The SIGMORPHON 2022 Shared Task on Cross-lingual and Low-Resource Grapheme-to-Phoneme Conversion

Arya D. McCarthy[♠], **Jackson L. Lee**, **Alexandra DeLucia**[♠], **Travis Bartley**[♡], **Milind Agarwal**[♢], **Lucas F.E. Ashby**[♡], **Luca Del Signore**[♡], **Cameron Gibson**[♡], **Reuben Raff**[♡], **Winston Wu**[♣]
[♠]Johns Hopkins University [♡]City University of New York
[♢]George Mason University [♣]University of Michigan

Abstract

Grapheme-to-phoneme conversion is an important component in many speech technologies, but until recently there were no multilingual benchmarks for this task. The third iteration of the SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion features many improvements from the previous year's task (Ashby et al., 2021), including additional languages, three subtasks varying the amount of available resources, extensive quality assurance procedures, and automated error analyses. Three teams submitted a total of fifteen systems, at best achieving relative reductions of word error rate of 14% in the cross-lingual subtask and 14% in the very-low resource subtask. The generally consistent result is that cross-lingual transfer substantially helps grapheme-to-phoneme modeling, but not to the same degree as in-language examples.

1 Introduction

Many speech technologies demand mappings between written words and their pronunciations. In open-vocabulary systems, as well as certain resource-constrained embedded systems, it is insufficient to simply list all possible pronunciations; these mappings must generalize to rare or unseen words as well. Therefore, the mapping must be expressed as a mapping from a sequence of orthographic characters—*graphemes*—to a sequence of sounds—*phones* or *phonemes*.[1]

Grapheme-to-phoneme (g2p) datasets vary in size across languages (van Esch et al., 2016). In low-resource scenarios, an effective way of "breaking the resource bottleneck" (Hwa et al., 2005) is cross-lingual transfer of information from a high-resource language, either by annotation projection

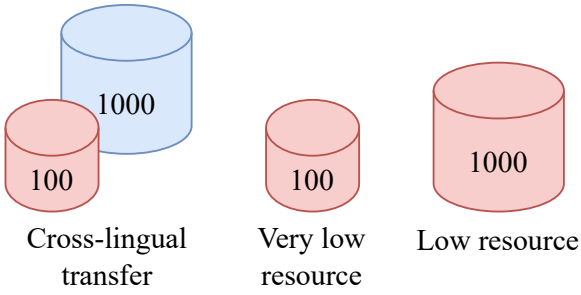

Figure 1: Training grapheme–phoneme pairs in the three subtasks. Transfer language is blue; target language is red. In all cases, the test set was 100 examples in the target language.

(Yarowsky and Ngai, 2001; Nicolai et al., 2020) or adapting a model to a new language (Zoph et al., 2016; Pino et al., 2019; McCarthy et al., 2019, 2020b; Mueller et al., 2020; Lee et al., 2022). The intent is that either the data or the learned representations and parameters carry across languages. Cross-lingual transfer shows promise for grapheme-to-phoneme conversion (Deri and Knight, 2016). Since this shared task began, *zero-shot* grapheme-to-phoneme procedures have been proposed, using no examples in the language of interest (Li et al., 2022).

SIGMORPHON in 2020 and 2021 hosted shared tasks on grapheme-to-phoneme conversion (Gorman et al., 2020; Ashby et al., 2021). The tasks have drawn wide participation, and in both years the participants outperformed the baseline systems by respectable margins. A major finding of the most recent iteration (Ashby et al., 2021) is that the largest improvements came from data augmentation, rather than alterations of the core model. Consequently, we have proposed a third edition of the shared task that explores data efficiency and language relatedness through cross-linguality.

This year's subtasks are designed so that, by contrasting these two aspects, we can answer two questions about data efficiency:

[1]We note that referring to elements of transcriptions as phonemes implies an ontological commitment which may or may not be justified; see Lee et al., 2020 (fn. 4) for discussion. Therefore, we use the term phone to refer to symbols used to transcribe pronunciations.

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 230–238
July 14, 2023 ©2023 Association for Computational Linguistics

1. How much does the transfer language data help?

2. How hard is it to model the language's grapheme-to-phoneme mapping, intrinsically?

This year, we study 10 language pairs, including two **surprise pairs** which were not released to the participants until close to the deadline as a challenge. Each pair of languages shares a script and some other relationship (e.g., phylogeny or hegemony). We investigate three data settings:[2]

Cross-lingual transfer A small amount of data (100 words) in the language of interest (the "target language") and a large amount of data (1000 words) in a nearby language (the "transfer language").

Very-low resource A small amount of data (100 words) in the target language and no data in the transfer language.

Low resource A large amount of data (1000 words) in the target language and no data in the transfer language.

In every case, we use the same 100-word test set, providing only the graphemes to the participants. Because the language pairs are consistent across the subtasks, we can draw meaningful contrasts.

Altogether, 15 systems were submitted, which allow substantial insights into our questions about data efficiency and g2p modelability. This third iteration of the SIGMORPHON shared task on graphme-to-phoneme conversion introduces transfer languages, new target languages, surprise languages, and stringent quality assurance, as the subtask structure which enables comparison.

2 Data

As in the two previous years, all pronunciation data was drawn from WikiPron (Lee et al., 2020), a massively multilingual pronunciation database extracted from the online dictionary Wiktionary. Depending on the language and script, Wiktionary pronunciations are either manually entered by human volunteers working from language-specific pronunciation guidelines or generated from the graphemic form via language-specific server-side scripting. WikiPron scrapes these pronunciations from Wiktionary, optionally applying case-folding to the graphemic form, removing any stress and syllable boundaries, and segmenting the pronunciation—encoded in the International Phonetic Alphabet—using the Python library segments (Moran and Cysouw, 2018). In all, 20 WikiPron languages were selected for the three subtasks. Only four of these were used in the 2021 iteration of the shared task. We give the twenty languages, as 10 target–transfer pairs, in Table 1.

Target language	Transfer language
Swedish	Norwegian Nynorsk
German	Dutch
Italian	Romanian
Ukrainian	Belarusian
Tagalog	Cebuano
Bengali	Assamese
Persian	Pashto
Thai	Eastern Lawa
Irish	Welsh
Burmese	Shan

Table 1: Language pairs used in the shared task. Irish–Welsh and Burmese–Shan were surprise pairs withheld until mid-April.

Morphological information from the UniMorph morphological lexicons (Kirov et al., 2018; McCarthy et al., 2020a) were again provided to participants; however, no participant made use of these, just like last year.

Language selection While the 2021 shared task considered both high-resource and low-resource settings, we did not control for the language itself. It was hard to extrapolate from the scores to claims about the resource requirements and difficulties of particular languages. This year, we use the same languages in all settings. This makes it reasonable and appropriate for the results to be directly compared, answering the two questions from Section 1.

These languages were chosen to avoid particularly pathological languages noted in previous years (English, Croatian) and those with unique and hard-to-predict phenomena, like *stød* in Danish.

Data quality assurance While the WikiPron data (Lee et al., 2020) that we use for the shared task is typically of high quality, some participants reported limitations in the English data. Consequently, we have omitted English data from the task. Beyond this, the data quality assurance pro-

[2]The data are available at https://github.com/sigmorphon/2022G2PST.

cedures are inspired by Ashby et al. (2021).

3 Task Definition

In this task, participants were provided with a collection of words and their pronunciations, and then scored on their ability to predict the pronunciation of a set of unseen words.

3.1 Subtasks

Last year, the task presented high-, medium-, and low-resource scenarios, each in different languages. This hampered cross-setting comparison, muddling whether differences in performance were due to data size, models, or languages.

This year, the same test sets are used across all settings, in the same set of languages. We offer a low-resource subtask, a very low-resource subtask, and a very low-resource subtask with more data available in a related (e.g., phylogenically or hegemonically) language. The relative error rates on each of three subtasks help to answer the research questions from Section 1. The design of these subtasks builds on McCarthy et al. (2019), which introduced the first shared task on cross-lingual transfer of information in morphological inflection.

Cross-lingual transfer This setting is meant to simulate a situation in which few data are available in the language of interest, but more are available in a related language, which can be leveraged. 100 words are given in each of the 10 languages, and an additional 1000 words are given in a related language for each language of interest. Throughout, we use the terms *transfer language* and *target language*, respectively, to refer to these. While it is realistic to have even more data available in a high-resource language, we constrain the size to enable comparison with the third setting.

Very-low resource This setting is designed to be extremely challenging. 100 words are given in each of the 10 languages. Comparing with the cross-lingual transfer setting gives insights about the value of the transfer data, and (indirectly) the similarities of the orthographic and phonetic systems present in the language pairs.

Low resource This setting matches the low resource condition from Ashby et al. (2021). 1000 words are given in each of the 10 languages. Comparing with the very-low resource setting gives insights about the learnability of the task. Comparing with both previous subtasks gives insights about the relevance of in-language data.

3.2 Data preparation

The procedures for sampling and splitting the data are similar to those used in the previous year's shared task; see Gorman et al. (2020, §3) and Ashby et al. (2021, §4.2). For each of the three subtasks, the data for each language are first randomly downsampled according to their frequencies in the Wortschatz (Goldhahn et al., 2012) norms. Words containing less than two Unicode characters or less than two phone segments are excluded, as are words with multiple pronunciations. The resulting data are randomly split into training data, development data, and test data. As in the previous year's shared task, these splits are constrained so that inflectional variants of any given lemma—according to the UniMorph (Kirov et al., 2018; McCarthy et al., 2020a) paradigms— can occur in at most one of the three shards. Training and development data was made available at the start of the task. The test words were also made available at the start of the task; test pronunciations were withheld until the end of the task.

Language-specific decisions The WikiPron data for Welsh has separate files for the North Wales and South Wales dialects. The South Wales dialect was chosen for there being slightly more data. Pashto, Eastern Lawa, and Shan do not have frequency data, so their "freq" file simply has the frequency of 1 for every word.

4 Evaluation

The primary metric for this task was word error rate (WER), the percentage of words for which the hypothesized transcription sequence is not identical to the gold reference transcription. As all three subtasks involve multiple languages, macro-averaged WER was used for system ranking. Participants were provided with two evaluation scripts: one which computes WER for a single language, and one which also computes macro-averaged WER across two or more languages. The 2020 shared task also reported another metric, phone error rate (PER), but this was found in the 2021 shared task to be highly correlated with WER and was not reported.

5 Baseline

The baseline system from 2021, the monotonic hard attention system from CLUZH (Makarov and Clematide, 2020), remained the baseline architecture in 2022. It is is a neural transducer system using an imitation learning paradigm (Makarov and Clematide, 2018).

All models were tuned to minimize per-language development-set WER. We reuse the best hyperparameter settings from last year. Alignments are computed using ten iterations of expectation maximization, and the imitation learning policy is trained for up to sixty epochs (with a patience of twelve) using the AdaDelta optimizer. A beam of size of four is used for prediction. Final predictions are produced by a majority-vote ten-component ensemble. Internal processing uses the decomposed Unicode normalization form (NFD), but predictions are converted back to the composed form (NFC). An implementation of the baseline was provided during the task and participating teams were encouraged to adapt it for their submissions.

In many cases, the baseline's loss did not improve over the course of training. We indicate this with a '-' in Tables 2 to 4.

6 Submissions

The shared task received 15 submissions from 3 teams. Below we provide brief descriptions of sub- missions to the shared task; more detailed descriptions of the first two—as well as various exploratory analyses and post-submission experiments—can be found in the system papers later in this volume.

Tü-G2P Girrbach (2022) evaluated three sequence labeling approaches to grapheme-to-phoneme conversion. In the supervised case, Girrbach trained a BiLSTM model to predict phoneme n-grams. The labels are derived from external alignments calculated by a custom neural aligner. Second, Girrbach trained a Gram-CTC model (Liu et al., 2017) to jointly predict and realign phoneme n-grams. Finally, the main approach is to use a standard BiLSTM sequence labeling model, but predict multiple ($\tau \in \{3, 4, 5\}$) phoneme unigrams from each grapheme. Girrbach uses standard CTC (Graves et al., 2006) to train the model, which is possible because predicting multiple phonemes from each grapheme causes the number of predicted symbols to always be greater than the number of target phonemes. Note that using CTC avoids relying on external alignments in any way. For the transfer task, Girrbach shares the same grapheme embeddings and BiLSTM encoder between target and transfer language, but uses different prediction layers.

Hammond Hammond (2022) submitted one system. He initially built a Transformer-based system, but because data are so minimal, it performed poorly. He switched to an HMM-based system (Novak et al., 2012).

For the transfer condition, which was his priority, he used the provided transfer data and augmented the system in two ways. First, he used a simplified version of the splicing augmentation scheme developed by Ryan and Hulden (2020) for the core data. Second, for the transfer languages, he only used data where the phonologies overlapped at the bigram level; in other words, he only included transfer training pairs that only included phonetic bigrams that occurred in the target languages.

mSLAM Garrette (2022) prepared a submission based on mSLAM (Bapna et al., 2022), a multilingual encoder model pretrained simultaneously on text from 101 languages and speech from 51 languages. The mSLAM team used the 600M parameter configuration of mSLAM. At fine-tuning time, they combined mSLAM's text encoder, which uses characters as ints input tokens, with an uninitialized RNN-T decoder (Graves, 2012) whose vocabulary was the set of all 384 phonemes appearing in the shared task data. Due to the extremely limited amount of training data for the tasks, the team found that the decoder needed to be very small. They used a single layer, with hidden dimension 8, model dimension of 16, and 4 heads. They also used a dropout rate of 0.3 and a label smoothing of 0.2.

They took an explicitly multilingual approach to modeling the G2P tasks, fine-tuning and evaluating a single model that covered all languages in the task. Having a single model for all languages made it necessary to tell the model, for each input, which language it was generating the pronunciation for, which was accomplished by prefixing each input string with the language's three-letter code (followed by a single space).

NFST Lin (2022) proposed a universal

Language	Baseline	Tü-G2P-1	-2	-3	-4	-5	Hammond	mSLAM
BEN	91.78	82.19	89.04	89.04	83.56	83.56	79.45	-
BUR	-	92.00	90.00	93.00	86.00	86.00	89.00	-
GER	97.00	79.00	74.00	74.00	74.00	74.00	85.00	-
GLE	-	78.00	74.00	80.00	81.00	81.00	85.00	-
ITA	44.00	41.00	41.00	38.00	40.00	40.00	41.00	-
PES	-	80.70	100.00	78.95	82.46	82.46	82.46	-
SWE	80.00	82.00	77.00	80.00	74.00	74.00	81.00	-
TGL	30.00	50.00	40.00	68.00	92.00	92.00	37.00	-
THA	-	91.00	83.00	81.00	94.00	94.00	91.00	-
UKR	96.00	77.00	74.00	76.00	92.00	92.00	86.00	-
Macro-average	83.48	75.29	74.20	75.80	79.90	79.90	75.69	-

Table 2: Results from the cross-lingual transfer subtask.

Language	Baseline	Tü-G2P-1	-2	-3	-4	-5	Hammond	mSLAM
BEN	-	90.41	83.56	83.56	86.30	91.78	91.78	-
BUR	-	90.00	87.00	86.00	87.00	95.00	93.00	-
GER	-	81.00	83.00	84.00	82.00	89.00	90.00	-
GLE	-	78.00	76.00	76.00	79.00	86.00	93.00	-
ITA	51.00	44.00	49.00	51.00	45.00	48.00	50.00	-
PES	-	75.44	80.70	85.96	82.46	80.70	80.70	-
SWE	79.00	84.00	81.00	81.00	81.00	86.00	82.00	-
TGL	29.00	40.00	35.00	37.00	32.00	42.00	24.00	-
THA	-	91.00	84.00	83.00	86.00	96.00	95.00	-
UKR	-	73.00	79.00	80.00	77.00	84.00	96.00	-
Macro-average	85.20	74.68	73.83	74.75	73.78	79.85	79.55	-

Table 3: Results from the very low resource subtask.

Language	Baseline	Tü-G2P-1	-2	-3	-4	-5	Hammond	mSLAM
BEN	67.12	68.49	72.60	69.86	68.49	71.23	71.23	-
BUR	29.00	37.00	31.00	37.00	35.00	51.00	46.00	-
GER	42.00	50.00	50.00	45.00	46.00	47.00	48.00	-
GLE	38.00	33.00	35.00	37.00	36.00	39.00	56.00	-
ITA	15.00	19.00	18.00	18.00	19.00	15.00	29.00	-
PES	59.65	57.89	100.00	57.89	56.14	61.40	59.65	-
SWE	45.00	54.00	53.00	51.00	52.00	51.00	62.00	-
TGL	20.00	15.00	16.00	18.00	15.00	14.00	16.00	-
THA	21.00	39.00	38.00	36.00	35.00	57.00	71.00	-
UKR	32.00	36.00	41.00	39.00	44.00	41.00	53.00	-
Macro-average	36.88	40.94	45.46	40.88	40.66	44.76	51.19	-

Table 4: Results from the low resource subtask.

grapheme-to-phoneme transduction model using neuralized finite-state transducers (NFST; Lin et al., 2019), a generalization of weighted finite-state transducers (WFSTs). The submission was not received by the published deadline. In fairness to other participants, scores are not listed.

7 Results

Overall, teams were able to outperform the baseline in the cross-lingual and very-low resource settings, at best achieving relative reductions of word error rate of 14% in the cross-lingual subtask and 14% in the very-low resource subtask. The best results for each setting are given in Tables 2 to 4. Non-neural approaches like HMMs with data augmentation were particularly successful in regimes where Transformer models often founder, mirroring findings in machine translation and morphological inflection (McCarthy et al., 2019).

7.1 Error analysis

Error analysis can help identify strengths and weaknesses of existing models, suggesting future improvements and guiding the construction of ensemble models. Prior experience using gold crowd-sourced data extracted from Wiktionary suggests that a non-trivial portion of errors made by top systems are due to errors in the gold data itself. For example, Gorman et al. (2019) report that a substantial portion of the prediction errors made by the top two systems in the 2017 CoNLL–SIGMORPHON shared task on morphological reinflection[3] are due to target errors, i.e., errors in the gold data. (These observations led to the development of cleaner data in UniMorph 3.0 (McCarthy et al., 2020a).)

To facilitate ensemble construction and further error analysis, we release all submissions' test set predictions to the research community.[4]

8 Discussion

We once again see an enormous difference in language difficulty. In particular, Hammond (2022) provides examples from the Welsh/Irish language pair to suggest that phylogenic or hegemonic similarity of languages does not entail similarity of orthography and phonology. Moreover, phoneme OOVs were a problem in the very-low resource setting: many phonemes and phenomena were simply not observed in 100 randomly sampled examples. This suggest room for typological information to improve modeling.

As mentioned above, the data here are a mixture of broad and narrow transcriptions. At first glance, this might explain some of the variation in language difficulty; for example, it is easy to imagine that the additional details in narrow transcriptions make them more difficult to predict. However, for many languages, only one of the two levels of transcription is available at scale, and other languages, divergence between broad and narrow transcriptions is impressionistically quite minor, as asserted in Ashby et al. (2021). However, this impression ought to be quantified.

The inclusion of the very-low resource subtask is intended to be a challenging case for participants; however, we did not anticipate the degree to which it would be challenging. In many cases, the baseline and participants' systems achieve a word error rate of zero or one. Clearly, there is room for improvement in minimally supervised grapheme-to-phoneme conversion.

Participants were permitted in all three subtasks to make use of lemmas and morphological tags from UniMorph as additional features. However, no team made use of these resources. Some prior work (e.g., Demberg et al., 2007) has found morphological tags highly useful, and Ashby et al. (2021) suggests this information would make an impact in French.

The results of the shared task suggest several next steps for carrying out a g2p shared task:

1. Split evaluation into frequent and infrequent test sets, as infrequent words may exhibit greater regularity.
2. Evaluate downstream performance for ASR.
3. Provide pointers to linguistic resources detailing phylogenic/hegemonic relationships, etc.

9 Conclusion

The third iteration of the shared task on multilingual grapheme-to-phoneme conversion is structured to provide answers to questions about the value of cross-lingual transfer and data availability.

Three teams submitted fifteen systems, achieving substantial reductions in both absolute and relative error over the baseline in two of three subtasks. We hope the code and data, released under permissive licenses,[5] will be used to benchmark grapheme-to-phoneme conversion and sequence-to-sequence modeling techniques more generally—especially in challenging low-resource scenarios.

[3] https://sigmorphon.github.io/sharedtasks/2017/

[4] https://drive.google.com/drive/folders/1qXKjMqtlgtNtT38o2uSZozLlo-7F_R0w?usp=sharing

[5] https://github.com/sigmorphon/2022G2PST

Acknowledgments

Kyle Gorman served as a consultant in the design of this task. We are grateful for his service. We thank Peter Makarov for discussions relating to the baseline model. We also thank the many Wiktionary contributors whose efforts made this task possible. A.D.M. is supported by an Amazon Fellowship.

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

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text.

Vera Demberg, Helmut Schmid, and Gregor Möhler. 2007. Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 96–103, Prague, Czech Republic. Association for Computational Linguistics.

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

Dan Garrette. 2022. Fine-tuning mSLAM for the SIGMORPHON 2022 shared task on grapheme-to-phoneme conversion. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Seattle, USA. Association for Computational Linguistics. Non-archival; abstract only.

Leander Girrbach. 2022. SIGMORPHON 2022 shared task on grapheme-to-phoneme conversion submission description: Sequence labelling for g2p. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Seattle, USA. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Mike Hammond. 2022. Low-resource grapheme-to-phoneme mapping with phonetically-conditioned transfer. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Seattle, USA. Association for Computational Linguistics.

R. Hwa, Philip Resnik, and Amy Weinberg. 2005. Breaking the resource bottleneck for multilingual parsing.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.

Chu-Cheng Lin. 2022. A future for universal grapheme-phoneme transduction modeling with neuralized finite-state transducers. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Seattle, USA. Association for Computational Linguistics. Non-archival; abstract only.

Chu-Cheng Lin, Hao Zhu, Matthew R. Gormley, and Jason Eisner. 2019. Neural finite-state transducers: Beyond rational relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 272–283, Minneapolis, Minnesota. Association for Computational Linguistics.

Hairong Liu, Zhenyao Zhu, Xiangang Li, and Sanjeev Satheesh. 2017. Gram-CTC: Automatic unit selection and target decomposition for sequence labelling. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2188–2197. JMLR.org.

Peter Makarov and Simon Clematide. 2018. Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877–2882, Brussels, Belgium. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2020. CLUZH at 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 171–176, Online. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020a. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong. 2020b. Addressing posterior collapse with mutual information for improved variational neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8512–8525, Online. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Steven Moran and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020. Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.

Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49, Donostia–San Sebastián. Association for Computational Linguistics.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D. McCarthy, and Deepak Gopinath. 2019. Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Zach Ryan and Mans Hulden. 2020. Data augmentation for transformer-based G2P. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 184–188, Online. Association for Computational Linguistics.

Daan van Esch, Mason Chua, and Kanishka Rao. 2016. Predicting pronunciations with syllabification and stress with recurrent neural networks. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 2841–2845. ISCA.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.