G&P2P: A Multi-Source Approach to Grapheme-to-Phoneme Conversion

Chun-Yi Peng
Borough of Manhattan Community College, CUNY Graduate Center
199 Chambers St. New York, NY, 365 5th Ave. New York, NY
cpeng@bmcc.cuny.edu

Abstract

Grapheme-to-phoneme (G2P) conversion plays a central role in speech technologies. This paper introduces G&P2P, a multi-source framework that integrates multiple pronunciation dictionaries to enhance G2P modeling. We evaluate both expert-curated and crowd-sourced resources using attentive LSTM, pointer-generator LSTM, and transformer architectures. Results indicate that combining high-quality expert dictionaries yields substantial improvements, achieving an 11.26-point absolute (22% relative) reduction in word error rate. In contrast, incorporating noisy crowd-sourced resources may degrade performance. Statistical analyses further suggest that dataset quality exerts a greater influence on outcomes than the choice of fusion strategy, offering practical guidance for the design of multi-source G2P systems.

Keywords: grapheme-to-phoneme conversion, side-pronunciation, multi-source learning

1. Introduction

Grapheme-to-phoneme (G2P) conversion maps written word forms to their phonemic representations. It plays a crucial role in text-to-speech systems, automatic speech recognition, and low-resource language documentation. Despite significant progress in neural sequence modeling, G2P performance remains highly dependent on the quality and coverage of training lexicons.

Most existing systems are trained on a single pronunciation dictionary. However, multiple lexical resources often exist for a language, including expert-curated dictionaries and crowd-sourced repositories. These resources vary in coverage, transcription conventions, and annotation quality. While combining them may increase lexical coverage and improve out-of-vocabulary generalization, naive integration may introduce inconsistencies and noise.

This paper proposes G&P2P, a multi-source framework that integrates multiple pronunciation lexicons under controlled fusion strategies. The study investigates: 1) whether multi-source supervision improves G2P performance, 2) how dataset quality influences learning, 3) Whether fusion strategy significantly affects outcomes, 4) how LSTM and transformer architectures compare under multi-source training.

The results show that improvements primarily depend on dataset quality rather than fusion strategy, and that pointer-generator LSTMs outperform transformers in most settings.

2. Related Work

With the rise of deep learning, neural sequence models have since become dominant in G2P tasks (Ashby et al., 2021; Gorman et al., 2020; Kłosowski, 2022). Recurrent Neural Networks (RNNs) capture sequential dependencies (Bahdanau et al., 2015; Luong et al., 2015), and Long Short-Term Memory (LSTM) networks address vanishing gradient issues through gated memory mechanisms (Hochreiter and Schmidhuber, 1997; Eisenstein, 2019; Rao et al., 2015). Encoder–decoder LSTMs with attention further improved G2P by allowing dynamic alignment between input graphemes and output phonemes (Luong et al., 2015).

Transformer architectures (Vaswani et al., 2017) replaced recurrence with self-attention, enabling parallel computation and effective long-range modeling (Yolchuyeva et al., 2019b), though they often require larger datasets and greater computational resources (Yolchuyeva et al., 2019a,b). Evidence from the SIGMORPHON 2020 shared task suggests LSTM baselines can outperform transformers in medium-resource settings (Gorman et al., 2020). To better handle rare forms, pointer-generator architectures combine neural generation with explicit copying mechanisms (Vinyals et al., 2015; Prabhu and Kann, 2020; See et al., 2017), a design particularly well suited to G2P tasks.

However, prior work typically assumes a single lexicon source. Multi-source supervision remains underexplored in G2P research.

3. Data

3.1. Sources

The following English pronunciation resources were used for this study:

- **CELEX** (Baayen et al., 1995) transcribes English phonemes using DISC, where each

Proceedings of the Third Workshop on Computation and Written Language (CAWL 2026) @LREC 2026, pages 89–94
12 May, 2026.

phoneme is represented by exactly one ASCII character.

- **PronLex** (Kingsbury et al., 1994) provides ARPAbet transcriptions of Mainstream American English from the CALLHOME Project.
- **NETTalk** (Sejnowski and Rosenberg, 1988) transcribes English phonemes using a novel ASCII-based transcription system that assigns unique symbols or character combinations to represent different English sounds.
- **WikiPron** (Lee et al., 2020) provides crowdsourced pronunciation dictionaries from Wiktionary. This study uses separate dictionaries for British (WikiPron_UK) and American (WikiPron_US) English.

The first three are expert-curated and relatively consistent in transcription standards. WikiPron-US and WikiPron-UK offer broader coverage but contain heterogeneous and potentially noisy annotations. Table 1 shows the size of each dictionary.

Dataset	Size	Dataset	Size
CELEX	72,995	WikiPron_US	78,633
NETTalk	20,006	WikiPron_UK	79,520
PronLex	96,941		

Table 1: Size of each dictionary.

Sample transcriptions from these sources are shown in Table 2. Consonants are transcribed similarly across dictionaries, whereas the transcriptions of vowels vary. Take the word *lesson* for example, *on* is transcribed as 'H' in CELEX, 'ih0 n' in PronLex, 'N' in NETTalk, and 'ən' in Wiktionary.

Lexicon	Transcription of ***lesson***
CELEX	' l E s H
PronLex	l eh1 s ih0 n
NETTalk	l E1 s N
WikiPron-UK	l ɛ s ə n
WikiPron-US	l ɛ s ə n

Table 2: Transcriptions of the word *lesson* as it appears in the five databases.

3.2. Quality Assurance

Since pronunciation data on Wiktionary is crowdsourced, transcription consistency cannot always be guaranteed. To address this, a custom English extractor was developed to standardize the transcription of the English "r." The extractor incorporates two modifications: (a) replacing the trill /r/ with /ɹ/, and (b) replacing word-final /əɹ/ with /ɚ/.

3.3. Preprocessing

To incorporate auxiliary information, primary grapheme–phoneme pairs were fused with secondary pronunciations using three types of concatenation: columns, control symbols, and subscripts. In the examples that follow, CELEX serves as the secondary pronunciation and NETTalk as the target dictionary. Table 3 shows the size of each dataset after concatenation.

In the **column-based** fusion approach, the dataset is structured into three columns: the first column contains the grapheme, the second column contains the feature (i.e., side pronunciation), and the third column contains the target for prediction.
Source:

```
a b a c k    @ ' b { k
```

Target:

```
x0 b @1 k
```

In the **control-based** fusion approach, the dataset is structured into two columns: the first column is the grapheme and side-pronunciation, each with a control symbol, and the second column is the target column for prediction.
Source:

```
{grapheme} a b a c k {CELEX} @ ' b { k
```

Target:

```
x0 b @1 k
```

In the **subscript-based** fusion approach, the dataset is also structured into two columns: the first column is the grapheme and side-pronunciation, with a subscript after each letter. The second column is the target column for prediction, also with a subscript after each letter. The token suffixes (e.g., /grapheme, /CELEX, /NETTalk) are informally referred to as "subscripts." Their purpose is to keep the grapheme and phoneme vocabularies mutually disjoint, ensuring that the model can distinguish between the same surface character appearing in different roles — for instance, b/grapheme, b/PronLex, and b/celex are three distinct token types despite sharing the same surface form.
Source:

```
a/grapheme b/grapheme a/grapheme
c/grapheme k/grapheme @/CELEX
'/CELEX b/CELEX {/CELEX k/CELEX
```

Target:

```
x0/NETTalk b/NETTalk @1/NETTalk
k/NETTalk
```

4. Methods

This study implemented and evaluated four architectures using Yoyodyne (Wiemerslage et al., 2025): 1) attentive LSTM, 2) pointer-generator

Dataset	Size
CELEX → NETTalk	20,459
CELEX → PronLex	98,007
CELEX → WikiPron_UK	80,220
PronLex → CELEX	79,078
PronLex → NETTalk	21,887
PronLex → WikiPron_US	83,652
WikiPron_UK → CELEX	78,507
WikiPron_UK → WikiPron_US	102,793
WikiPron_US → NETTalk	24,348
WikiPron_US → PronLex	105,872
WikiPron_US → WikiPron_UK	104,990

Table 3: Size of each dataset.

LSTM, 3) transformer, and 4) pointer-generator transformer.

Hyperparameters were first optimized via a deep sweep (200 replicates) using a fixed random seed. The fine-tuning sweeps were synced on Weights & Biases for tracking and visualization. The configuration of the sweep with the highest validation accuracy score was selected as the best set of hyperparameters for the experiments. The experiments were run with five random seeds. Each model was evaluated using word error rate (WER) as the primary metric. The median of the five WERs for each model was reported in the next section.

5. Results

5.1. Comparison Across datasets

5.1.1. G2P vs. G&P2P

Table 4 compares the WERs of different models trained on the G2P and G&P2P datasets. The right arrow "→" denotes the direction of concatenation, where the dictionary on the left serves as the side-pronunciation being merged to the dictionary on the right. The baseline pointer-generator LSTM model trained solely on CELEX yields a median WER of 26.90. When concatenated with PronLex, the WER decreases substantially to 19.81, resulting in a 7.09 absolute (26.35 relative) reduction in error. Even when combined with WikiPron_UK, a crowd-sourced dataset, the WER still declines modestly by 0.13. A similar pattern is observed for the PronLex baseline. Concatenating PronLex with CELEX achieves an 11.81 absolute (37.37 relative) reduction in error. In contrast, concatenating PronLex with WikiPron_US results in only a marginal improvement, with the WER decreasing by 0.1 points to 31.50.

A comparable trend is observed for transformer-based models. However, for pointer-generator transformer models, incorporating WikiPron_UK or WikiPron_US actually degrades performance. This pattern suggests that transformer-based models may be more sensitive to noise in the training data.

5.1.2. Expert-curated vs. Crowd-sourced Datasets

The effects of data cleanliness become more pronounced when comparing expert-curated and crowd-sourced datasets. As shown in Table 5, in the pointer-generator LSTM models, fusing two expert-curated dictionaries yields substantially better performance than combining an expert-curated dictionary with a crowd-sourced one. When NETTalk serves as the target dictionary, and CELEX is used as a side pronunciation, the WER is 20.95. A similar pattern emerges when CELEX is fused with PronLex, yielding a WER of 19.81. In contrast, performance degrades sharply when WikiPron_UK or WikiPron_UK serves as the target dataset.

5.1.3. Between Fusion Techniques

Differences in WERs across the three techniques are generally within 1 point. A Friedman test confirms that the choice of fusion strategy does not substantially affect overall performance, $\chi^2(2) = 2.40$, $p = .30$.

5.2. Comparisons Across Architectures

5.2.1. LSTM vs. Transformer

Overall, LSTM-based models consistently outperform transformer-based models on the grapheme-to-phoneme conversion task (Table 6). For instance, under the CELEX→PronLex configuration, the LSTM model achieves a WER of 19.79, compared to 20.37 for the transformer. Although the absolute difference is modest, this pattern is consistent across configurations, suggesting that LSTM architectures are better suited to G&P2P tasks under the present experimental setup. A Wilcoxon signed-rank test conducted on WERs from the pointer-generator LSTM and pointer-generator transformer confirms the results ($p < .00$).

The comparison between pointer-generator LSTMs and transformers highlights the importance of computational cost. As shown in Table 7, the LSTM model required over three and a half hours of training (3h 42m), while the transformer completed training in just over an hour (1h 1m). This nearly threefold reduction in training time underscores the transformer's advantage in parallelizability. The efficiency gains are also reflected in model size. The pointer-generator LSTM contained approximately 53.6 million trainable parameters, corresponding to an estimated parameter size of 214.51 MB. The

Dataset	PG LSTM	PG Transformer	Transformer
CELEX	26.90	36.59	34.38
PronLex → CELEX	19.81	22.78	23.12
WikiPron_UK → CELEX	26.77	38.39	31.88
PronLex	31.60	37.45	36.88
CELEX → PronLex	19.79	20.37	20.72
WikiPron_US → PronLex	30.01	42.25	35.99
WikiPron_UK	53.11	58.29	57.02
CELEX → WikiPron_UK	52.06	56.52	53.59
WikiPron_US → WikiPron_UK	45.58	43.82	43.13

Table 4: G2P vs. G&P2P Comparison.

Dataset	WER
CELEX → NETTalk	20.95
CELEX → PronLex	19.79
CELEX → WikiPron_UK	52.06
PronLex → CELEX	19.81
PronLex → NETTalk	20.36
PronLex → WikiPron_US	55.71

Table 5: WERs from pointer-generator LSTM models (median values).

pointer-generator transformer, on the other hand, contained only 3.2 million trainable parameters, with a parameter size of 12.73 MB. This dramatic reduction in model size — by more than a factor of 16 — suggests that pointer-generator transformers can achieve competitive performance with far fewer parameters, thereby reducing both memory consumption and the computational burden of optimization.

5.2.2. Attentive vs. Pointer-Generator LSTMs

Attentive and pointer-generator LSTM-based models exhibit largely comparable performance (see Table 8). A Wilcoxon signed-rank test conducted on WERs from the two model variants reveals no statistically significant difference ($p = .09$). Nevertheless, when an expert-curated dictionary is fused with a noisy crowd-sourced dictionary (e.g., CELEX→WikiPron_UK and PronLex→WikiPron_US), the attentive LSTM consistently outperforms the pointer-generator LSTM, with improvements of 14.23 points in WER for CELEX→WikiPron_UK and 11.44 points in WER for PronLex→WikiPron_US.

6. Conclusion

The results suggest that incorporating side pronunciations from external dictionaries can improve performance, particularly when the additional resource is expert-curated. However, gains appear to depend on data quality, as adding a crowd-sourced and potentially noisy dictionary may not yield consistent improvements and may, in some cases, reduce performance.

LSTM- and transformer-based models show broadly comparable performance on expert-curated datasets. LSTM-based architectures appear particularly well-suited for G&P2P tasks when a crowd-sourced dictionary is combined with an expert-curated one. Transformer-based models, though more sensitive to noise, may serve as a viable alternative when computational efficiency and parameter reduction are primary considerations.

Looking ahead, a key direction is the integration of large language models (LLMs) into TTS pipelines, where their ability to capture broader context may improve G2P and prosody prediction. Hybrid approaches that combine LLMs with pointer-generator networks, along with transfer learning, could further enhance performance while reducing reliance on curated resources. In addition, expanding pronunciation lexicons through cross-resource projection offers a promising way to increase coverage and improve robustness, particularly for out-of-vocabulary words.

7. Acknowledgements

I would like to thank Kyle Gorman and the two anonymous reviewers for their valuable feedback. Any remaining errors are my own.

8. Bibliographical References

Lucas F. E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Work-*

Dataset	PG LSTM	PG transformer
CELEX → NETTalk	20.95	23.09
CELEX → PronLex	19.79	20.37
CELEX → WikiPron_UK	52.06	56.52
PronLex → CELEX	19.81	22.78
PronLex → NETTalk	20.36	22.34
PronLex → WikiPron_US	55.71	59.56
WikiPron_UK → CELEX	26.77	38.39
WikiPron_UK → WikiPron_US	44.60	45.40
WikiPron_US → NETTalk	24.03	55.93
WikiPron_US → PronLex	30.01	42.25

Table 6: WERs from pointer-generator LSTM and pointer-generator transformer models.

	PG LSTM	PG Trans.
Trainable params	53.6 M	3.2 M
Total params	53.6 M	3.2 M
Total size (MB)	214.512	12.732
Training Time	3h 42m	1h 1m 53s

Table 7: Comparison between pointer-generator LSTM and pointer-generator transformer models trained on the WikiPron_UK-CELEX-column dataset.

shop on Computational Research in Phonetics, Phonology, and Morphology, pages 115–125, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015), Conference Track Proceedings*, San Diego, USA.

Jacob Eisenstein. 2019. *Introduction to Natural Language Processing*. MIT Press.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Piotr Kłosowski. 2022. A rule-based grapheme-to-phoneme conversion system. *Applied Sciences*, 12(5).

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Nikhil Prabhu and Katharina Kann. 2020. Making a point: Pointer-generator transformers for disjoint vocabularies. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 85–92, Suzhou, China. Association for Computational Linguistics.

Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. 2015. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229, South Brisbane, Australia.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, USA. Curran Associates, Inc.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2692–2700, Montréal, Canada. Curran Associates, Inc.

Adam Wiemerslage, Kyle Gorman, and Travis M. Bartley. 2025. Yoyodyne: Small-vocabulary neu-

Dataset	Att. LSTM	PG LSTM
CELEX → NETTalk	24.52	20.95
CELEX → PronLex	19.78	19.79
CELEX → WikiPron_UK	**37.83**	**52.06**
PronLex → CELEX	20.03	19.81
PronLex → NETTalk	19.75	20.36
PronLex → WikiPron_US	**44.27**	**55.71**
WikiPron_UK → CELEX	24.60	26.77
WikiPron_UK → WikiPron_US	48.08	44.60
WikiPron_US → NETTalk	23.53	24.03
WikiPron_US → PronLex	31.50	30.01

Table 8: WERs from attentive and pointer-generator LSTM models.

ral sequence-to-sequence models. Computer software.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019a. Grapheme-to-phoneme conversion with convolutional neural networks. *Applied Sciences*, 9(6).

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019b. Transformer Based Grapheme-to-Phoneme Conversion. In *Interspeech 2019*, pages 2095–2099, Graz, Austria.

9. Language Resource References

Baayen, R. H. and Piepenbrock, R. and Gulikers, L. 1995. *CELEX2*. Linguistic Data Consortium. LDC Catalog No.: LDC96L14, ISBN: 1-58563-085-3.

Kingsbury, Paul and Strassel, Stephanie and McLemore, Cynthia and MacIntyre, Robert. 1994. *CALLHOME American English Lexicon (PRONLEX)*. Linguistic Data Consortium. LDC Catalog No.: LDC97L20, ISBN: 1-58563-110-8.

Lee, Jackson L. and Ashby, Lucas F.E. and Garza, M. Elizabeth and Lee-Sikka, Yeonju and Miller, Sean and Wong, Alan and McCarthy, Arya D. and Gorman, Kyle. 2020. *Massively Multilingual Pronunciation Modeling with WikiPron*. European Language Resources Association.

Sejnowski, Terry and Rosenberg, Charles. 1988. *Connectionist Bench (Nettalk Corpus)*. DOI: https://doi.org/10.24432/C5VP6T.