Fast, Not Fancy: Rethinking G2P with Rich Data and Statistical Models

Mahta Fetrat and Zahra Dehghanian and Hamid R. Rabiee
Dept. of Computer Engineering
Sharif University of Technology / Tehran
m.fetrat@sharif.edu, zahra.dehghanian97@sharif.edu, rabiee@sharif.edu

Abstract

Homograph disambiguation remains a significant challenge in grapheme-to-phoneme (G2P) conversion, especially for low-resource languages. This challenge is twofold: (1) creating balanced and comprehensive homograph datasets is labor-intensive and costly, and (2) specific disambiguation strategies introduce additional latency, making them unsuitable for real-time applications such as screen readers and other accessibility tools. In this paper, we address both issues. First, we propose a semi-automated pipeline for constructing homograph-focused datasets, introduce the HomoRich dataset generated through this pipeline, and demonstrate its effectiveness by applying it to enhance a state-of-the-art deep learning-based G2P system for Persian. Second, we advocate for a paradigm shift—utilizing rich offline datasets to inform the development of fast, statistical methods suitable for latency-sensitive accessibility applications like screen readers. To this end, we improve one of the most well-known rule-based G2P systems, eSpeak, into a fast homograph-aware version, HomoFast eSpeak. Our results show an approximate 30 percentage-point improvement in homograph disambiguation accuracy for the deep learning-based and eSpeak systems.

1 Introduction

Grapheme-to-phoneme (G2P) conversion is a crucial step in many fast text-to-speech (TTS) models (Ren et al., 2020). It refers to the task of converting a given written text into its corresponding sequence of phonemes—how it is pronounced. There are several formats for representing phoneme sequences, one of the most widely used being the International Phonetic Alphabet (IPA) (Association, 1999). As an example, the phoneme sequence for the sentence *"I will read it"* is /aɪ wɪl riːd ɪt/ in IPA format.

The complexity of G2P conversion varies by language. Some languages like Turkish and Spanish are highly phonetic, meaning a near one-to-one correspondence between spelling and pronunciation (Koşaner et al., 2013; Delattre, 1945). In contrast, in many other languages, such as Persian, G2P is more complex due to exceptions and rules that depend on context (Qharabagh et al., 2025a).

One such challenge is handling homographs—words spelled the same but pronounced differently depending on context. For example, the word *"read"* is pronounced /rɛd/ in the past tense (*"I read this book yesterday"*) and /riːd/ in the present tense (*"I read the book every night"*).

Unfortunately, sentence-level G2P datasets are extremely scarce in low-resource languages. This scarcity stems from the fact that phonemization is a time-consuming and costly process that requires expert annotators. Homograph-specific datasets are even rarer, as they depend on source corpora that must meet strict conditions: they should contain a wide range of homographs and provide a balanced number of examples for each pronunciation. Without this balance, the resulting models will fail to learn some homographs and tend to default to the more frequent pronunciation in ambiguous cases.

Beyond data scarcity, there is also a methodological challenge in G2P conversion. The two primary approaches are neural models (Ploujnikov, 2024; Řezáčková et al., 2024b; Gao, 2024) and non-neural methods (Silva et al., 2012; Alayiaboozar et al., 2019; Riahi and Sedghi, 2012). While neural methods have gained popularity due to their flexibility and learning capacity, they often suffer from high inference latency, making them unsuitable for real-time applications such as screen readers that serve accessibility needs. This motivates a renewed focus on non-neural approaches, aiming to improve their accuracy while preserving their inherent speed.

This work proposes a practical approach for generating a rich and balanced homograph dataset. We demonstrate that such a dataset not only boosts

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22382–22408
November 4-9, 2025

the homograph disambiguation accuracy of neural G2P models but also significantly enhances the performance of rule-based systems. Specifically, we show that by incorporating a simple, fast statistical method that leverages the proposed dataset, rule-based models can be equipped with context understanding, leading to improved handling of homographs without sacrificing speed.

Our key contributions are as follows:

- We propose a practical and cost-efficient recipe for constructing rich and balanced homograph datasets in low-resource languages by leveraging LLMs for G2P annotation and homograph sample generation.
- We release HomoRich, the first and largest Persian homograph dataset, and demonstrate its effectiveness by improving the homograph disambiguation accuracy of a state-of-the-art neural G2P model by 29.72 percentage-points.
- We introduce a lightweight statistical method that enhances G2P systems for homograph disambiguation, using datasets generated by our proposed approach.
- We integrate this method into the open-source eSpeak engine, resulting in HomoFast eSpeak, a variant that achieves a 30.66 percentage-point improvement in homograph disambiguation without compromising real-time performance.

2 Related Works

In this section, we review homograph disambiguation from two perspectives: the common methods and the datasets developed to address challenges in low-resource settings like Persian.

2.1 Approaches

There are multiple approaches to addressing the homograph challenge, including neural and rule-based methods, various machine learning algorithms, hybrid techniques, and the use of large language models (LLMs). We briefly highlight only the works most relevant to our approach. A more comprehensive review is provided in Appendix A.

Rule-based approaches have been widely explored for homograph disambiguation across various languages. These methods often rely on morphosyntactic patterns, lexical cues, and contextual heuristics rather than deep semantic inference. For instance, Silva et al. (2012) and Alayiaboozar et al. (2019) utilized hand-crafted linguistic rules derived from syntactic and morphological features in Brazilian Portuguese and Persian, respectively. Hearst (1991) introduced a system based on shallow syntactic patterns and lexical co-occurrences in local contexts, while Yarowsky (1997) developed data-driven decision lists using log-likelihood-ranked contextual patterns. Riahi and Sedghi (2012) further extended these ideas by integrating rule-based decision lists into a tri-training framework.

Neural approaches have been widely adopted for homograph disambiguation and G2P conversion across languages, leveraging contextual embeddings, sequence modeling, and attention mechanisms. Early work in this area applied sequence-to-sequence neural network models, such as LSTMs, to the G2P task, demonstrating performance comparable to or surpassing previous n-gram and maximum entropy models (Yao and Zweig, 2015). Specifically, Yao and Zweig (2015) explored using encoder-decoder LSTMs and found that bi-directional LSTMs that utilize alignment information significantly advanced the state-of-the-art for monolingual G2P conversion. Additionally, Peters et al. (2017) introduced a massively multilingual neural approach that used a single shared encoder-decoder across hundreds of languages, leveraging a language ID token to manage different spelling-pronunciation patterns. Nicolis and Klimkov (2021) and Seale (2021) utilized pre-trained language models like BERT, ALBERT, and XLNet to extract contextual word embeddings and fine-tune token classifiers or logistic regressors for English homographs. SoundChoice, proposed by Ploujnikov (2024), employed a hybrid RNN-attention model with BERT embeddings and curriculum learning to predict phonemes in context. Similarly, Nanni (2023) adapted SoundChoice for Italian, integrating ChatGPT-generated data. Řezáčková et al. (2024a,b) adopted the T5 transformer for multilingual G2P, bypassing rule-based post-processing by modeling cross-word effects. Comini et al. (2025) combined GRUs, transformers, and knowledge distillation for efficient G2P in low-resource settings. Gao (2024) enhanced multilingual phonetic recognition using self-supervised learning models (e.g., wav2vec2, HuBERT) and

Title	Sample Type	Sample Count	Hom. Curated	Hom. Count	Availability	License
Semi-sup H.D. (Riahi and Sedghi, 2012)	Sent.	–	Yes	2	Not avail.	N.A.
AvashoG2P (Moghadaszadeh et al., 2024)	Sent.	–	Yes	54	Not avail.	N.A.
Multi-Module G2P (Rezaei et al., 2022)	Word	12,000	No	–	Available	N.A.
GE2PE (Rahmati and Sameti, 2024)	Sent.	42,540	No	–	Not avail.	N.A.
HomoRich (Ours)	Sent.	5,376,670	No	–	Available	MIT
	Sent.	**528,891**	**Yes**	**285**	**Available**	**CC0-1.0**

Table 1: Persian Homograph Datasets. *Hom. Count* shows the number of homographs covered in the dataset and *Hom. Curated* indicates if homograph samples were deliberately inserted or naturally occurring in a regular corpus.

synthetic data.

LLM-based approaches are increasingly demonstrating the potential of LLMs in G2P conversion. Suvarna et al. (2024) were the first to benchmark models like GPT-4 and Claude-3 on phonological tasks, including G2P, and found that while promising, they still lag behind traditional models in accuracy. Han et al. (2024) leveraged GPT-4's in-context retrieval to map homographs to dictionary pronunciations, combining automated generation with manual refinement for accuracy. Similarly, Qharabagh et al. (2025a) applied LLMs to Persian G2P conversion through advanced prompting, achieving state-of-the-art results on custom datasets without model fine-tuning.

2.2 Datasets

Several studies have proposed various methods to address data scarcity in G2P for low-resource languages such as Persian. A comprehensive review of these studies is provided in Appendix A.2; however, here we summarize only the most relevant features of the datasets in Table 1. As shown, all of the referenced datasets are either not homograph-specific, not sentence-level, or not publicly available. This highlights a critical gap in homograph data for Persian—and likely for many other low-resource languages—which has resulted in the lack of G2P systems that outperform random chance in homograph disambiguation.

3 Methodology

Developing an effective G2P model requires both high-quality data and the tools to make use of it. This section outlines our data generation process and how we leveraged it to improve G2P models.

3.1 Data Preparation

The scarcity of homograph data arises from two main challenges. First, assembling a high-quality text corpus that provides broad and balanced coverage of homographs across diverse contexts is difficult. Second, phonemizing a text corpus is both time-consuming and costly, as it requires trained experts with linguistic knowledge. In this paper, we present a practical approach for collecting such data in a low-resource language like Persian and demonstrate its effectiveness in the next section.

To tackle the first challenge, we started with KaamelDict (Fetrat, 2024a), the most extensive Persian G2P dictionary introduced in Qharabagh et al., 2025a. We filtered for words with multiple valid pronunciations to identify potential homographs. Then, through manual review, we excluded words that either (1) had multiple commonly accepted pronunciations needing no disambiguation, or (2) included archaic, poetic, or rarely used forms. From this, we selected a list of 285 homograph words that were both comprehensive and practically relevant.

The next task was to generate a diverse and balanced set of sentences for each homograph, covering different usage contexts and ensuring equal representation of all pronunciations. To automate this, we experimented with prompting LLMs to generate sentences for each pronunciation or meaning. However, the results were often skewed toward the dominant pronunciation, even with explicit instructions. We found that embedding the homograph in a full sentence that implied its intended meaning significantly improved accuracy.

As a result, we adopted a hybrid approach, com-

bining manual and LLM-generated sentences. We first shared a list of selected homographs with about 200 native speakers[1], asking each to write five contextually varied sentences for every pronunciation. We then used some of these human-written examples as few-shot prompts to guide LLM-based sentence generation (see Figure 1).

To further enhance the dataset and support downstream TTS and G2P tasks, we integrated sentences from three widely used Persian corpora: ManaTTS (Qharabagh et al., 2025b), GPTInformal (Fetrat, 2025), and CommonVoice (Ardila et al., 2019). These additions were meant to improve overall G2P accuracy—particularly phoneme error rate (PER)—and enrich the corpus with phoneme-annotated examples from diverse registers.

To address the second challenge—phonemization—we leveraged prior work on LLM-powered G2P conversion (Qharabagh et al., 2025a). In that study, it was demonstrated that LLMs can assist in labeling graphemes with their phonemes, thanks to their phonetic knowledge and contextual understanding, which is particularly helpful in disambiguating homographs. The study introduced several techniques to enhance LLM performance in G2P tasks without requiring any training, benchmarking state-of-the-art models to guide future dataset generation.

We use the most effective method from that study to phonemize our corpus. It prompts the model with Finglish—a more accessible but slightly ambiguous phonemic representation of Persian—instead of the less common IPA format. The method combines in-context learning, few-shot examples, hints from a G2P dictionary, and a final mapping step to produce the target phoneme format (see Figure 2). To balance cost, availability, and quality, we use GPT-4o (Hurst et al., 2024) as the LLM, which achieved a Phoneme Error Rate (PER) of 6.43% and a homograph disambiguation accuracy of 64%, outperforming many existing Persian G2P systems (see Section 4 for details).

Figure 3 summarizes the structure of the generated dataset. For compatibility with previous work, we mapped the phonemes of all sentences to an alternative phoneme format (see Appendix B). We release our dataset, named HomoRich, under a permissive CC0-1 license, making it freely available for both academic and commercial use.[2]

[1]Details about these human annotators can be found in Appendix F.

[2]The HomoRich dataset is available at https://huggingface.co/datasets/MahtaFetrat/HomoRich-G2P-Persian.

3.1.1 Data Statistics

The HomoRich dataset, generated using our proposed recipe, contains 528,891 annotated Persian sentences. As mentioned, it consists of both homograph-focused and general-purpose G2P data collected from multiple sources. Figure 4 illustrates the composition of the dataset. The exact count of samples from each source is also available in Table 4 in the Appendix.

To ensure diversity, both human annotators and language models were instructed to generate data across a wide range of contexts. The dataset comprises 75,715 unique words, and the distribution of sentence lengths is shown in Appendix Figure 10. To further assess diversity, we calculated three metrics for the corpus: 1) the average cosine similarity of ParsBERT (Farahani et al., 2020) embeddings for sentences of each homograph, 2) the average unique word count for the samples of each homograph, and 3) the average unique sentence ratio (USR) for the dataset. The results can be found in Table 2.

The HomoRich dataset includes 285 homograph words, each associated with multiple pronunciations: 257 have two variants, 21 have three, and 7 have four. On average, each homograph appears in over 1,000 distinct sentence contexts. To avoid bias toward more frequent pronunciations, we maintained a balanced number of samples for each variant. Figure 5 shows the pronunciation distribution, confirming the dataset's high balance.

To evaluate the overall correctness of the HomoRich dataset, we manually reviewed a random sample of 1,219 instances from the homograph-specific subset—arguably the most error-prone section—ensuring coverage of all homograph words and pronunciations. The review process identified and filtered out samples that misused the required pronunciation. The resulting accuracy of 91.38% demonstrates a high level of reliability, which we deem acceptable for our purposes.

3.1.2 Data Augmentation

To further address data scarcity—particularly in homograph disambiguation—we proposed three augmentation methods (Figure 6) aimed at enhancing the model's understanding of context and increasing data diversity.

1. Synonym Replacement (Figure 6a): We iden-

System
You are a helpful assistant skilled in generating sentences in Persian, especially for words with multiple meanings and pronunciations. Your responses should demonstrate appropriate use of homographs and provide context for each pronunciation.

User
Your task is to generate sentences for a Persian homograph word that has multiple pronunciations and meanings.
The word is "مرد" with **pronunciations:** */mard/, /mord/*.
Here are some **sample sentences** for the pronunciation */mord/*:

- او در هشتاد سالگی مرد
- تصادف کرد و مرد

Please generate **50 additional** sentences with "مرد" with this pronunciation. Note that these are only sample sentences; you must generate **different and diverse sentences** with this meaning.
Feel free to append prefixes/suffixes to the word as long as you include the word itself.
Please return each output sentence **within brackets**.

Figure 1: Prompt for generating homograph sentences.

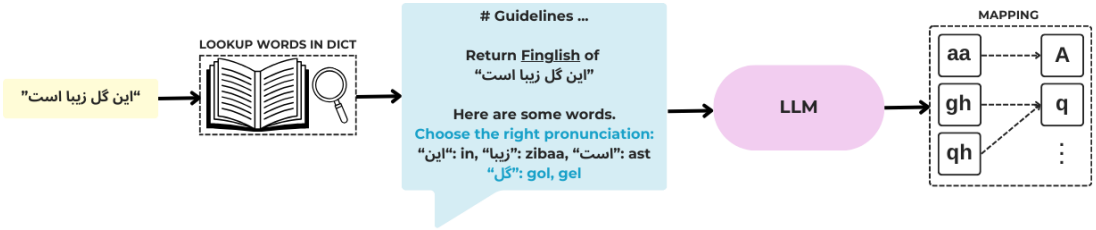


Figure 2: LLM-powered G2P workflow (Qharabagh et al., 2025a)

Source	human
Source Index	1
Grapheme	شاید مرد خوبی است
Homograph Grapheme	مرد
Phoneme	SAyad mard-e xubi ?ast
Homograph Phoneme	mard
Alternative Phoneme	$ay/d m/rde1 xubi @/st
Alternative Homograph Phoneme	m/rd

Figure 3: Dataset structure with example entry.

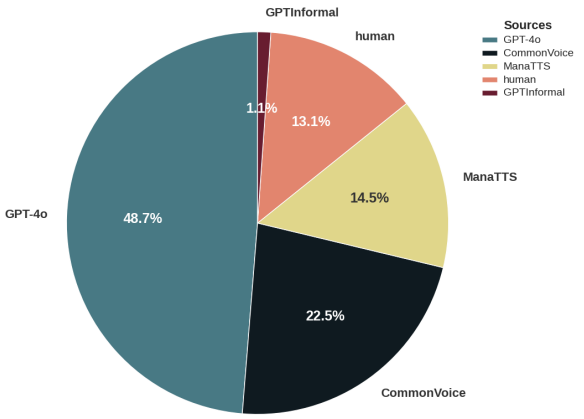


Figure 4: Data source distribution in HomoRich dataset.

tified the most frequently occurring words in the dataset and mapped each to a set of synonyms with equivalent meaning. For each sentence, we replaced these words with their alternatives to generate new samples.

2. Sentence Reordering (Figure 6b): In most cases, the order of context words does not affect the pronunciation of the homograph. Thus, we split sentences at random words and swapped the resulting segments, updating their corresponding phoneme sequences. However, in Persian and similar languages like Arabic, Ezafe (a phoneme that connects grammatically related words) must be preserved. We employed a POS tagger (Group, 2023) to detect Ezafe constructions and ensured no splits occurred within them.

3. Homograph-focused Concatenation (Figure 6c): we further augmented homograph samples by appending randomly selected short sentences (without homographs) to the homograph samples.

Using combinations of these methods, we were able to scale the dataset by up to 10x, depending on the augmentation configuration.

3.2 Proposed G2P Tools

Having generated a large, rich, and balanced homograph dataset using the proposed method, we introduce both neural and non-neural G2P tools

Avg. Cosine Similarity ↓	Avg. Unique Word Count ↑	Avg. Unique Sentence Ratio (USR) ↑
0.5984 ± 0.0407	1441.8090 ± 307.4205	0.9837 ± 0.0124

Table 2: Diversity evaluation of the HomoRich dataset.

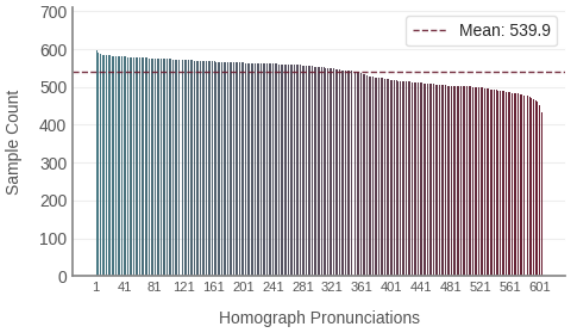


Figure 5: Sample counts per pronunciation.

that build upon this data and demonstrate how this dataset can be used to enhance homograph disambiguation in each approach.

3.2.1 Homo-GE2PE (Neural)

As reviewed in Section 2, ByT5 (Xue et al., 2022) has been successfully fine-tuned for G2P tasks in multiple studies (Řezáčková et al., 2024b,a; Rahmati and Sameti, 2024). In a recent study (Rahmati and Sameti, 2024), this approach resulted in GE2PE, a model achieving state-of-the-art performance in Persian G2P. We further fine-tuned GE2PE on our dataset using a three-phase process:

1. Initial fine-tuning on the regular G2P subset
2. Second-phase fine-tuning on LLM-generated homograph sentences
3. Final fine-tuning on high-quality, human-authored homograph sentences

We used a learning rate of 5e-4 and a batch size of 32 across all phases, with 5, 20, and 50 training epochs respectively, trained on an NVIDIA GTX TITAN X (12GB VRAM, CUDA 12.2) with Intel i7-5820K CPU. The full training process took approximately 24 hours in total. The learning curves for all phases, including training and validation metrics, are shown in Figure 7. The resulting enhanced model, named Homo-GE2PE, is publicly available under an open license.[3]

[3]Complete training scripts, model files and usage instructions are available at https://github.com/MahtaFetrat/Homo-GE2PE-Persian.

3.2.2 HomoFast eSpeak (Non-neural)

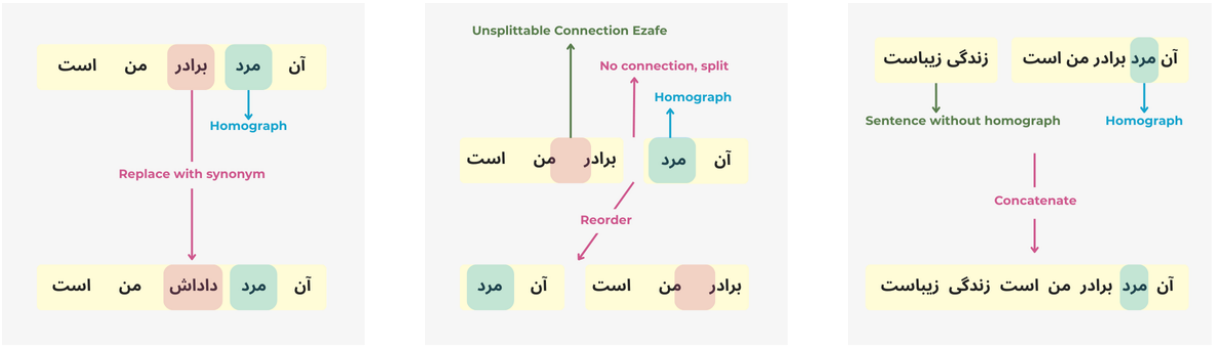
As discussed earlier, one of the main motivations for favoring non-neural methods in certain applications is their low latency. Neural models, while powerful, often incur high inference times, making them less suitable for real-time systems such as screen readers. In contrast, rule-based and statistical systems are extremely fast and lightweight, enabling them to operate effectively in low-latency environments. Therefore, despite the advances in neural G2P systems, it remains important to continue exploring and enhancing non-neural approaches, particularly when speed and responsiveness are critical.

However, a key limitation of non-neural systems is their difficulty in disambiguating homographs, due to their limited or nonexistent semantic or contextual understanding. In this work, we introduce a strategy to enhance the homograph disambiguation ability of G2P systems using datasets generated by our proposed approach. This strategy is purely statistical and does not rely on neural models or even embeddings, making it a perfect solution for improving the homograph accuracy of rule-based methods without compromising their key advantage—speed and low latency. While straightforward in design, this approach has not been explored in prior homograph disambiguation research.

The approach begins by tokenizing the sentences in our dataset, removing stopwords, and constructing a database that maps different pronunciations of homographs to lists of context words that frequently co-occur with each pronunciation.

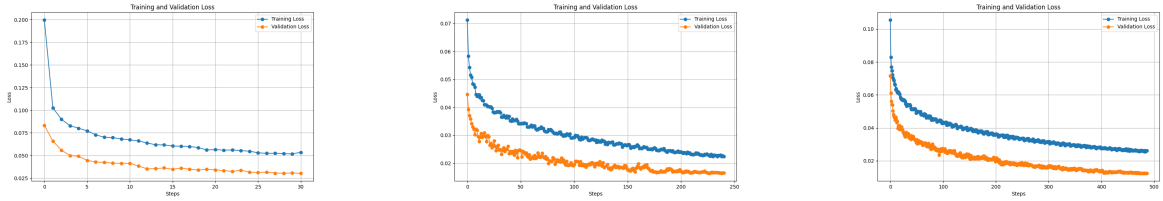
For a new sentence, we compute a weighted overlap between its context words and each pronunciation's context list to derive a similarity score. To mitigate bias toward longer lists, we normalize each score by the length of the corresponding context list. The pronunciation with the highest normalized score is then selected as the most contextually appropriate. For a schematic overview of this method, see Figure 8.

We applied this approach to the widely used eSpeak NG project (Duddington, 2024), selected for its relevance to real-world applications. eSpeak NG



(a) Synonym Replacement (b) Reordering (c) Concatenation

Figure 6: Illustration of our three data augmentation methods for homograph disambiguation.



(a) Phase 1 (5 epochs) (b) Phase 2 (20 epochs) (c) Phase 3 (50 epochs)

Figure 7: Learning curves across fine-tuning phases.

is a compact, open-source text-to-speech synthesizer available on Linux, Windows, Android, and other platforms. It supports over 100 languages and accents, benefiting from contributions by various linguistic communities. Notably, it has an add-on in the open-source NVDA screen reader (nva, 2007), and its Persian G2P module is extensively used in screen readers by a large portion of the blind community in Iran (NV Access Limited and contributors, 2023; Gooshkon, 2022). We name the enhanced version HomoFast eSpeak, which, as shown in the following sections, demonstrated outstanding results, indicating a viable path for enhancing rule-based TTS systems in Persian.[4]

4 Results

The first public sentence-level dataset for benchmarking homograph accuracy in G2P systems, termed SentenceBench, was introduced by Fetrat (2024b). We adopted this dataset as our primary benchmark. This constitutes a stronger methodological choice than using a test split from our own HomoRich dataset, as the latter is predominantly generated by a single LLM and may therefore contain inherent biases toward specific contextual patterns.

[4]The HomoFast eSpeak is available at https://github.com/MahtaFetrat/HomoFast-eSpeak-Persian.

Evaluation of Baseline G2P Tools: Table 3 presents the performance of previously available G2P tools on the SentenceBench benchmark. As shown, the only two models that perform well in terms of phoneme error rate (PER) are the neural GE2PE model (Rahmati and Sameti, 2024) and the rule-based eSpeak tool (Duddington, 2024). However, even these models perform worse than random when it comes to homograph disambiguation.

Evaluation of the Proposed Improved G2P Tools: To address the challenge of homograph disambiguation in Persian G2P systems, we utilized a curated homograph dataset to enhance both neural and rule-based models. Specifically, we fine-tuned the GE2PE (Rahmati and Sameti, 2024) model and proposed a statistical disambiguation module integrated into eSpeak (Duddington, 2024), resulting in two improved variants: Homo-GE2PE and HomoFast eSpeak.

As depicted in Table 3, our improved GE2PE model achieves a 29.72 percentage-point increase in homograph accuracy with a concurrent reduction in phoneme error rate (PER). Notably, our statistical disambiguation module—devoid of any neural components or learned embeddings—delivers the same level of homograph accuracy improvement when integrated into rule-based models, all while maintaining their inference speed. This un-

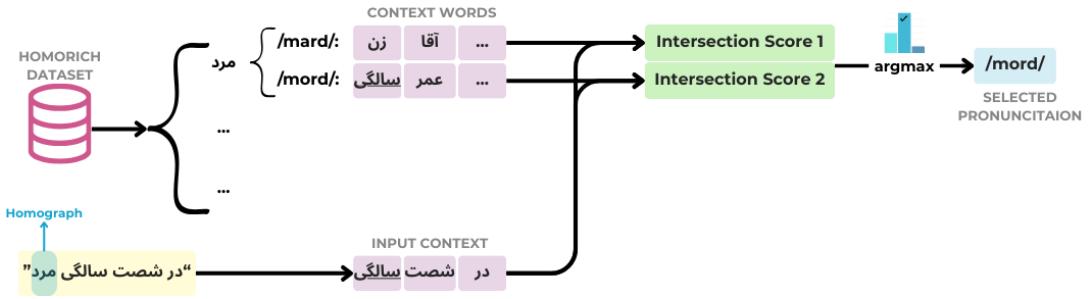


Figure 8: Overview of the proposed statistical homograph disambiguation approach.

Model	PER (%) ↓	Homograph Acc. (%) ↑	Avg. Inf. Time (s) ↓
persian-phonemizer (Dehghani, 2022)	25.27 ± 0.09	29.25 ± 0.47	0.1803 ± 0.04
PersianG2P (Pascal, 2020)	15.04 ± 0.00	37.74 ± 0.00	2.1686 ± 0.10
Persian_G2P (Rabiee, 2019)	35.23 ± 0.00	21.23 ± 0.00	11.1374 ± 0.56
G2P (Ajini, 2022)	19.63 ± 1.83	29.91 ± 0.72	28.0039 ± 0.42
G2P with Transformer (Alipour, 2023)	12.85 ± 0.09	40.00 ± 0.21	0.9685 ± 0.03
Epitran (Mortensen et al., 2018)	45.12 ± 0.00	0.00 ± 0.00	**0.0003 ± 0.00**
eSpeak (Duddington, 2024)	6.92 ± 0.00	43.87 ± 0.00	0.0169 ± 0.00
GE2PE (Rahmati and Sameti, 2024)	4.81 ± 0.00	47.17 ± 0.00	0.4464 ± 0.03
Homo-ByT5	4.12 ± 0.13	76.32 ± 0.52	0.4141 ± 0.09
HomoFast eSpeak	6.33 ± 0.00	74.53 ± 0.00	0.0084 ± 0.00
Homo-GE2PE	**3.98 ± 0.00**	**76.89 ± 0.00**	0.4473 ± 0.02

Table 3: Comparison of Persian G2P tools in terms of Phoneme Error Rate (PER), Homograph Accuracy, and Average Inference Time. Results are reported as mean ± standard deviation across 5 independent runs. Best results are in **bold**, and second-best are underlined.

derscores the value of high-quality data and shows that even simple statistical techniques can be highly effective when supported by strong datasets.

Also, to provide a comprehensive evaluation of our neural tool and to isolate the effect of training data exposure, we performed an 80-20 stratified train-test split on the homograph subset of the HomoRich dataset. This split ensures that approximately 80% of the instances for each pronunciation of every homograph are allocated to the training set, with the remaining 20% reserved for testing. We trained the base GE2PE model on this training partition using the same settings and scripts as in our primary experiments. Evaluation on the held-out test set yielded a PER of 5.36% and a homograph accuracy of 87.64%. This partitioned dataset is available in the aforementioned HomoRich repository.

Fine-tuning ByT5 on Our Dataset: To evaluate the effectiveness of our dataset in improving both the general phoneme error rate (PER) and homograph disambiguation, we fine-tuned the base GE2PE model (ByT5) using only our data with the same hardware setup and training configuration as for Homo-GE2PE, referring to this variant as Homo-ByT5. The learning curves can be seen in Figure 11. Despite our dataset being an order of magnitude smaller than the 5-million-sample synthetic dataset used in the original GE2PE study (Rahmati and Sameti, 2024), Homo-ByT5 achieves competitive phoneme error rate (PER) and high homograph accuracy (Table 3), demonstrating the quality and utility of our approach.

Evaluation of Inference Speed: Another critical factor is inference speed. While the Homo-GE2PE model outperforms HomoFast eSpeak in accuracy, it is orders of magnitude slower, making it impractical for real-time applications such as screen readers. Figure 12 in the appendix presents the speed and accuracy of all available and pro-

posed G2P tools. All inference tests were conducted on Google Colab (CPU runtime).[5] The color heatmap highlights lower-performing models in red and higher-performing models in green. As shown, eSpeak and HomoFast eSpeak are the fastest models, with the latter benefiting from a newly added feature that enables processing of larger text segments in a single run.

5 Conclusion

In this work, we tackled two persistent challenges in homograph disambiguation for low-resource languages: the high cost of dataset creation and the latency constraints of real-time G2P applications. We proposed a semi-automated pipeline for building homograph-rich datasets and introduced HomoRich, the first large-scale, openly licensed Persian homograph dataset. Using this resource, we achieved a 29.72 percentage-point improvement in homograph accuracy for a state-of-the-art neural G2P model.

To bridge the gap between accuracy and real-time performance, we further developed a lightweight, context-aware statistical method that enhances homograph handling with minimal computational overhead. Integrated into the widely used eSpeak engine, this method led to HomoFast eSpeak, a fast, homograph-aware G2P system that improves disambiguation accuracy by 30.66 percentage-points while retaining the responsiveness crucial for screen readers and other accessibility tools.

Our results highlight the potential of using high-quality offline datasets not only to train neural models, but also to enrich and modernize traditional rule-based systems. By releasing all resources under a CC0-1.0 license, we aim to foster further research and practical adoption in accessibility technologies for low-resource languages.

6 Limitations

Homograph disambiguation is not the only context-dependent challenge in Persian. Another notable challenge is the correct phonemization of the Ezafe, a linking phoneme that grammatically and semantically connects words. This is a major weakness in current non-neural systems.

Addressing such context-sensitive phenomena requires further research, particularly in designing fast yet linguistically aware non-neural methods. Tackling challenges like Ezafe handling could bring rule-based G2P models significantly closer to the naturalness of neural models—while maintaining the speed advantage crucial for real-world deployment.

[5] Inference scripts and benchmarking code are available at https://github.com/MahtaFetrat/Persian-G2P-Tools-Benchmark.

Acknowledgments

We gratefully acknowledge AvalAI (AvalAI, 2023) for providing API credits that helped accelerate this research in leveraging large language models for dataset preparation and analysis.

References

2002. Wiktionary free dataset. https://www.wiktionary.org/. Accessed: 2024-09-05.

2004. Persian zaya dictionary. https://peykaregan.ir/dataset/%D9%88%D8%A7%DA%98%DA%AF%D8%A7%D9%86-%D8%B2%D8%A7%DB%8C%D8%A7%DB%8C-%D8%B2%D8%A8%D8%A7%D9%86-%D9%81%D8%A7%D8%B1%D8%B3%DB%8C. Accessed: 2024-09-05.

2007. NV Access - home of the nvda screen reader. https://www.nvaccess.org/. Accessed: 2025-04-29.

2017. Persian jame glossary. http://farhang.apll.ir/. Accessed: 2024-09-05.

2019. Tihu persia dictionary. https://github.com/tihu-nlp/tihudict.

2022. Ipa-translator. https://github.com/lotusfa/IPA-Translator.

Mohammad Hasan Sohan Ajini. 2022. Attention based grapheme to phoneme. https://github.com/mohamad-hasan-sohan-ajini/G2P. Accessed: 2025-04-22.

Elham Alayiaboozar, Amirsaeid Moloodi, and Manouchehr Kouhestani. 2019. Word sense disambiguation focusing on pos tag disambiguation in persian: A rule-based approach. *International Journal of Information*, 17(2):119–134.

Sajad Alipour. 2023. Persian grapheme to phoneme with transformer. https://github.com/sajadalipour7/Persian-Grapheme-To-Phoneme-With-Transformer. Accessed: 2025-04-22.

Sawsan Alqahtani, Hanan Aldarmaki, and Mona Diab. 2019. Homograph disambiguation through selective diacritic restoration. *arXiv preprint arXiv:1912.04479*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

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

AvalAI. 2023. https://avalai.ir/. Accessed: 2025-04-22. [link].

Seyed Moein Ayyoubzadeh and Kourosh Shahnazari. 2024. Persian homograph disambiguation: Leveraging parsbert for enhanced sentence understanding with a novel word disambiguation dataset. *arXiv preprint arXiv:2406.00028*.

Giulia Comini, Heereen Shim, and Sam Ribeiro. 2025. Lightweight neural front-ends for low-resource on-device text-to-speech.

Hafez Dehghani. 2022. persian_phonemizer: A tool for translating persian text to ipa. https://github.com/de-mh/persian_phonemizer. Accessed: 2025-04-22.

Pierre Delattre. 1945. Spanish is a phonetic language. *Hispania*, 28(4):511–516.

Jonathan Duddington. 2024. espeak: Compact open source speech synthesizer. https://espeak.sourceforge.net/. Accessed: 2025-04-22.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *ArXiv*, abs/2005.12515.

Mahta Fetrat. 2024a. Kaameldict: A dictionary dataset. Hugging Face Dataset. Accessed: 2025-04-22.

Mahta Fetrat. 2024b. Sentencebench: A benchmark for sentence-level g2p in persian. https://huggingface.co/datasets/MahtaFetrat/SentenceBench. Accessed: 2025-04-30.

Mahta Fetrat. 2025. Gptinformal-persian: Informal persian text dataset. Hugging Face Dataset. Accessed: 2025-04-22.

Heting Gao. 2024. *Unsupervised speech technology for low-resource languages*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Masood Ghayoomi. 2019. Identifying persian words' senses automatically by utilizing the word embedding method. *Iranian Journal of Information Processing & Management*, 35(1):25–50.

Gooshkon. 2022. New version of espeak for android and windows notice. https://gooshkon.ir/1401/03/%D9%86%D8%B3%D8%AE%D9%87-%D8%AC%D8%AF%DB%8C%D8%AF-%D8%A7%DB%8C%D8%B3%D9%BE%DB%8C%DA%A9-%D8%A8%D8%B1%D8%A7%DB%8C-%D8%A7%D9%86%D8%AF%D8%B1%D9%88%DB%8C%D8%AF-%D9%88-%D9%88%DB%8C%D9%86%D8%AF%D9%88%D8%B2-e/. Accessed: 2025-04-29.

Kyle Gorman, Gleb Mazovetskiy, and Vitaly Nikolaev. 2018. Improving homograph disambiguation with supervised machine learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Roshan Research Group. 2023. Hazm: A python library for digging into persian text. GitHub repository. Version 0.7.0. Accessed: 2025-04-22.

Maria-Loulou Hajj, Martin Lenglet, Olivier Perrotin, and Gérard Bailly. 2022. Comparing nlp solutions for the disambiguation of french heterophonic homographs for end-to-end tts systems. In *International Conference on Speech and Computer*, pages 265–278. Springer.

Dongrui Han, Mingyu Cui, Jiawen Kang, Xixin Wu, Xunying Liu, and Helen Meng. 2024. Improving grapheme-to-phoneme conversion through in-context knowledge retrieval with large language models. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 631–635. IEEE.

Marti Hearst. 1991. Noun homograph disambiguation using local context in large text corpora. *Using Corpora*, pages 185–188.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Maria Karamihaylova. 2023. Neural network vs. rule-based g2p: A hybrid approach to stress prediction and related vowel reduction in bulgarian.

Özgün Koşaner, Çağdas Can Birant, and Özlem Aktaş. 2013. Improving turkish language training materials: Grapheme-to-phoneme conversion for adding phonemic transcription into dictionary entries and course books. *Procedia-Social and Behavioral Sciences*, 103:473–484.

Mohamadreza Mahmoodvand and Maryam Hourali. 2015. Persian word sense disambiguation corpus extraction based on web crawler method. *Advances in Computer Science: an International Journal*, 4(5):101–106.

Mohamadreza Mahmoodvand and Maryam Hourali. 2017. Semi-supervised approach for persian word sense disambiguation. In *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 104–110. IEEE.

Ali Moghadaszadeh, Fatemeh Pasban, Mohsen Mahmoudzadeh, Maryam Vatanparast, and Amirmohammad Salehoof. 2024. Avashog2p: A multi-module g2p converter for persian. In *2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 343–348. IEEE.

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

Matilde Nanni. 2023. Disambiguating italian homographic heterophones with soundchoice and testing chatgpt as a data-generating tool.

Marco Nicolis and Viacheslav Klimkov. 2021. Homograph disambiguation with contextual word embeddings for tts systems.

NV Access Limited and contributors. 2023. Speech Player in eSpeak add-on for nvda. `https://addons.nvda-project.org/addons/speechPlayerInEspeak.en.html`. Accessed: 2025-04-29.

Demetry Pascal. 2020. Simple persian (farsi) grapheme-to-phoneme converter. `https://github.com/PasaOpasen/PersianG2P`. Accessed: 2025-04-22.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. *arXiv preprint arXiv:1708.01464*.

Artem Ploujnikov. 2024. Towards a unified model for speech and language processing.

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R Rabiee. 2025a. Llm-powered grapheme-to-phoneme conversion: Benchmark and case study. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R Rabiee. 2025b. Manatts persian: A recipe for creating tts datasets for lower-resource languages. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Mexico City, Mexico.

Azam Rabiee. 2019. Persian g2p. `https://github.com/AzamRabiee/Persian_G2P`. Accessed: 2025-04-22.

Elnaz Rahmati and Hossein Sameti. 2024. Ge2pe: Persian end-to-end grapheme-to-phoneme conversion. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3426–3436.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Markéta Řezáčková, Daniel Tihelka, and Jindřich Matoušek. 2024a. Homograph disambiguation with text-to-text transfer transformer. In *Proc. Interspeech 2024*, pages 2785–2789.

Markéta Řezáčková, Daniel Tihelka, and Jindřich Matoušek. 2024b. T5g2p: Text-to-text transfer transformer based grapheme-to-phoneme conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Mahdi Rezaei, Negar Nayeri, Saeed Farzi, and Hossein Sameti. 2022. Multi-module g2p converter for persian focusing on relations between words. *arXiv preprint arXiv:2208.01371*.

Noushin Riahi and Fatemeh Sedghi. 2012. A semi-supervised method for persian homograph disambiguation. In *20th Iranian Conference on Electrical Engineering (ICEE2012)*, pages 748–751. IEEE.

Jennifer M Seale. 2021. *Label imputation for homograph disambiguation: theoretical and practical approaches*. Ph.D. thesis, City University of New York.

Denilson C Silva, Daniela Braga, and Fernando Gil V Resende Jr. 2012. A rule-based method for homograph disambiguation in brazilian portuguese text-to-speech systems. *Journal of Communication and Information Systems*, 27(1).

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. Phonologybench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14.

Virongrong Tesprasit, Paisarn Charoenpornsawat, and Virach Sornlertlamvanich. 2003. A context-sensitive homograph disambiguation in thai text-to-speech synthesis. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pages 103–105.

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

Kaisheng Yao and Geoffrey Zweig. 2015. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv preprint arXiv:1506.00196*.

David Yarowsky. 1997. Homograph disambiguation in text-to-speech synthesis. In *Progress in speech synthesis*, pages 157–172. Springer.

J Zhu, C Zhang, and D Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. *Interspeech 2022*.

A Extended Related Work Review

This appendix provides a more detailed review of prior work on homograph disambiguation. We organize the discussion into two parts: first, we survey general approaches used across languages, including rule-based, statistical, neural, and hybrid methods. Then, we turn our focus to Persian-specific efforts, particularly those that involve the creation or use of datasets aimed at addressing the scarcity of resources for homograph disambiguation in low-resource settings.

A.1 Homograph Disambiguation Approaches

This subsection reviews the main approaches proposed for homograph disambiguation across languages. To provide a clear structure, we divide the methods into five categories: rule-based, neural, hybrid, LLM-based, and other approaches. This organization reflects both the chronological evolution and methodological diversity of the field.

A.1.1 Rule-based and statistical Approaches

Silva et al. (2012) proposed a rule-based algorithm set as their core method for homograph disambiguation in Brazilian Portuguese text-to-speech systems. Their approach utilizes linguistic rules based on morphosyntactic and semantic analysis, employing information from the surrounding context, including part-of-speech, morphology, lemmas, and semantic relations from Wordnets, along with restrict lexical combinations. The authors tested their algorithms on existing text databases, namely a newspaper corpus (CETENFolha), the Holy Bible in BP, and Brazilian literature.

Yarowsky (1997) developed a corpus-driven approach for English homograph disambiguation, utilizing a 400-million-word multi-domain dataset that included news articles, scientific texts, and literary works. Their method employed statistical decision lists that ranked contextual patterns (including adjacent words and part-of-speech tags) by their log-likelihood ratios to determine correct pronunciations, effectively addressing seven major categories of homographs through data-driven rules rather than neural networks. The work demonstrated how large-scale, diverse training data could be leveraged to resolve lexical ambiguities with high accuracy.

Hearst (1991) proposed a method for noun homograph disambiguation in English using a large unrestricted text corpus, the Academic American Encyclopedia, which contains approximately 8.6 million words. To address the lack of sense-annotated data, the author manually labeled a small set of training instances for each homograph—testing the method on five English nouns (e.g., bank, tank, bass)—and further improved performance through an unsupervised learning phase that incorporated high-confidence predictions without additional manual effort. The core method, called CatchWord, is rule-based and relies on shallow contextual cues such as syntactic patterns, orthographic features (e.g., capitalization), and lexical co-occurrence information extracted from local context windows. This approach avoids deep semantic resources or inference and demonstrates that coarse-grained disambiguation can be effectively achieved using lightweight, corpus-driven statistical techniques.

A.1.2 Neural Approaches

Yao and Zweig (2015) explored applying sequence-to-sequence neural network models to the grapheme-to-phoneme (G2P) task, which is distinguished from machine translation and image captioning by its small vocabulary and the need for exactly correct outputs. The authors investigated two main approaches: a simple encoder-decoder LSTM model and alignment-based models. The simple encoder-decoder LSTM was found to perform well and was close to the state-of-the-art without requiring explicit alignment information. However, by allowing the neural network to use the same alignment information as conventional methods, the authors were able to significantly advance the state-of-the-art with a bi-directional LSTM architecture. The bi-directional LSTM uses one RNN to process the input from left-to-right and another to process it right-to-left, combining their outputs to predict the next phoneme. They also found that deeper bi-directional networks further improved performance.

Peters et al. (2017) introduced a massively multilingual neural approach for grapheme-to-phoneme (G2P) conversion, which aims to address the lack of resources for low-resource languages by training a single system on data from hundreds of languages. The model, which is based on an encoder-decoder architecture with attention, shares a single encoder and decoder across all languages. To handle the different pronunciation patterns of various languages, the system prepends an artificial language ID token (e.g., <eng>) to the input grapheme sequence. This approach exploits the intrinsic similarities between different writing systems and improves

performance on low-resource languages by allowing them to implicitly share parameters with high-resource languages. The authors demonstrated an 11% improvement in phoneme error rate (PER) over a baseline approach that adapts high-resource monolingual models to low-resource languages. They also noted that the language ID token was highly beneficial for performance, especially when an embedding had been learned for it, and that the model was much more compact compared to previous approaches.

Nicolis and Klimkov (2021) proposed a homograph disambiguation system for American English text-to-speech (TTS) applications, focusing primarily on neural methods rather than rule-based ones. They used a publicly available dataset comprising 138 homograph words, each with around 90 training and 10 test sentences, and addressed data imbalance by manually augmenting the training set for underrepresented homograph variants using an internal fiction-based corpus. This augmentation, which added about 10 examples per weak variant, led to a relative accuracy improvement of over 11%, demonstrating the effectiveness of targeted data enrichment. Their method relies on contextual word embeddings (CWEs) extracted from pretrained BERT and ALBERT models, which are then fed into lightweight logistic regression classifiers trained separately for each homograph. This fully ML-based approach achieves state-of-the-art performance without the need for hand-crafted rules.

Seale (2021) addressed the challenge of low-resource data in homograph disambiguation by exploring label imputation techniques. To mitigate this, the author generated four homograph disambiguation datasets and made them available for the research community. The author also used the Wikipedia Homograph Data (WHD) released by Gorman et al. (2018) to conduct the research. Their core method involved employing regularized, multinomial logistic regression and fine-tuning pretrained ALBERT, BERT, and XLNet language models as token classifiers to improve model performance, particularly in classes with low prevalence samples.

Ploujnikov (2024) proposed SoundChoice, a sentence-level Grapheme-to-Phoneme (G2P) model aimed at improving homograph disambiguation in English. To address the challenge of context-aware phoneme prediction, they constructed the LibriG2P dataset, which integrates lexicon-based word pronunciations from CMUDict, phoneme alignments from LibriSpeech, and Wikipedia homograph data. This dataset includes approximately 10259 homograph-labeled samples, addressing inconsistencies between lexicon-based and audio-derived phoneme annotations. Their model employs a hybrid neural architecture, leveraging LSTMs, GRUs, and content-based attention, alongside CTC loss and curriculum learning—progressing from individual word training to sentence-level fine-tuning for enhanced contextual phoneme prediction. Additionally, BERT word embeddings are incorporated to inject semantic knowledge for better homograph resolution, achieving a phoneme error rate (PER) of 2.65% and 94% homograph classification accuracy. This work contributes to dataset development and model innovations in grapheme-to-phoneme conversion.

Řezáčková et al. (2024b), Řezáčková et al. (2024a) introduced a grapheme-to-phoneme (G2P) conversion approach using a Text-to-Text Transfer Transformer (T5) model. To capture cross-word context and assimilation effects, their models for English and Czech were trained on proprietary datasets of several hundred thousand sentences provided by language experts, mitigating the need for explicit rule-based post-processing. The T5-based model achieved high conversion accuracy across the tested languages.

Comini et al. (2025) present a neural-based lightweight front-end for on-device TTS in English, Polish, and Russian, using internal pronunciation dictionaries and the Kaggle text normalization dataset to address data limitations. Their dataset includes 53.1k, 42.4k, and 31.9k words for G2P and 6.4k, 6.5k, and 11.2k tokens for TN. They employ transformer-based and GRU-based models, leveraging knowledge distillation from pre-trained teacher models to train compact student models, optimizing for low latency and scalability in low-resource scenarios.

Gao (2024) tackle speech processing for low-resource languages using neural methods, particularly self-supervised learning (SSL) with models like wav2vec2 and HuBERT. They use existing speech datasets (e.g., LibriSpeech, VoxPopuli, CommonVoice) and enhance SSL pretraining with synthetic speech generated by diffusion models to address data scarcity. Their approach improves multilingual and zero-shot phonetic recognition without requiring labeled data.

Nanni (2023) investigated homographic heterophone disambiguation in Italian Text-To-Speech (TTS) systems using the SoundChoice model, which includes an RNN (LSTM + GRU) and a transformer version. Given the scarcity of Italian homograph datasets, the study generated 9,916 sentences with ChatGPT, supplementing a 1,700-sentence corpus dataset. The ChatGPT-generated data was created through iterative prompting, where sentences were crafted to include homographs in varying syntactic contexts. These sentences were manually validated for linguistic accuracy and context relevance before phonetic transcription using a ReadSpeaker transcription tool, which had a 59.56% accuracy in homograph resolution. The model integrates semantic disambiguation via BERT embeddings and a weighted homograph loss, enabling sentence-level pronunciation prediction. Evaluation showed the transformer model outperformed the RNN, highlighting the feasibility of neural methods for Italian homograph disambiguation.

A.1.3 Hybrid Approaches

Gorman et al. (2018) addressed homograph disambiguation for English TTS by creating a labeled dataset of 163 homographs (including morphosyntactic, lexical, and mixed types), with 100 sentences per homograph sampled from Wikipedia and annotated via crowdsourcing. To mitigate data scarcity, they employed rigorous adjudication for label disagreements and released the dataset publicly. Their hybrid system combined rule-based heuristics (e.g., context-triggered pronunciation rules, POS tags) with supervised ML (per-homograph maxent classifiers using word-context, POS, and capitalization features), showing that hybridization outperformed either approach alone.

Karamihaylova (2023) developed a hybrid grapheme-to-phoneme (G2P) system for Bulgarian, combining rule-based finite-state transducers (FSTs) for consonant mapping and vowel reduction rules with an LSTM-based seq2seq model for stress prediction. To address inconsistencies in publicly available data, they scraped and filtered 38,000 word-pronunciation pairs from Bulgarian Wiktionary using WikiPron, then standardized consonant transcriptions while preserving vowel variations to study stress-induced reduction. The dataset included homographs, where stress position disambiguates meaning. Their hybrid approach achieved performance comparable to pure neural methods, demonstrating the viability of curated rule-neural integration for medium-resource languages.

A.1.4 LLM-based Approaches

Suvarna et al. (2024) introduced PhonologyBench, evaluating Large Language Models (LLMs) on English phonological tasks, including homographs. Their dataset includes 3,000 words for grapheme-to-phoneme conversion, sourced from SIGMORPHON 2021, ensuring phonemic transcriptions. They tested GPT-4, Claude-3-Sonnet, and LLaMA-2-13B, using a zero-shot neural approach, showing that LLMs struggle with homograph pronunciation. Their findings highlight the need for phonology-aware datasets to improve text-based pronunciation models.

Han et al. (2024) explored the use of Large Language Models (LLMs) for grapheme-to-phoneme conversion, focusing on leveraging the in-context knowledge retrieval capabilities of GPT-4 to disambiguate homographs. To facilitate this, the authors constructed a dictionary by combining the Librig2p training dataset and the CMU dictionary. For homograph words, they used GPT-4 to generate cases automatically. Each homograph contains multiple cases and was later manually refined. The core of their method involves prompting GPT-4 to analyze the input sentence, identify the most relevant meaning and part-of-speech for the target word, and then retrieve the corresponding phoneme pronunciation from the constructed dictionary.

Qharabagh et al. (2025a) proposed an LLM-powered approach to Grapheme-to-Phoneme (G2P) conversion in Persian, addressing challenges posed by polyphone words and context-sensitive phonemes. To improve phonetic accuracy and benchmark sentence-level G2P performance, two datasets were introduced: Kaamel-Dict, a unified phonetic dictionary with over 120,000 entries, and Sentence-Bench, a sentence-level dataset containing 400 annotated sentences, including 100 polyphone words used in various contexts. The method leverages large language models (LLMs) without additional training, applying advanced prompting and post-processing techniques to enhance phonetic predictions. Benchmarking results demonstrate that LLMs can outperform traditional models, highlighting the potential of LLMs in low-resource G2P tasks.

A.1.5 Other Approaches

Tesprasit et al. (2003) addressed the challenges posed by word boundary and homograph ambiguity in Thai Text-to-Speech, noting the absence of word delimiters in the language. To conduct their research, they created their own 25K-word corpus where sentences were manually segmented, and part-of-speech tags and pronunciations were manually annotated by linguists. Their core method is a unified machine learning framework based on the Winnow algorithm, a statistical technique that learns to combine local and long-distance contextual features like context words and collocations to disambiguate word pronunciations without relying on predefined rules or standard neural network architectures.

Alqahtani et al. (2019) addressed homograph disambiguation in Arabic by proposing unsupervised, data-driven methods to selectively restore diacritics, balancing lexical disambiguation and sparsity. They leveraged existing corpora (50M tokens, including Gigaword and Arabic Treebank) without new data collection, using the MADAMIRA tool for automated diacritization and morphological analysis. Their approach identified 33.8% of words as homographs (e.g., 168K ambiguous types) by clustering diacritized variants (Brown, K-means) and analyzing translation divergences in parallel text. Unlike rule-based or neural methods, their work focused on distributional similarity and morphological variants to guide selective diacritization, demonstrating improved performance in downstream tasks like machine translation and POS tagging.

Hajj et al. (2022) addressed the challenge of disambiguating French heterophonic homographs for TTS systems by creating a custom dataset. They collected 8137 sentences from the web, ensuring a balanced representation of 34 pairs of prototypical homographs, with roughly one hundred instances per pair. To enhance disambiguation, they employed Linear Discriminant Analysis (LDA) classifiers, utilizing contextual word embeddings as input features, and experimented with the FlauBERT transformer for POS tagging.

A.2 Persian Homograph Disambiguation and Dataset Development

Several recent works have introduced or curated datasets specifically for Persian homograph disambiguation and word sense disambiguation (WSD). Notably, Moghadaszadeh et al. (2024) presented a dataset collected through a cluster-based sampling strategy to mitigate phoneme imbalance. Another valuable dataset is by Ghayoomi (2019), who developed a manually annotated gold standard for 20 Persian ambiguous words, each with 100 sentences, totaling 2000 sentences. These sentences were extracted from the Persian Language Database and annotated according to SemEval2010 guidelines. Similarly, Rahmati and Sameti (2024) generated over 5 million sentence-phoneme pairs, including manually and automatically labeled data which was a valuable source for general G2P task not homograph challenge.

Other works focused on smaller, curated datasets. Ayyoubzadeh and Shahnazari (2024) created a dataset containing 63 homograph words, with sentence-level phonetic annotations developed through careful selection. Mahmoodvand and Hourali (2017) extracted 5368 documents/sentences using a web crawler for three Persian homographs ("Shir", "Rast", "Tar") from Iranian news agency websites, partially labeled (2133 documents). Riahi and Sedghi (2012) used the Hamshahri corpus and manually tagged instances of two homographs, with training sizes ranging from 10 to 1500 words for their tri-training framework. Additionally, Nanni (2023) created an Italian homograph dataset, including 9,916 ChatGPT-generated sentences supplemented with 1,700 corpus examples.

The following paragraphs provide a more detailed examination of each study.

Riahi and Sedghi (2012) addressed the challenge of limited manually tagged data for Persian Word Sense Disambiguation (WSD) by proposing a semi-supervised method. To conduct their experiments, they utilized the raw Hamshahri corpus and created their own tagged data by manually annotating instances of two Persian homographs. Their core method employs a statistical approach based on tri-training with decision lists. The decision lists classify homographs by analyzing the distribution of collocations (surrounding words), and the tri-training framework iteratively leverages a small tagged corpus and a larger untagged corpus to improve disambiguation accuracy.

Moghadaszadeh et al. (2024) introduced AvashoG2P, a multi-module system for Persian grapheme-to-phoneme (G2P) conversion that primarily employs neural network-based approaches. For out-of-vocabulary word prediction, their core

method utilizes a sequence-to-sequence model with a GRU-based recurrent unit and an attention mechanism. Addressing the lack of labeled data for homograph disambiguation in Persian, the authors collected and labeled their own homograph data. To mitigate the challenge of data imbalance in their collected homograph data, they first clustered the data for each homograph before labeling a selection of samples from each cluster. Their homograph disambiguation module leverages a classification approach that uses a single model for all 54 supported Persian homographs, with experiments highlighting the superior performance of transformer-based models like XLMRoberta.

Ghayoomi (2019) proposed an unsupervised neural method for Persian word sense induction using word embeddings and hierarchical clustering. They trained embeddings on a combined corpus (529M words) and evaluated on a manually annotated dataset of 20 ambiguous words (100 sentences each). Their approach leveraged context windows (8 surrounding words) and sentence-level embeddings, clustering them without predefined rules.

Ayyoubzadeh and Shahnazari (2024) introduced a novel dataset for Persian homograph disambiguation, addressing the challenges posed by words with identical spellings but different meanings in Persian[1]. Their dataset includes diverse sentences containing homographs, which are carefully annotated to facilitate detailed analysis and model training. The authors trained both lightweight machine learning and deep learning models, leveraging embeddings and cosine similarity to disambiguate homographs and evaluated model performance using accuracy, recall, and F1 score.

Mahmoodvand and Hourali (2017) addressed the challenge of limited labeled data for Persian word sense disambiguation by implementing a semi-supervised machine learning approach. They created their own corpus by developing a crawler to extract sentences containing target ambiguous words from news agency websites, building a dataset specifically designed for WSD tasks. Their method leverages a small set of labeled seed data combined with a larger volume of unlabeled data in a collaborative learning framework, focusing on defined features of target words to disambiguate their meanings. The researchers evaluated their approach on three Persian homograph words ("Shir," "Rast," and "Tar"), achieving impressive results with 88% recall, 95% precision, and 93% accuracy across 5,368 documents, demonstrating the effectiveness of their semi-supervised approach for Persian language processing despite the inherent challenges of Persian's rich metaphorical nature and complex writing style.

Mahmoodvand and Hourali (2015) presented a method for building a Persian word sense disambiguation (WSD) dataset by employing a web crawler to gather documents containing specific ambiguous words. Addressing the lack of suitable WSD corpora for Persian, their approach focuses on extracting relevant phrases for ambiguous words from web data to create a dataset that can be used in WSD tasks. The authors used three prevalent Persian ambiguous words to extract appropriate phrases. This research provides a foundation for supervised WSD methods in Persian by offering a means to generate training data where it was previously scarce.

Rahmati and Sameti (2024) proposed GE2PE, a Persian end-to-end grapheme-to-phoneme conversion model that addresses the challenges of Persian homographs and missing short vowels by leveraging sentence-level context. To support this, they created two large datasets comprising over five million sentences with corresponding phoneme sequences, including both manually labeled and machine-generated data, and designed evaluation sets specifically for tasks like Kasre-Ezafe detection and homograph disambiguation. Their core approach is a ByT5 (Xue et al., 2022) model trained in a two-step process, building on advances in transformer architectures shown to be effective for G2P tasks. This work stands out for its extensive data creation tailored to Persian linguistic challenges and its end-to-end neural modeling strategy.

Rezaei et al. (2022) proposed a multi-module G2P system for Persian that addresses the challenges of homographs, OOV words, and ezafe constructions. To handle homographs, they extracted a homograph dictionary from the Ariana lexicon. Their core method involves a combination of GRU and Transformer architectures within separate modules to handle different aspects of G2P conversion. The system operates at the sequence level, capturing cross-word relations crucial for homograph disambiguation and ezafe recognition.

Alayiaboozar et al. (2019) proposed a rule-based approach for disambiguating Persian noun and adjective homographs ending in (/i/), leveraging context-sensitive syntactic rules (e.g., preposition + quantifier patterns) derived from three existing corpora: the Peykare corpus, Farsi Linguis-

tic Database, and Persian Dependency Treebank. They extracted 36 rules based on 10-word contextual windows, achieving high accuracy (e.g., 94% for some rules), but did not create new labeled data. Their method focused on morphological and syntactic patterns (e.g., adjacent POS tags) to resolve ambiguity in a language with prevalent homography due to orthographic constraints.

B Phoneme Representaion Mapping

There are two common representations for Persian phonemics. The first representation is the one used in many of the G2P glossaries, including KaamelDict (Fetrat, 2024a; tih, 2019; IPA, 2022; wik, 2002; zay, 2004; jam, 2017; Ajini, 2022; Pascal, 2020; Rabiee, 2019; Zhu et al., 2022) and benchmarks like SentenceBench (Fetrat, 2024b). The second representation is used in one of the state-of-the-art G2P models for Persian, namely GE2PE, which is fine-tuned and enhanced in this work. Our HomoRich dataset includes the sentence phoneme sequences in both of these formats for compatibility. Figure 9 shows these two representations.

A key challenge in mapping the Ezafe phoneme between these representations was its inconsistent annotation. The Ezafe is a short vowel /e/ used to indicate possession, relation, or description in Persian noun phrases. For instance, in the sentences "This is Ziba's flower" (/in gol-e zibA ast/) and "This flower is beautiful" (/in gol zibA ast/), the Ezafe appears as a linking /e/ sound, but its presence or absence can alter the meaning of the sentence. In the GE2PE representation, the Ezafe is denoted by an additional '1' symbol after the '/e/' phoneme, while in our dataset, the '/e/' phoneme alone may indicate either a regular vowel or an Ezafe.

To resolve this ambiguity, we employed a POS tagger (Group, 2023) with 99.249% accuracy to identify Ezafe constructions based on the grapheme sequence. For each Ezafe occurrence, we retrieved its phonemic form from the KaamelDict (Fetrat, 2024a) glossary and searched for the corresponding '/e/' phoneme in the phoneme sequence. A '1' symbol was then appended to the '/e/' to maintain consistency with the GE2PE representation.

C Additional Figures

Dataset Sentence Length Distribution A well-designed dataset for a G2P model should include sentences of varying lengths to ensure the model can accurately transcribe both short and long utterances. Sentence length is also an indicator of linguistic diversity and complexity. Figure 10 illustrates the distribution of sentence lengths in the Homorich dataset.

Learning Curves for ByT5 Training Figure 11 shows the training dynamics across all phases when fine-tuning ByT5 (Xue et al., 2022), with identical hyperparameters as described in Section 3.2.1.

Inference speed vs. performance Figure 12 demonstrates the trade-off between inference speed and performance for various G2P tools. The best-performing tools are relatively slower, while the fastest tool is low in performance. The eSpeak versions offer a balance, with fast inference and favorable performance.

D Additional Tables

Data sources in HomoRich Table 4 details the composition of the HomoRich dataset, listing the count of samples from each data source.

Source	Count
GPT-4o	257,915
CommonVoice	118,983
ManaTTS	76,561
human	69,560
GPTInformal	5,872
Homograph Samples (Human + GPT-4o)	**327,475**
Total	**528,891**

Table 4: Breakdown of the data sources and their sample counts in the HomoRich dataset.

E Statistical Analysis of Experimental Results

To provide a comprehensive view of the variability in the reported metrics, we present error bar plots for the Phoneme Error Rate (PER), Homograph Accuracy, and Inference Time across the evaluated G2P tools and proposed models. Figures 13, 14, and 15 illustrate these metrics, with error bars representing standard deviations across five runs. The inference time plot is rendered on a logarithmic scale to highlight differences across models with varying computational requirements.

Symbol	Persian Sound	IPA Equivalent	Example
A	ا, آ (long vowel)	ɑː	ماه: mAh
a	ـَ (short vowel)	æ	درد: dard
u	او (long vowel)	uː	دوست: dust
i	ای (long vowel)	iː	میز: miz
o	ـُ (short vowel)	o	ظهر: zohr
e	ـِ (short vowel)	e	ذهن: zehn
S	ش (consonant)	ʃ	شهر: Sahr
C	چ (consonant)	tʃʰ	چتر: Catr
Z	ژ (consonant)	ʒ	ژاله: ZAle
q	غ، ق (consonant)	ɣ, q	غذا: qazA, قند: qand
x	خ (consonant)	x	خاک: xAk
r	ر (consonant)	r	روح: ruh
y	ی (consonant)	j	یار: yAr
j	ج (consonant)	dʒ	نجات: nejAt
v	و (consonant)	v	ورم: varam
?	ع، ء، ئ (consonant)	ʔ	عمر: ?omr, آینده: ?Ayande

(a) Repr. 1 (used in this work and related studies)

Symbol	Persian Sound	IPA Equivalent	Example
a	ا, آ (long vowel)	ɑː	ماه: mah
/	ـَ (short vowel)	æ	درد: d/rd
u	او (long vowel)	uː	دوست: dust
i	ای (long vowel)	iː	میز: miz
o	ـُ (short vowel)	o	ظهر: zohr
e	ـِ (short vowel)	e	ذهن: zehn
$	ش (consonant)	ʃ	شهر: $/hr
C	چ (consonant)	tʃʰ	چتر: C/tr
;	ژ (consonant)	ʒ	ژاله: ;ale
q	غ، ق (consonant)	ɣ, q	غذا: q/za, قند: q/nd
x	خ (consonant)	x	خاک: xak
r	ر (consonant)	r	روح: ruh
y	ی (consonant)	j	یار: yar
j	ج (consonant)	dʒ	نجات: nejat
v	و (consonant)	v	ورم: v/r/m
@	ع، ء، ئ (consonant)	ʔ	عمر: @omr, آینده: @ay/nd

(b) Repr 2 (used in other prior literature)

Figure 9: Comparison of two commonly used phoneme representations for Persian sounds.

F Details of Human Subject Participation in Data Collection

As part of this study, we engaged approximately 200 human participants to contribute to the human-generated portion of our homograph sentence corpus. Specifically, we curated a list of 285 Persian homograph words, each with multiple valid pronunciations. These were organized into several Google Sheets, where each sheet listed a subset of homograph words along with their pronunciations, followed by five empty rows designated for sentence creation for each alternative.

Each homograph word appeared in multiple sheets to ensure that it was annotated by different individuals, and each participant received a subset of words—thus distributing the workload and encouraging diversity in linguistic expression. The instructions, originally provided in Persian, asked participants to compose Persian sentences that naturally incorporate the target homograph with the specified pronunciation. A translated excerpt of the instruction reads:

> "Please write five different Persian sentences using the given word with the pronunciation indicated below it. Try to make the sentences as natural and diverse as possible. Avoid repeating sentence structures or vocabulary."

Participants were explicitly encouraged to avoid sentence repetition, maintain lexical diversity, and write fluent, meaningful examples.

Each participant completed multiple such entries, and collectively, this process yielded a total of 69,560 high-quality, human-written sentences. The sentences form a valuable component of our dataset for disambiguating homograph pronunciation in context.

G Broader Impact

The ultimate goal of our work is to improve the quality of fast, rule-based G2P models—and neural G2P systems in general—so they can be effectively integrated into low-latency text-to-speech (TTS)

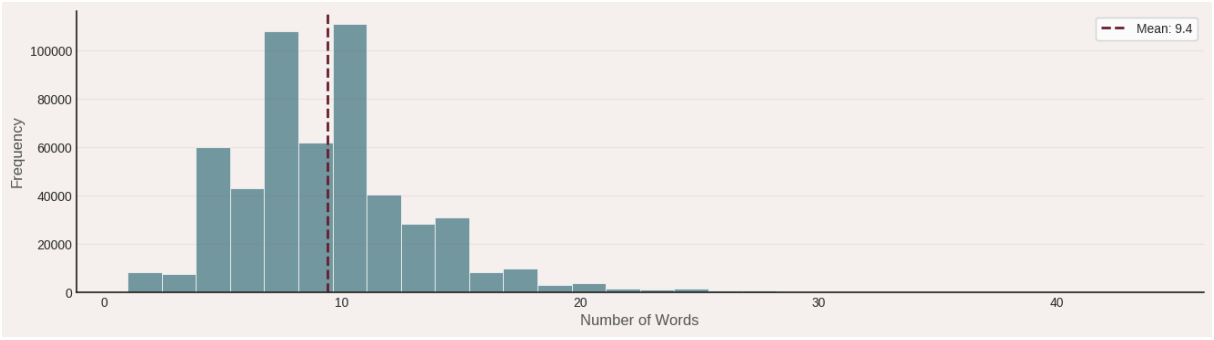
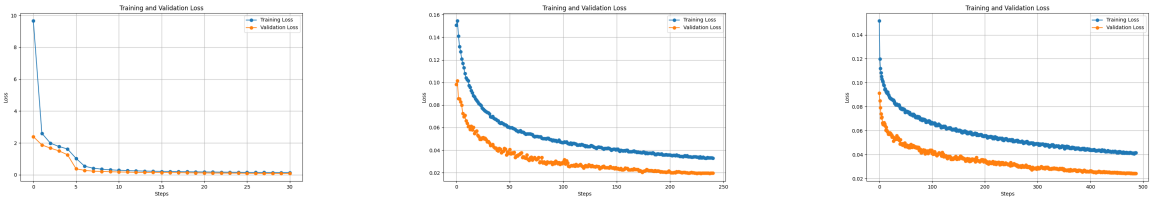


Figure 10: Distribution of sentence word counts.



(a) Phase 1 (5 epochs) (b) Phase 2 (20 epochs) (c) Phase 3 (50 epochs)

Figure 11: Learning curves across fine-tuning phases of ByT5.

pipelines, particularly for screen readers and other real-time accessibility tools. By enhancing homograph disambiguation and overall phonetic accuracy, we enable more natural and reliable speech synthesis, which is critical for users who rely on assistive technologies.

A key practical outcome of our research is the development of HomoFast eSpeak, an enhanced version of the widely used open-source eSpeak NG speech synthesizer. Our experiments show that HomoFast eSpeak achieves a 30.66 percentage-point improvement in homograph disambiguation accuracy while maintaining the low-latency performance critical for real-time applications. This advancement has the potential to elevate the intelligibility and naturalness of synthesized speech in screen readers used by the blind community in Iran.

Beyond immediate applications, we hope this work encourages further development of open, high-quality, and performant TTS systems for low-resource languages. By releasing our dataset (HomoRich), models (Homo-GE2PE), and enhancements to eSpeak under permissive licenses, we lower barriers for researchers and developers working on accessibility-focused speech technologies. Our contributions demonstrate that even simple, data-informed statistical methods can significantly improve rule-based systems—making high-quality G2P more scalable and sustainable for languages with limited resources.

H Disclosure of LLM usage

We used large language models (LLMs) for language refinement, including grammar correction, paragraph rephrasing, and other minor edits, based on drafts written by the authors. In the related work section, LLMs assisted in summarizing prior works after the authors had identified, read summaries of, and grouped the relevant literature; this use was limited to generating low-novelty text describing pre-existing methods and data. The generated text was subsequently reviewed for accuracy. Additionally, LLMs were used for fine-grained coding tasks such as generating individual functions or single-purpose scripts, which were then validated and integrated by the authors.

I Data Sheet

In the rest of this document, we present the datasheet for the HomoRich dataset, adhering to the guidelines outlined by Gebru et al. (2021).

I.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets

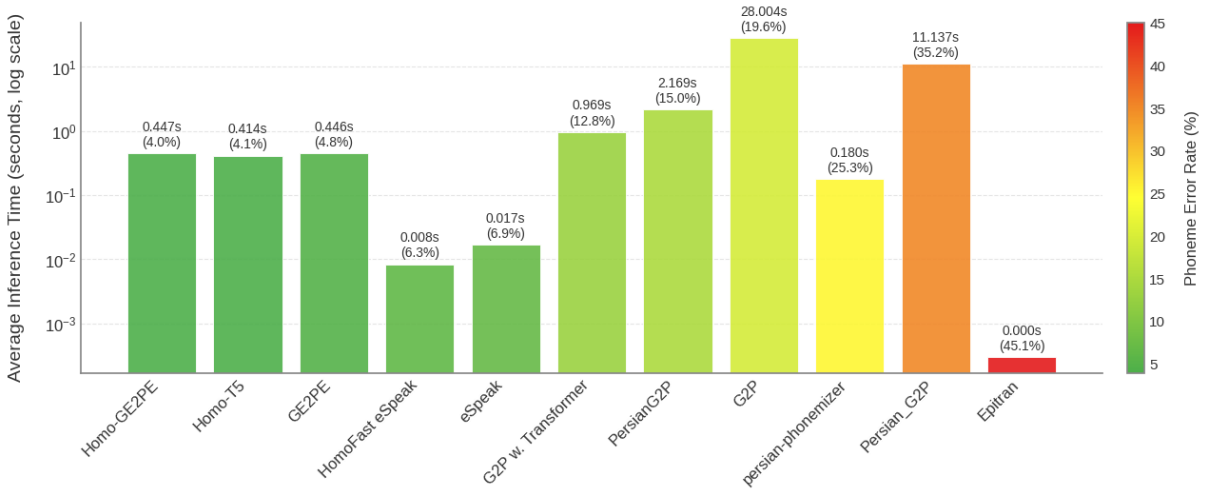


Figure 12: Inference speed and phoneme error rate (PER) of available and proposed G2P tools.

created for research purposes.

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

ANS: The dataset was created to address the scarcity of open-source datasets and models for grapheme-to-phoneme (G2P) conversion, with a focus on homograph disambiguation in Persian. These resources aim to support the development of open text-to-speech (TTS) and screen reader tools, enhancing accessibility for Persian-speaking communities, including individuals with visual impairments.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

ANS: The dataset was created by the speech processing team of the Data Science and Machine Learning (DML) Laboratory at Sharif University of Technology.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANS: The dataset creation received no external funding and is provided free of charge.

Any other comments?

ANS: No.

I.2 Composition

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

ANS: The dataset consists of Persian sentences (text) paired with their corresponding phoneme sequences in two formats (text). A subset of the dataset includes carefully curated Persian sentences containing homograph words, where each homograph and its pronunciation are explicitly annotated (text). All samples include metadata indicating their source (human, GPT-4o, CommonVoice, ManaTTS, or GPTInformal) and a unique identifier within each source category.

How many instances are there in total (of each type, if appropriate)?

ANS: The dataset contains 528,891 Persian sentences in total, with 327,475 specifically curated for homograph disambiguation.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of

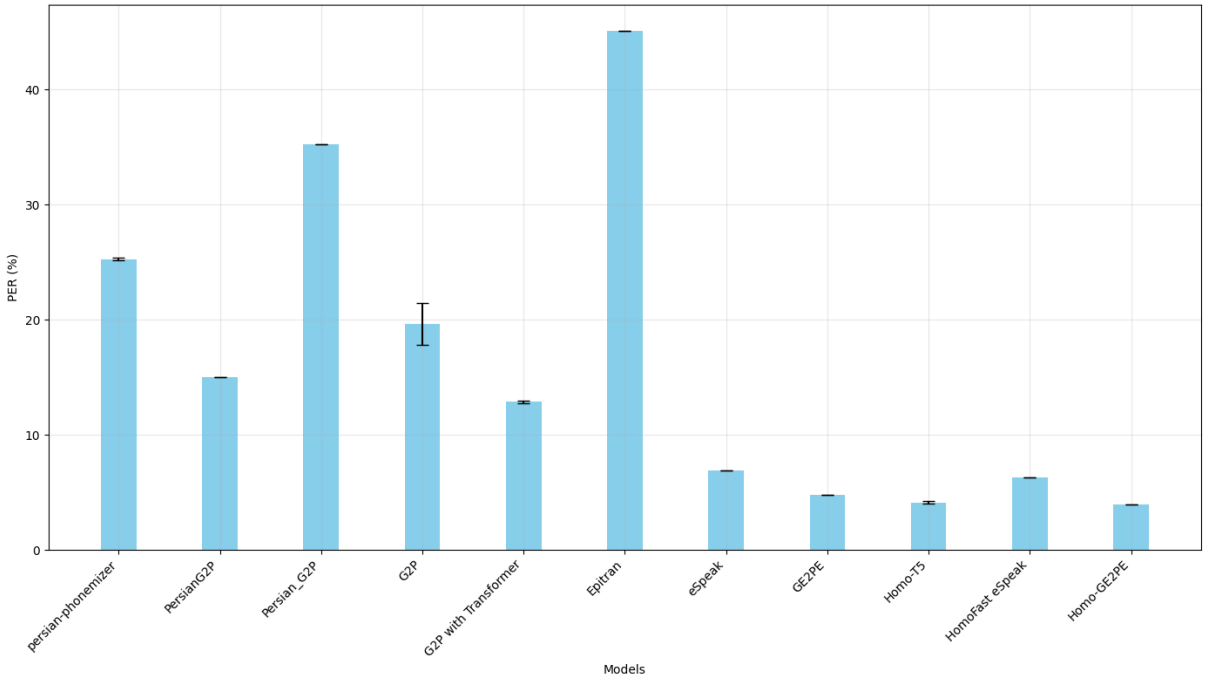


Figure 13: Phoneme Error Rate (PER) of previous and proposed G2P tools/models with error bars indicating standard deviations across five runs.

instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

ANS: The dataset incorporates: (1) complete samples from ManaTTS and GPTInformal (covering all available data at study time), and (2) a non-random subset of CommonVoice selected by availability (prioritizing validated samples while respecting original data ordering). GPT-4o generations and human annotations were collected specifically for this study.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

ANS: Each instance contains processed Persian text along with its corresponding phoneme sequence represented in two formats: a primary phonemic transcription and an alternative standardized representation mapped for compatibility. For instances containing homographs, the data additionally includes the identified homograph word and its correct pronunciation in both representation formats.

Is there a label or target associated with each instance? If so, please provide a description.

ANS: Yes, each instance serves multiple labeling purposes. The complete phoneme sequence of the sentence acts as the primary label. For homograph-containing instances, additional labels include the specific homograph word and its contextually correct pronunciation, enabling the dataset to support both general grapheme-to-phoneme conversion and specialized homograph disambiguation tasks.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

ANS: Due to our semi-automated data creation pipeline, sentences containing multiple homograph words only have the target homograph (the focus of that particular instance) annotated.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

ANS: No.

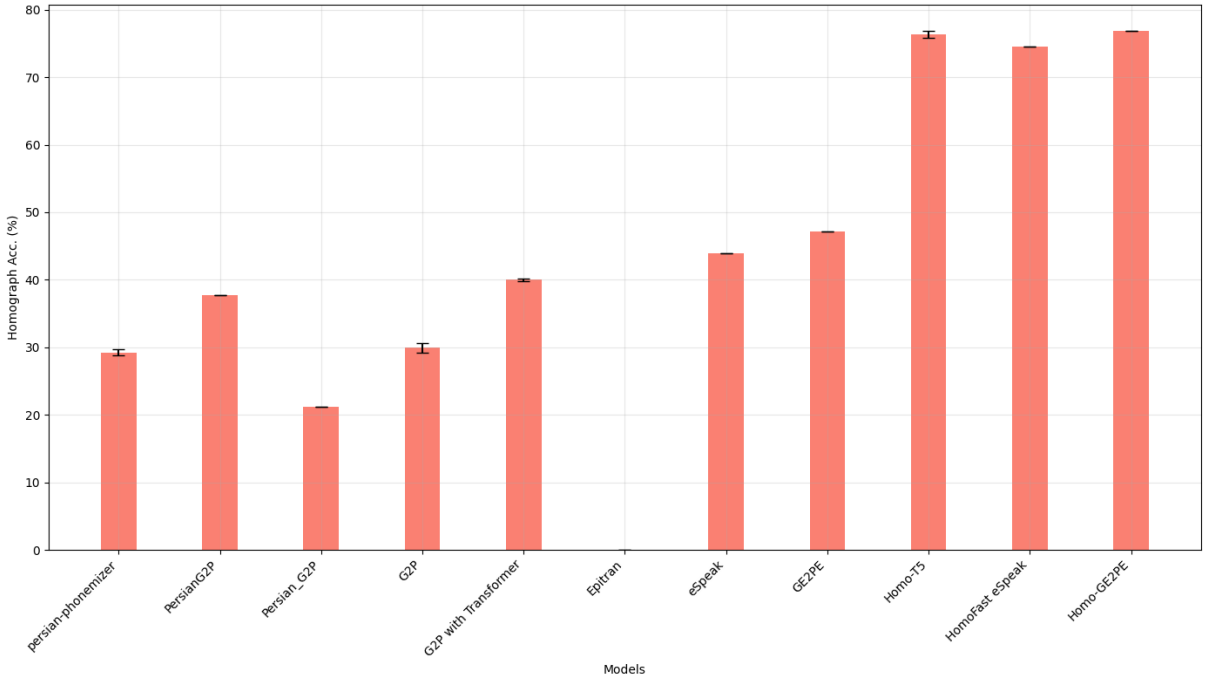


Figure 14: Homograph Accuracy of previous and proposed G2P tools/models with error bars indicating standard deviations across five runs.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

ANS: The dataset does not come with predefined splits. We recommend using the entire dataset for training while evaluating performance on the dedicated SentenceBench test set, following the methodology established in our work. This approach ensures consistent benchmarking across studies.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

ANS: As detailed in the data creation process, some sentences were generated by GPT-4o with prompts targeting specific homograph pronunciations. While we implemented techniques to prevent these issues, the approach carries inherent limitations including potential hallucinated sentences and occasional incorrect homograph usage. Additionally, phonemization was performed using the LLM-based method from prior work, which achieves a phoneme error rate of 6.43% and homograph accuracy of 64%, representing another source of potential noise in the phonetic transcriptions.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

ANS: The dataset is self-contained and doesn't rely on external resources.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

ANS: The dataset contains no confidential or personal information. All data originates from three sources: (1) established public datasets (CommonVoice, ManaTTS, and GPTInformal), (2) GPT-4o generated content, and (3) contributions from

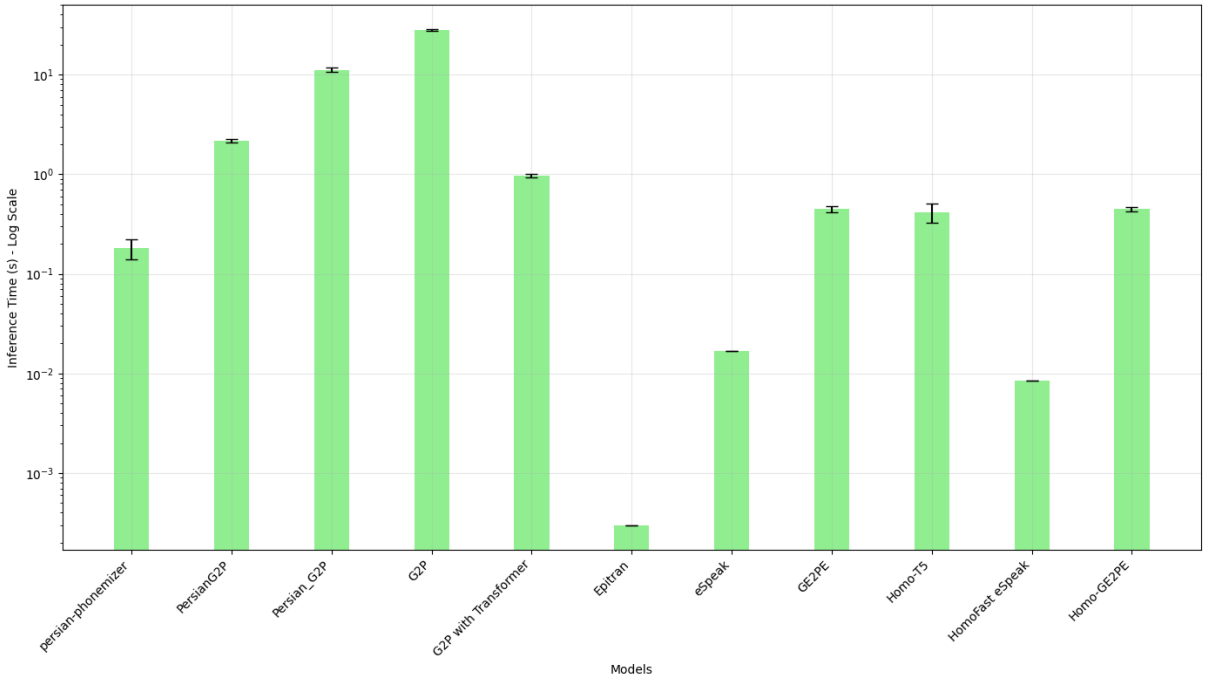


Figure 15: Inference Time (s) of previous and proposed G2P tools/models plotted on a logarithmic scale with error bars indicating standard deviations across five runs.

voluntary human participants who provided non-sensitive example sentences.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

ANS: The dataset is derived from well-known public datasets, the safeguarded GPT-4o model, and voluntary human subjects in an academic environment who were specifically asked to generate example sentences. Given these controlled sources and collection methods, we believe it is unlikely to contain offensive or harmful content. However, as with any language dataset, we recommend users review the content for their specific application needs.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

ANS: The dataset does not identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

ANS: We believe identification is not possible, as the data consists of voluntarily provided sample sentences generated for specific words.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

ANS: The dataset consists of linguistic examples derived from established public datasets, the safeguarded GPT-4o model, and voluntary contributions from participants in an academic setting. Given these controlled sources and collection methods focused solely on language patterns, we believe it is unlikely to contain sensitive information. However, as with any textual dataset, we recommend users assess the content for their specific requirements.

Any other comments?

ANS: No.

I.3 Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

ANS: The data combines three acquisition methods: (1) directly observable text from public datasets (CommonVoice, ManaTTS, GPTInformal), (2) GPT-4o-generated sentences with targeted homograph usage (indirectly derived through prompting), and (3) human-authored sentences voluntarily contributed in an academic setting. No specific validation was performed on the LLM-generated or human-provided data beyond the collection methods described in the paper.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

ANS: For the GPT-4o generated portion, data was collected through API calls using Python scripts. The human-authored content was gathered via online Google Sheets containing the target homograph words and detailed instructions, as documented in our methodology. No additional validation procedures were applied to these collection mechanisms.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

ANS: The dataset incorporates: (1) complete samples from ManaTTS and GPTInformal (covering all available data at study time), and (2) a non-random subset of CommonVoice selected by availability (prioritizing validated samples while respecting original data ordering). GPT-4o generations and human annotations were collected specifically for this study.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

ANS: The human-annotated portion of the dataset was collected through voluntary participation of native Persian speakers from diverse backgrounds. While we did not collect detailed demographic information about participants, their native language proficiency was the primary qualification for contribution. Participants were not financially compensated, as the data collection was conducted as part of an academic research initiative.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

ANS: The dataset was compiled in 2024-2025, combining newly generated GPT-4o outputs and human annotations with existing public corpora. The ManaTTS, GPTInformal, and CommonVoice components originate from their 2024 releases.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

ANS: No ethical review processes were conducted.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

ANS: The data was obtained from the individuals directly.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

ANS: The data was not collected from a pre-existing source; instead, individuals were explicitly instructed to generate the data, eliminating the need for notification.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

ANS: Similar to the previous response, since the data was generated based on explicit instructions provided to the individuals, consent was inherently obtained through participation, and no additional consent process was necessary.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

ANS: As the data generation was based on direct instructions and not from pre-existing sources or personal information, the issue of consent revocation does not apply in this context.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

ANS: No such analysis has been conducted.

Any other comments?

ANS: No.

I.4 Preprocessing/cleaning/labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

ANS: Yes, the underlying text corpora sourced from previous datasets and generated through GPT-4o or human annotators were phonemized as labels using the LLM prompting method outlined in a prior study, as referenced in the paper.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

ANS: Yes, the raw data includes the underlying text corpora from previous datasets (ManaTTS, GPTInformal, CommonVoice), as well as data generated using GPT-4o and contributions from human subjects. These data remain accessible and were only augmented with the phoneme labels as described earlier.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

ANS: Yes, the complete code for data generation and labeling is publicly accessible at https://github.com/MahtaFetrat/HomoRich-G2P-Persian.

Any other comments?

ANS: No.

I.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

Has the dataset been used for any tasks already? If so, please provide a description.

ANS: Yes, it has been employed to finetune two neural G2P models and enhance a rule-based G2P tool in our research, which is used to evaluate data efficiency.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

ANS: The dataset was not publicly available before this work, and as far as we know, it hasn't been utilized in any other projects.

What (other) tasks could the dataset be used for?

ANS: The dataset can be utilized for both general G2P conversion and specific homograph pronunciation disambiguation. Additionally, it could be valuable for tasks involving context understanding, such as word sense disambiguation. While not all sense disambiguations involve pronunciation differences, words with multiple pronunciations often convey distinct meanings that require contextual clarification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

ANS: We do not believe that the dataset carries such risks.

Are there tasks for which the dataset should not be used? If so, please provide a description.

ANS: We do not foresee any specific limitations regarding potential uses of the dataset.

Any other comments?

ANS: No.

I.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

ANS: Yes, the dataset is available to the public under a CC-0 license.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

ANS: The dataset is publicly available on Hugging Face and GitHub. A permanent DOI (10.57967/hf/6420) has been assigned to ensure citability. Links to the repositories are provided in the paper.

When will the dataset be distributed?

ANS: The dataset is made publicly available, and the repository links are included in the paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

ANS: The dataset will be shared under the CC-0 license, allowing free use.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

ANS: No, there are no IP-based or other restrictions imposed on the data associated with the instances.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

ANS: No, there are no export controls or other regulatory restrictions applicable to the dataset or individual instances.

Any other comments?

ANS: No.

I.7 Maintenance

The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

Who will be supporting/hosting/maintaining the dataset?

ANS: The dataset will be stored on public data repositories and maintained by the authors for updates.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

ANS: The authors can be contacted via email addresses provided in the author list.

Is there an erratum? If so, please provide a link or other access point.

ANS: There is currently no erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

ANS: We intend to update the dataset if significant errors are identified or if valuable community contributions can be incorporated. However, we do not plan to establish a formal mechanism for communicating changes. Updates can be tracked through the version history available on the hosting platforms (e.g., GitHub).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

ANS: There are no retention limits specified for the dataset.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

ANS: No, older versions will not be maintained. We do not plan to implement a specific mechanism to notify consumers of updates. Instead, changes can be observed through the version history available on the hosting platforms (e.g., GitHub).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

ANS: Contributions are very welcome. Contributors can open issues or submit pull requests on GitHub, or contact the authors directly for error reports or improvements.

Any other comments?

ANS: No.