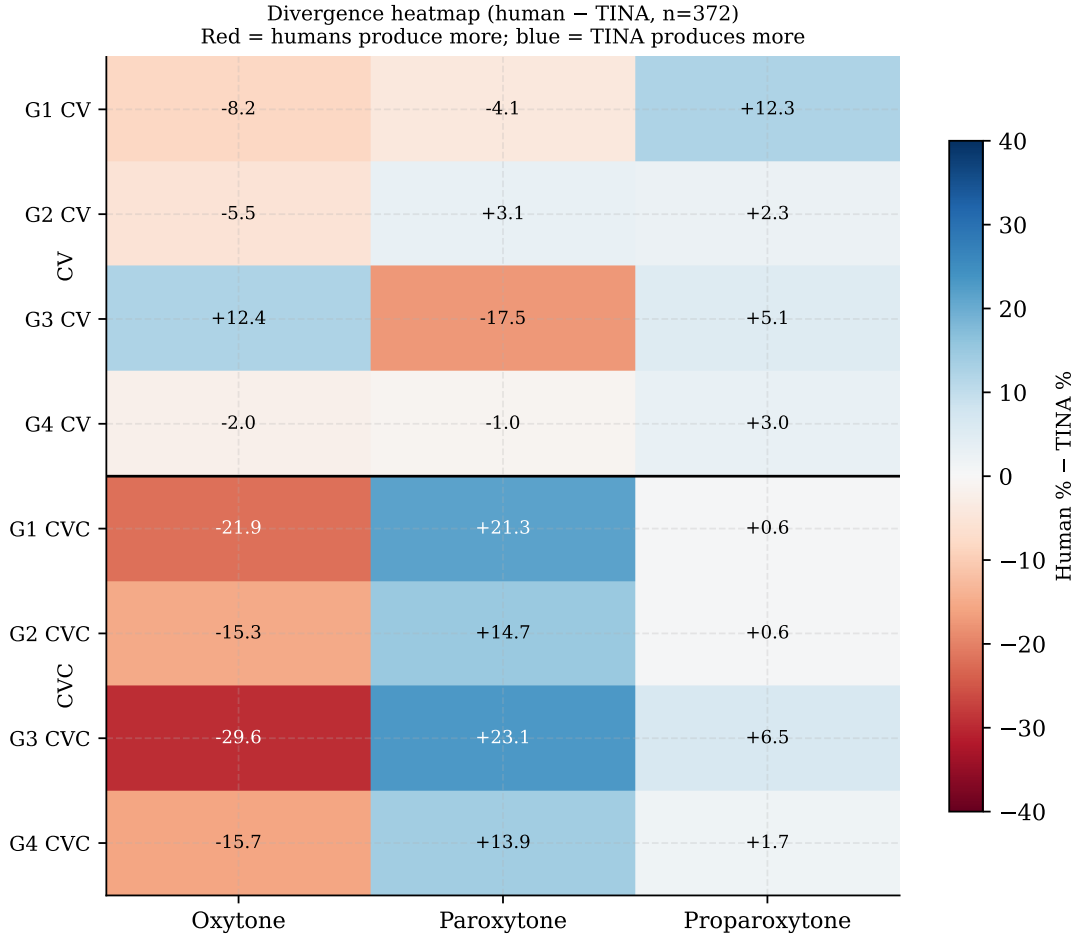


Figure 3 synthesizes the same imbalance from another angle. The mismatches are smallest where the model’s structural generalization is directionally correct, but they widen whenever speaker behavior remains non-categorical. In closed-final conditions, TINA tends to overproduce oxytones and underproduce the remaining paroxytone and proparoxytone responses. In open-final conditions, the main residual divergence lies in the human production of some proparoxytones that the model never generates. The figure therefore supports the same interpretation suggested earlier: TINA captures the main structural cue, but it sharpens that cue more than speakers do, whereas the speaker data in (Benevides, 2022, 2024) remain visibly gradient.

These limits clarify both the value and the scope of the hybrid design. TINA is at its best when the task can be decomposed into stable structure plus controlled variant expansion: syllable count, broad stress placement, and archiphonemic output are handled reliably, and the rule layer adds interpretable surface alternatives. It is weaker when the target behavior depends on gradient analogy, lexical comparison, semantic memory, partial productivity, or low-frequency patterns that speakers preserve without making fully categorical. That profile is still useful. It means that TINA is better understood as a compact model of Brazilian Portuguese pronunciation structure than as a system that exhausts the phonetics of speaker behavior, especially when the spoken realization of stress in BP is distributed across multiple acoustic dimensions rather than reducible to a single categorical marker, when at least some segmental domains such as rhotics show prosodically conditioned variability that is richer than the model’s discrete output space, and when BP codas themselves can involve gradient coordination and vocalization effects rather than simple slot filling (Barbosa et al., 2013; Fontes, 2013; Medeiros, 2012; Silva & Albano, 1999).

6 Discussion

The results in Sections 4 and 5 make it possible to locate TINA more precisely within the space of Brazilian Portuguese pronunciation modeling. The system is clearly more than a string transducer that maps orthography to an undifferentiated phone sequence. It predicts syllable structure, stress, and segmental material in a coordinated way, and it does so robustly enough that novel pseudowords still receive well-formed outputs. At the same time, the system is not a full model of spoken realization. Its neural output is deliberately abstract, and its rule-based expansion produces controlled surface alternatives rather than a gradient phonetic space. The main interpretive question, then, is not whether TINA “solves” Brazilian Portuguese pronunciation, but what kind of phonological and phonetic knowledge it appears to encode and where that knowledge stops.

6.1 What kind of pronunciation knowledge TINA appears to learn

The strongest positive result is that the model generalizes structurally. On the pseudoword batch, every item receives a transcription, almost all outputs remain trisyllabic, and the overall stress distribution is split almost evenly between final and penultimate stress. Taken together, these facts suggest that the model has learned more than lexical recall. Faced with unfamiliar forms, it still preserves a stable sense of what a plausible Brazilian Portuguese word should look like in terms of syllable count, syllable grouping, and stressable positions. That is an important result because the paper’s central claim is not that TINA reaches a new performance ceiling, but that it functions as a compact structured model of pronunciation. It is also exactly the kind of behavior one would want in the non-verbal domain of Brazilian Portuguese, where stress on new forms is expected to reflect prosodic structure rather than verbal morphology (Wetzels, 2007).

The pseudoword results also suggest that the model has learned a real right-edge structural asymmetry. Open-final forms are mapped overwhelmingly to penultimate stress, while closed-final forms are mapped categorically to final stress. This is not a trivial artifact of output formatting. It indicates that the system treats the final syllable as structurally informative for stress assignment and that it organizes this information at the syllabic level rather than only through local segment-to-segment correspondences. In other words, the model seems to encode a broad generalization of Brazilian Portuguese stress behavior: final-syllable structure matters, and it matters enough to reorganize the distribution of outputs on novel forms. Read against Wetzels’s (2007) analysis of BP non-verbal stress, this is precisely the productive core one would expect a structured model to learn first: the right edge is decisive, and heavy final rhymes strongly attract stress.

This interpretation fits the architecture described earlier in the paper. Stress is not predicted independently of the rest of the representation. It is computed over syllable-oriented outputs, with access to phoneme-derived information and explicit supervision over syllable count and boundary structure. The model therefore has multiple opportunities during training to connect prominence to syllabification and segmental shape. The discussion point is not simply that these components exist in the code, but that the pseudoword behavior is exactly what one would expect from a system

that has internalized stress as part of a structured pronunciation object. TINA behaves as though orthographic input is first converted into a syllabified and prominence-bearing representation, and only then interpreted segmentally at a more detailed level. That representation, however, remains phonological rather than acoustic: in Brazilian Portuguese lexical-stress perception, duration is an important cue and formant pattern also contributes, while production work across speaking styles identifies duration as the most robust stress correlate, with F0 variability and spectral emphasis providing additional support. Fully matching speaker behavior would therefore require more than assigning an abstract stress mark to one syllable (Barbosa et al., 2013; Fontes, 2013).

A second positive result concerns the hybrid division of labor itself. The neural component produces a relatively abstract output, while the rule layer expands that output into variant realizations through epenthesis, vowel harmony and alternation, affrication, and dialect-profile substitutions. In this respect, TINA aligns with an older hybrid computational line in which data-driven mapping is combined with explicit phonological structure instead of forcing all orthographic–phonological regularities into a single flat representation. For Brazilian Portuguese, this is not merely an implementation convenience. It means that the system distinguishes between a stable structural analysis of the word and a later stage of surface-oriented expansion. That distinction is especially valuable in a language where codas, rhotics, and unstressed vowels are often better treated as positions of controlled variability than as fixed one-to-one segmental facts. In Medeiros’s (2012) gestural account, BP codas are not just symbolic terminal slots but loci of reduced, overlapping, and sometimes vocalized coordination, which is exactly the kind of behavior that makes a post-neural expansion layer more plausible than a single fully concrete output representation. For rhotics in particular, Brazilian Portuguese variation reaches the phonetics–phonology boundary and may be conditioned by prosodic context rather than by a single invariant segmental realization, which makes a separation between structural analysis and later surface expansion especially appropriate (Silva & Albano, 1999). Even when the model does not capture the full phonetics of those domains, it at least preserves the insight that abstract structure and realized form should not always be collapsed into a single output tier.

6.2 Where the model remains too categorical

The same results also show where TINA becomes too rigid. The model captures the main oxytone–paroxytone opposition on pseudowords, but it eliminates the smaller antepenultimate zone entirely. That absence matters. It means that the learned stress system is productive, but narrower than the speaker benchmark, where pseudoword stress remains sensitive not only to broad prosodic regularities but also to lexical analogy, frequency, and final-syllable composition (Benevides, 2024). From the perspective of BP stress theory, this looks like a model that has absorbed the productive side of the system more strongly than its exceptional residue: Wetzels (2007) argues that non-verbal stress is right-edge oriented and weight-sensitive, while antepenultimate patterns occupy a restricted exceptional space and are excluded altogether when the prefinal rhyme is heavy. As discussed before, this is an artifact from the training data itself. In practical terms, TINA behaves as if the major regularities of the system were enough and the weaker residual patterns could be ignored, as it did not have access to non-marked antepenultimate-stressed syllable patterns, due to BP formal orthographic conventions. This is also where the gap between phonological regularity and phonetic implementation becomes relevant: perceptual and acoustic work on BP stress suggests that prominence is recovered from a bundle of cues, especially duration, vowel-quality-related formant structure, and other prosodic correlates that remain robust across styles, rather than from a single binary prominence decision. A model that predicts only symbolic stress positions will therefore tend to preserve the main structural opposition while losing part of the weaker residual space (Barbosa et al., 2013; Fontes, 2013).

A similar sharpening appears in the treatment of final syllables. For open-final pseudowords, TINA still leaves room for a minority of final-stressed outputs, which keeps the distribution from collapsing into a single default. For closed-final pseudowords, however, the system becomes categorical: every item is mapped to final stress. This is stronger than the speaker pattern reported by Benevides (2024), where final CVC syllables strongly favor oxytone outputs but do not eliminate penultimate and residual antepenultimate responses altogether. This is analytically revealing. It suggests that the model has indeed detected the importance of final closure, but it has turned that

cue into a deterministic rule. That strengthening is especially noteworthy in light of analyses that treat BP coda behavior itself as phonetically reduced and gradient rather than as a uniformly stable consonantal endpoint (Medeiros, 2012). Human behavior in the benchmark is much more constrained than free, yet it still preserves penultimate and even occasional antepenultimate responses in those same environments, including the experimentally attested case in which proparoxytonic patterns survive even with final CVC structure when phonological similarity to stored lexical patterns is strong enough (Benevides, 2024). The model therefore learns the main structural pressure while underrepresenting the residual flexibility that remains visible in speaker productions. Here again the relation to Wetzels’s (2007) account is instructive: TINA mirrors the productive attraction of heavy final syllables, but in a strengthened form that suppresses the lexically restricted prefinal-stress residue that BP still allows.

The group comparisons point in the same direction. In the human benchmark, outputs vary with the broader lexical environment of the pseudoword, including factors tied to similarity, lexical frequency, and final-syllable shape (Benevides, 2024). In TINA, by contrast, the same final-syllable structure tends to produce the same stress distribution across groups. This does not mean that the model has no relation to lexical statistics; it was trained on a large lexicon, and some regularities clearly reflect experience with word forms. But in the present design, the system shows little sign of preserving the finer analogical pressures that lead speakers to keep low-frequency or similarity-sensitive alternatives partially active. It is better at broad structural generalization than at retaining a graded memory of marginal patterns. This is precisely where the comparison with Benevides’s pseudoword data becomes most revealing: native speakers do not simply apply a blind default, but extend stored accentual patterns *analogically* when phonological similarity is high, especially when lexical support is frequent enough for those patterns to remain strongly available.

The same limitation extends to surface variation. The rule layer is interpretable and linguistically motivated, but it is still a finite symbolic expansion mechanism. It can generate epenthetic variants, vowel harmony, alternations, and dialect-specific substitutions, yet it does not assign probabilities, track register, or represent gradient realization. This matters most for domains

such as BP rhotics, where variation can be partly categorical, partly continuous, and sensitive to boundary strength rather than only to segment identity (Silva & Albano, 1999). It also matters for codas more generally: Medeiro's (2012) argues that BP codas, especially nasal ones, involve specific patterns of gestural overlap and gradient vocalization at the phonetics–phonology interface, whereas TINA's repairs remain discrete symbolic options. Its dialect handling is profile-based rather than learned as a continuum, and its coda repairs are categorical options rather than graded tendencies. This is a defensible design choice for a transcription system, but it sets a clear limit on what kind of phonetics the model can claim to represent. TINA can enumerate plausible surface realizations; it cannot yet model how frequently, how strongly, or under which speaking conditions speakers realize one option instead of another.

6.3 Why the hybrid design is still the right level of ambition

These limitations do not weaken the main contribution of the system. If anything, they make its real contribution easier to state. TINA is most successful when it is understood as a model of structured pronunciation rather than as a complete model of spoken output. It learns stable relationships among orthography, syllabification, stress, and abstract segmental structure; it then passes that structure to a rule component that makes dialect-sensitive surface expansion explicit. This architecture is modest, but it is also conceptually coherent, and it belongs to a longer computational tradition in which phonological parsing and hierarchical lexical structure are treated as useful intermediates for pronunciation modeling rather than as expendable scaffolding around direct string conversion (Meng et al., 1994). A purely flat end-to-end transcription would hide exactly the distinctions that make Brazilian Portuguese interesting here, especially the interaction among syllable structure, prominence, and right-edge weight in non-verbs (Wetzels, 2007).

The evaluation strategy confirms that this is the right level at which to read the model. Lexical output inspection suggests that the system produces coherent structured transcriptions on ordinary items, but the pseudoword benchmark reveals what is actually distinctive about it. On unfamiliar forms, TINA does not collapse into noise. It continues to produce structured outputs,

and its stress assignments track meaningful aspects of Brazilian Portuguese word shape. That is why pseudowords are so informative in this paper. They expose not only whether the model can transcribe, but what kind of generalization it uses when memory alone is no longer enough. In this respect, the benchmark is especially valuable because Brazilian Portuguese pseudoword reading has been shown to probe exactly the interaction among lexical analogy, token frequency, and final-syllable structure that a structured pronunciation model ought either to capture or to miss in interpretable ways (Benevides, 2024). In that respect, the pseudoword results are not a side experiment. They are the clearest evidence that the system has internalized a broad pronunciation grammar.

At the same time, the discussion should resist turning this into a stronger claim than the evidence supports. TINA does not model the acoustic implementation of stress, even though BP lexical-stress perception depends importantly on duration and also on formant pattern, and acoustic studies further show duration, F0 variability, and spectral emphasis to be reliable correlates across speaking styles (Barbosa et al., 2013; Fontes, 2013). Nor does it learn a gradient sociophonetic space of rhotic or coda realization, even though BP rhotics themselves show that productive variation may sit exactly at the phonetics–phonology boundary and be prosodically conditioned, and BP codas more generally have been analyzed as gradient coordination patterns rather than purely categorical segmental slots (Medeiros, 2012; Silva & Albano, 1999). It does not preserve all of the low-probability alternatives that speakers still produce on novel forms. Its success is therefore selective. The system captures broad structural regularities, but it simplifies the phonetics–phonology interface at exactly the points where speaker behavior remains variable, analogical, or gradient.

That selective success is still scientifically useful. It shows that a relatively compact hybrid model can encode a meaningful portion of Brazilian Portuguese pronunciation knowledge without reducing the task to either pure symbolic rules or an opaque flat neural mapping. The model’s strengths are interpretable: stable syllable parsing, strong control over stress placement, robust output formation on unseen items, and a transparent mechanism for expanding underspecified representations into dialect-oriented variants. Its weaknesses are interpretable as well: overregularization, categorical

handling of cues that remain probabilistic for speakers, and limited access to gradient phonetic realization. For the purposes of this paper, that is precisely the right outcome. TINA is most valuable not as a final account of Brazilian Portuguese speech, but as a workable and linguistically legible model of Brazilian Portuguese pronunciation structure.

7 Conclusion

This paper has argued that automated phonetic transcription for Brazilian Portuguese is better treated as a structured pronunciation problem than as a flat grapheme-to-phone mapping. TINA was designed accordingly. Instead of predicting only a single undifferentiated output string, it combines a neural component that learns syllabification, stress, and abstract segmental structure with a rule layer that expands those outputs into surface-oriented variants. That architecture is deliberately modest. Its purpose is not to present a radically new neural method, but to show that a compact hybrid system can represent linguistically meaningful regularities of Brazilian Portuguese in a way that remains usable both computationally and analytically.

The empirical picture supports that claim. On ordinary lexical items, TINA produces well-formed structured outputs. More importantly, on pseudowords it continues to produce well-formed outputs instead of collapsing when lexical memory alone is no longer sufficient. The model preserves stable syllabic organization, assigns stress in a way that tracks broad Brazilian Portuguese right-edge regularities, and responds strongly to final-syllable structure. Those are not trivial properties. They indicate that the system has learned a productive core of pronunciation knowledge rather than only a list of stored correspondences. In that sense, the pseudoword results are not an auxiliary evaluation added after the fact. They are the clearest evidence that the model internalizes a structured generalization about Brazilian Portuguese word form.

At the same time, the same benchmark makes the model’s limits unmistakable. TINA regularizes too strongly where speaker behavior remains gradient. It captures the major opposition between final and penultimate stress, but it suppresses the smaller antepenultimate domain. It

tracks the importance of final closure, but it tends to convert that cue into a categorical decision. It generates interpretable surface variants, but those variants remain symbolic possibilities rather than probabilistic realizations. The comparison with native-speaker pseudoword behavior is valuable precisely because it reveals these boundaries clearly: the model learns broad structural pressures better than it preserves marginal patterns, analogical effects, and low-probability alternatives.

That boundary should be read as a result, not as a defect to be hidden. TINA is most convincing when it is understood as a hybrid model of Brazilian Portuguese pronunciation structure. It captures a meaningful middle layer between spelling and spoken realization: not just letters, and not yet full phonetic detail, but syllabified, stress-bearing, archiphonemically organized representations that can later be expanded into controlled surface forms. For Brazilian Portuguese, that level of analysis is scientifically useful in its own right, because it keeps visible the interaction among syllable structure, stress, codas, and dialect-sensitive realization instead of flattening them into a single opaque transduction problem.

The broader methodological point of the paper follows from this. Pseudowords provide a principled way to evaluate whether a computational pronunciation model is actually behaving like speakers on unseen forms, rather than merely reproducing familiar lexical items. They therefore do more than add another benchmark: they help distinguish memorization from structured generalization, and they make overregularization visible when it occurs. In the case of TINA, that test shows a system that is useful, linguistically legible, and productively constrained, even though it is not a full model of gradient spoken variation. The model is available as a working transcription tool, but the main scientific contribution is more specific: a compact hybrid architecture can go a long way in Brazilian Portuguese when the target representation is structured, and pseudoword evaluation is one of the clearest ways to show exactly how far that architecture really goes.

Limitations and future work

TINA remains a modest system, and several of its current limits also point directly to promising next steps. First, its outputs still simplify aspects of Brazilian Portuguese that are gradient,

variable, and socially or regionally conditioned. Second, the present evaluation, while informative, is still based on a single pseudoword dataset. These limits do not undermine the main result, but they suggest a clear path forward.

Future work can therefore expand TINA in two complementary directions: broader empirical coverage and richer linguistic modeling. On the empirical side, larger lexica (including non-marked proparoxytones and non-standard orthographic variants), additional pseudoword datasets, and more explicit speaker-based comparisons would help test the robustness of its generalizations. On the modeling side, future versions can move beyond categorical outputs toward more fine-grained treatment of variation, uncertainty, and dialect-sensitive realization. Even in its current form, however, TINA already shows that a compact hybrid system can serve as a useful model of Brazilian Portuguese pronunciation, a useful tool for Brazilian Portuguese scholars and as a productive basis for further work.

Data accessibility statement

The complete results, corpora, and analytical routines supporting this article are available in an open repository. The repository includes all files needed for full reproducibility, including the trained models/weights used in the analysis. Repository link: <https://github.com/alexandre-barroso/TINA>. Persistent identifier (DOI) of the repository: [10.5281/zenodo.19509618](https://doi.org/10.5281/zenodo.19509618). Web app available at: <https://tina.alexandrebarroso.com>.

Ethics and consent

This article reports analyses of previously published datasets and does not involve new data collection from human participants. Ethical review and participant-consent procedures for the original data collections are described in the source studies cited in the manuscript.

Competing interests

The author has no competing interests to declare.

Authors' contributions

The author is solely responsible for the conceptualization, methodology, formal analysis, software, visualization, and writing of this article.

References

- Alencar, M. S. M. d. (2007). *Aspectos sócio-dialetais da língua falada em fortaleza* [Tese (Doutorado em Linguística)]. Universidade Federal do Ceará. https://repositorio.ufc.br/bitstream/riufc/6109/1/2007_tese_smalencar.pdf
- Barbosa, P. A., Eriksson, A., & Åkesson, J. (2013). On the Robustness of some Acoustic Parameters for Signalling Word Stress across Styles in Brazilian Portuguese. *Interspeech 2013*, 282–286. <https://doi.org/10.21437/Interspeech.2013-84>
- Benevides, A. d. L. (2022). *O acento em pseudopalavras* [Thesis]. Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. <https://doi.org/10.11606/T.8.2022.tde-31012023-180157>
- Benevides, A. d. L. (2024). O processamento lexical do acento em pseudopalavras a partir da leitura de falantes do português brasileiro. *Filol. Linguíst. Port.*, 26(2), 241–259. <https://doi.org/10.11606/issn.2176-9419.v26i2p241-259>
- Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. *Advances in Neural Information Processing Systems* 28, 1171–1179.
- Callou, D. (2009). Um perfil da fala carioca. In S. S. C. Ribeiro, S. B. B. Costa, & S. A. M. Cardoso (Eds.), *Dos sons às palavras: Nas trilhas da língua portuguesa* (pp. 130–152). EDUFBA. <https://books.scielo.org/id/vr2dr/pdf/ribeiro-9788523211851-08.pdf>

- Callou, D., Serra, C., & Cunha, C. (2015). Mudança em curso no português brasileiro: O apagamento do r no dialeto nordestino. *Revista da ABRALIN*, 14(1), 195–219. <https://doi.org/10.5380/rabl.v14i1.42491>
- Cardoso, S. A. M. (1986). Tinha nascentes razão? considerações sobre a divisão dialetal do Brasil. *Estudos Linguísticos e Literários*, (5), 47–59.
- Caruana, R. (1997). Multitask Learning. *Machine Learning*, 28, 41–75. <https://doi.org/10.1023/A:1007379606734>
- Fontes, M. A. S. (2013). Pistas Acústicas e a Percepção do acento lexical em Português Brasileiro. *Revista Intercâmbio*, 27, 100–109.
- Lin, B., Wang, L., Feng, X., & Zhang, J. (2020). Joint detection of sentence stress and phrase boundary for prosody. *Interspeech 2020*, 4392–4396. <https://doi.org/10.21437/Interspeech.2020-1284>
- Lo, R. Y.-H., & Nicolai, G. (2021). Linguistic Knowledge in Multilingual Grapheme-to-Phoneme Conversion. *Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 131–140.
- Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421.
- Medeiros, B. R. d. (2012). Uma proposta sobre a coda do português brasileiro a partir da fonologia gestual, com foco especial na nasal. *Revista da ABRALIN*, 11(1), 89–138.
- Meng, H. M., Seneff, S., & Zue, V. W. (1994). Phonological Parsing for Bi-directional Letter-to-Sound/Sound-to-Letter Generation. *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. <https://doi.org/10.3115/1075812.1075876>
- Rao, K., Peng, F., Sak, H., & Beaufays, F. (2015). Grapheme-to-Phoneme Conversion Using Long Short-Term Memory Recurrent Neural Networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4225–4229. <https://doi.org/10.1109/ICASSP.2015.7178767>

- Ronanki, S., Watts, O., & King, S. (2017). A Hierarchical Encoder-Decoder Model for Statistical Parametric Speech Synthesis. *Interspeech 2017*, 1133–1137. <https://doi.org/10.21437/Interspeech.2017-628>
- Silva, A. H. P., & Albano, E. (1999). Brazilian Portuguese Rhotics and the Phonetics/Phonology Boundary. *Proceedings of the 14th International Congress of Phonetic Sciences*, 2211–2214.
- Toshniwal, S., & Livescu, K. (2016). Jointly Learning to Align and Convert Graphemes to Phonemes with Neural Attention Models. *2016 IEEE Spoken Language Technology Workshop (SLT)*, 76–82.
- van Esch, D., Chua, M., & Rao, K. (2016). Predicting Pronunciations with Syllabification and Stress with Recurrent Neural Networks. *Interspeech 2016*, 2841–2845. <https://doi.org/10.21437/Interspeech.2016-1419>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems* 30, 5998–6008.
- Wetzels, W. L. (2007). Primary Word Stress in Brazilian Portuguese and the Weight Parameter. *Journal of Portuguese Linguistics*, 5/6, 9–58.
- Yao, K., & Zweig, G. (2015). Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion. *Interspeech 2015*, 3330–3334. <https://doi.org/10.21437/Interspeech.2015-134>
- Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2019). Transformer based Grapheme-to-Phoneme Conversion. *Interspeech 2019*, 2095–2099. <https://doi.org/10.21437/Interspeech.2019-1954>