Improved pronunciation prediction accuracy using morphology

Dravyansh Sharma, Saumya Yashmohini Sahai, Neha Chaudhari[†], Antoine Bruguier
[†]Google LLC[*]
dravyans@andrew.cmu.edu, sahai.17@osu.edu,
neha7.chaudhari@gmail.com, bruguier@almuni.caltech.edu

Abstract

Pronunciation lexicons and prediction models are a key component in several speech synthesis and recognition systems. We know that morphologically related words typically follow a fixed pattern of pronunciation which can be described by language-specific paradigms. In this work we explore how deep recurrent neural networks can be used to automatically learn and exploit this pattern to improve the pronunciation prediction quality of words related by morphological inflection. We propose two novel approaches for supplying morphological information, using the word's morphological class and its lemma, which are typically annotated in standard lexicons. We report improvements across a number of European languages with varying degrees of phonological and morphological complexity, and two language families, with greater improvements for languages where the pronunciation prediction task is inherently more challenging. We also observe that combining bidirectional LSTM networks with attention mechanisms is an effective neural approach for the computational problem considered, across languages. Our approach seems particularly beneficial in the low resource setting, both by itself and in conjunction with transfer learning.

1 Introduction

Morphophonology is the study of interaction between morphological and phonological processes and mostly involves description of sound changes that take place in morphemes (minimal meaningful units) when they combine to form words. For example, the plural morpheme in English appears as '-s' or '-es' in orthography and as [s], [z], and [ɪz] in phonology, e.g. in *cops*, *cogs* and *courses*. The different forms can be thought to be derived from a common plural morphophoneme which undergoes context dependent transformations to produce the correct phones.

A *pronunciation model*, also known as a grapheme to phoneme (G2P) converter, is a system that produces a phonemic representation of a word from its written form. The word is converted from the sequence of letters in the orthographic script to a sequence of phonemes (sound symbols) in a pre-determined transcription, such as IPA or X-SAMPA. It is expensive and possibly, say in morphologically rich languages with productive compounding, infeasible to list the pronunciations for all the words. So one uses rules or learned models for this task. Pronunciation models are important components of both speech recognition (ASR) and synthesis (text-to-speech, TTS) systems. Even though end-to-end models have been gathering recent attention (Graves and Jaitly, 2014; Sotelo et al., 2017), often state-of-the-art models in industrial production systems involve conversion to and from an intermediate phoneme layer.

A single system of morphophonological rules which connects morphology with phonology is well-known (Chomsky and Halle, 1968). In fact computational models for morphology such as the two-level morphology of Koskenniemi (1983); Kaplan and Kay (1994) have the bulk of the machinery designed to handle phonological rules. However, the approach involves encoding language-specific rules as a finite-state transducer, a tedious and expensive process requiring linguistic expertise. Linguistic rules are augmented computationally for small corpora in Ermolaeva (2018), although scalability and applicability of the approach across languages is not tested.

We focus on using deep neural models to improve the quality of pronunciation prediction using

[*] Part of the work was done when D.S., N.C. and A.B. were at Google.

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology,pages 222–228
August 5, 2021. ©2

morphology. G2P fits nicely in the well-studied sequence to sequence learning paradigms (Sutskever et al., 2014), here we use extensions that can handle supplementary inputs in order to inject the morphological information. Our techniques are similar to Sharma et al. (2019), although the goal there is to lemmatize or inflect more accurately using pronunciations. Taylor and Richmond (2020) consider improving neural G2P quality using morphology, our work differs in two respects. First, we use morphology class and lemma entries instead of morpheme boundaries for which annotations may not be as readily available. Secondly, they consider BiLSTMs and Transformer models, but we additionally consider architectures which combine BiLSTMs with attention and outperform both. We also show significant gains by morphology injection in the context of transfer learning for low resource languages where sufficient annotations are unavailable.

2 Background and related work

Pronunciation prediction is often studied in settings of speech recognition and synthesis. Some recent work explores new representations (Livescu et al., 2016; Sofroniev and Çöltekin, 2018; Jacobs and Mailhot, 2019), but in this work, *a pronunciation is a sequence of phonemes, syllable boundaries and stress symbols* (van Esch et al., 2016). A lot of work has been devoted to the G2P problem (e.g. see Nicolai et al. (2020)), ranging from those focused on accuracy and model size to those discussing approaches for data-efficient scaling to low resource languages or multilingual modeling (Rao et al., 2015; Sharma, 2018; Gorman et al., 2020).

Morphology prediction is of independent interest and has applications in natural language generation as well as understanding. The problems of lemmatization and morphological inflection have been studied in both contextual (in a sentence, which involves morphosyntactics) and isolated settings (Cohen and Smith, 2007; Faruqui et al., 2015; Cotterell et al., 2016; Sharma et al., 2019).

Morphophonological prediction, by which we mean viewing morphology and pronunciation prediction as a single task with several related inputs and outputs, has received relatively less attention as a language-independent computational task, even though the significance for G2P has been argued (Coker et al., 1991). Sharma et al. (2019) show improved morphology prediction using phonology, and Taylor and Richmond (2020) show the reverse. The present work aligns with the latter, but instead of requiring full morphological segmentation of words we work with weaker and more easily annotated morphological information like word lemmas and morphological categories.

3 Improved pronunciation prediction

We consider the G2P problem, i.e. prediction of the sequence of phonemes (pronunciation) from the sequence of graphemes in a single word. The G2P problem forms a clean, simple application of seq2seq learning, which can also be used to create models that achieve state-of-the-art accuracies in pronunciation prediction. Morphology can aid this prediction in several ways. One, we could use morphological category as a non-sequential side input. Two, we could use the knowledge of the morphemes of the words and their pronunciations which may be possible with lower amounts of annotation. For example, the lemma (and its pronunciation) may already be annotated for an out-of-vocabulary word. Often standard lexicons list the lemmata of derived/inflected words, lemmatizer models can be used as a fallback. Learning from the exact morphological segmentation (Taylor and Richmond, 2020) would need more precise models and annotation (Demberg et al., 2007).

Given the spelling, language specific models can predict the pronunciation by using knowledge of typical grapheme to phoneme mappings in the language. Some errors of these models may be fixed with help from morphological information as argued above. For instance, homograph pronunciations can be predicted using morphology but it is impossible to deduce correctly using just orthography.[1] The pronunciation of 'read' (/ɹiːd/ for present tense and noun, /ɹɛd/ for past and participle) can be determined by the part of speech and tense; the stress shifts from first to second syllable between 'project' noun and verb.

3.1 Dataset

We train and evaluate our models for five languages to cover some morphophonological diversity: (American) English, French, Russian, Spanish and Hungarian. For training our models, we use pronunciation lexicons (word-pronunciation pairs) and morphological lexicons (containing *lex-*

[1] Homographs are words which are spelt identically but have different meanings and pronunciations.

ical form, i.e. lemma and morphology class) of only inflected words of size of the order of 10^4 for each language (see Table 5 in Appendix A). For the languages discussed, these lexicons are obtained by scraping[2] Wiktionary data and filtering for words that have annotations (including pronunciations available in the IPA format) for both the *surface form* and the *lexical form*. While this order of data is often available for high-resource languages, in Section 3.3 we discuss extension of our work to low-resource settings using Finnish and Portuguese for illustration where the Wiktionary data is about an order of magnitude smaller.

Word (language)	Morph. Class	Pron.	LS	LP
masseuses (fr)	n-f-pl	/ma.søz/	masseur	/ma.sœʁ/
fagylaltozom (hu)	v-fp-s-in-pr-id	/ˈfɒɟlɒltozom/	fagylaltozik	/ˈfɒɟlɒltozik/

Table 1: Example annotated entries. (v-fp-s-in-pr-id: Verb, first-person singular indicative present indefinite)

We keep 20% of the pronunciation lexicons aside for evaluation using word error rate (WER) metric. WER measures an output as correct if the entire output pronunciation sequence matches the ground truth annotation for the test example.

3.1.1 Morphological category

The morphological category of the word is appended as an ordinal encoding to the spelling, separated by a special character. That is, the categories of a given language are appended as unique integers, as opposed to one-hot vectors which may be too large in morphologically rich languages.

3.1.2 Lemma spelling and pronounciation

Information about the lemma is given to the models by appending both, the lemma pronounciation ⟨LP⟩ and lemma spelling ⟨LS⟩ to the word spelling ⟨WS⟩, all separated by special characters, like, ⟨WS⟩§⟨LP⟩¶⟨LS⟩. Lemma spelling can potentially help in irregular cases, for example 'be' has past forms 'gone' and 'were', so the model can reject the lemma pronunciation in this case by noting that the lemma spellings are different (but potentially still use it for 'been').

3.2 Model details

The models described below are implemented in OpenNMT (Klein et al., 2017).

3.2.1 Bidirectional LSTM networks

LSTM (Hochreiter and Schmidhuber, 1997) allows learning of fixed length sequences, which is not a major problem for pronunciation prediction since grapheme and phoneme sequences (represented as one-hot vectors) are often of comparable length, and in fact state-of-the-art accuracies can be obtained using bidirectional LSTM (Rao et al., 2015). We use single layer BiLSTM encoder - decoder with 256 units and 0.2 dropout to build a character level RNN. Each character is represented by a trainable embedding of dimension 30.

3.2.2 LSTM based encoder-decoder networks with attention (BiLSTM+Attn)

Attention-based models (Vaswani et al., 2017; Chan et al., 2016; Luong et al., 2015; Xu et al., 2015) are capable of taking a weighted sample of input, allowing the network to focus on different possibly distant relevant segments of the input effectively to predict the output. We use the model defined in Section 3.2.1 with Luong attention (Luong et al., 2015).

3.2.3 Transformer networks

Transformer (Vaswani et al., 2017) uses self-attention in both encoder and decoder to learn rich text representaions. We use a similar architecture but with fewer parameters, by using 3 layers, 256 hidden units, 4 attention heads and 1024 dimensional feed forward layers with relu activation. Both the attention and feedforward dropout is 0.1. The input character embedding dimension is 30.

3.3 Transfer learning for low resource G2P

Both non-neural and neural approaches have been studied for transfer learning (Weiss et al., 2016) from a high-resource language for low resource language G2P setting using a variety of strategies including semi-automated bootstrapping, using acoustic data, designing representations suitable for neural learning, active learning, data augmentation and multilingual modeling (Maskey et al., 2004; Davel and Martirosian, 2009; Jyothi and Hasegawa-Johnson, 2017; Sharma, 2018; Ryan and Hulden, 2020; Peters et al., 2017; Gorman et al., 2020). Recently, transformer-based architectures have also been used for this task (Engelhart et al., 2021). Here we apply a similar approach of using representations learned from the high-resource languages as an additional input for low-resource models but for our BiLSTM+Attn architecture. We

[2] kaikki.org/dictionary/

Model	Inputs	*en*	*fr*	*ru*	*es*	*hu*
BiLSTM	(b/+c/+l)	(39.7/39.4/37.1)	(8.69/8.94/7.94)	(5.26/4.87/5.60)	(1.13/1.44/1.30)	(6.96/5.85/7.21)
BiLSTM+Attn	(b/+c/+l)	(36.9/36.1/**31.0**)	(4.45/4.20/**4.12**)	(5.06/**3.80**/4.04)	(0.32/0.32/**0.29**)	(1.78/1.31/**1.12**)
Transformer	(b/+c/+l)	(40.2/39.3/37.7)	(8.19/7.11/10.6)	(6.57/6.38/5.36)	(2.29/1.62/2.20)	(8.20/4.93/8.11)

Table 2: Models and their Word Error Rates (WERs). 'b' corresponds to baseline (vanilla G2P), '+c' refers to morphology class injection (Sec. 3.1.1) and '+l' to addition of lemma spelling and pronunciation (Sec. 3.1.2).

evaluate our model for two language pairs — *hu* (high) - *fi* (low) and *es* (high) and *pt* (low) (results in Table 3). We perform morphology injection using lemma spelling and pronunciation (Sec. 3.1.2) since it can be easier to annotate and potentially more effective (per Table 2). *fi* and *pt* are not really low-resource, but have relatively fewer Wiktionary annotations for the lexical forms (Table 5).

Model	*fi*	*fi+hu*	*pt*	*pt+es*
BiLSTM+Attn (base)	18.53	9.81	62.65	58.87
BiLSTM+Attn (+lem)	9.27	**8.45**	59.63	**55.48**

Table 3: Transfer learning for vanilla G2P (base) and morphology augmented G2P (+lem, Sec. 3.1.2).

4 Discussion

We discuss our results under two themes — the efficacy of the different neural models we have implemented, and the effect of the different ways of injecting morphology that were considered.

We consider three neural models as described above. To compare the neural models, we first note the approximate number of parameters of each model that we trained:

- BiLSTM: ∼1.7M parameters,
- BiLSTM+Attn: ∼3.5M parameters,
- Transformer: ∼5.2M parameters.

For BiLSTM and BiLSTM+Attn, the parameter size is based on neural architecture search i.e. we estimated sizes at which accuracies (nearly) peaked. For transformer, we believe even larger models can be more effective and the current size was chosen due to computational restrictions and for "fairer" comparison of model effectiveness. Under this setting, BiLSTM+Attn models seem to clearly outperform both the other models, even without morphology injection (cf. Gorman et al. (2020), albeit it is in the multilingual modeling context). Transformer can beat BiLSTM in some cases even with the sub-optimal model size restriction, but is consistently worse when the sequence lengths are larger which is the case when we inject lemma spellings and pronunciations.

We also look at how adding lexical form information, i.e. morphological class and lemma, helps with pronunciation prediction. We notice that the improvements are particularly prominent when the G2P task itself is more complex, for example in English. In particular, ambiguous or exceptional grapheme subsequence (e.g. `ough` in English) to phoneme subsequence mappings, may be resolved with help from lemma pronunciations. Also morphological category seems to help for example in Russian where it can contain a lot of information due to the inherent morphological complexity (about 25% relative error reduction). See Appendix B for more detailed comparison and error analysis for the models.

Our transfer learning experiments indicate that morphology injection gives even more gains in low resource setting. In fact for both the languages considered, adding morphology gives almost as much gain as adding a high resource language to the BiLSTM+Attn model. This could be useful for low resource languages like Georgian where a high resource language from the same language family is unavailable. Even with the high resource augmentation, using morphology can give a significant further boost to the prediction accuracy.

5 Conclusion

We note that combining BiLSTM with attention seems to be the most attractive alternative in getting improvements in pronunciation prediction by leveraging morphology, and hence correspond to the most appropriate 'model bias' for the problem from among the alternatives considered. We also note that all the neural network paradigms discussed are capable of improving the G2P prediction quality when augmented with morphological information. Since our approach can potentially support partial/incomplete data (using appropriate ⟨`MISSING`⟩ or ⟨`N/A`⟩ tokens), one can use a single model which injects morphology class and/or lemma pronunciation as available. For languages where neither is available, our results suggest building word-lemma lists or utilizing effective lemmatizers (Faruqui et al., 2015; Cotterell et al., 2016).

6 Future work

Our work only leverages the inflectional morphology paradigms for better pronunciation prediction. However in addition to inflection, morphology also results in word formation via derivation and compounding. Unlike inflection, derivation and compounding could involve multiple root words, so an extension would need a generalization of the above approach along with appropriate data. An alternative would be to learn these in an unsupervised way using a dictionary augmented neural network which can efficiently refer to pronunciations in a dictionary and use them to predict pronunciations of polymorphemic words using pronunciations of the base words (Bruguier et al., 2018). It would be interesting to see if using a combination of morphological side information and dictionary-augmentation results in a further accuracy boost. Developing non-neural approaches for the morphology injection could be interesting, although as noted before, the neural approaches are the state-of-the-art (Rao et al., 2015; Gorman et al., 2020).

One interesting application of the present work would be to use the more accurate pronunciation prediction for morphologically related forms for efficient pronunciation lexicon development (useful for low resource languages where high-coverage lexicons currently don't exist), for example annotating the lemma pronunciation should be enough and the pronunciation of all the related forms can be predicted with high accuracy. This is hugely beneficial for languages where there are hundreds or even thousands of surface forms associated with the same lemma. Another concern for reliably using the neural approaches is explainability (Molnar, 2019). Some recent research looks at explaining neural models with orthographic and phonological features (Sahai and Sharma, 2021), an extension for morphological features should be useful.

References

Antoine Bruguier, Anton Bakhtin, and Dravyansh Sharma. 2018. Dictionary Augmented Sequence-to-Sequence Neural Network for Grapheme to Phoneme prediction. *Proc. Interspeech 2018*, pages 3733–3737.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4960–4964. IEEE.

Noam Chomsky and Morris Halle. 1968. The sound pattern of English.

Shay B Cohen and Noah A Smith. 2007. Joint morphological and syntactic disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Cecil H Coker, Kenneth W Church, and Maik Y Liberman. 1991. Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis. In *The ESCA Workshop on Speech Synthesis*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22.

Marelie Davel and Olga Martirosian. 2009. Pronunciation dictionary development in resource-scarce environments.

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.

Eric Engelhart, Mahsa Elyasi, and Gaurav Bharaj. 2021. Grapheme-to-Phoneme Transformer Model for Transfer Learning Dialects. *arXiv preprint arXiv:2104.04091*.

Marina Ermolaeva. 2018. Extracting morphophonology from small corpora. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 167–175.

Daan van Esch, Mason Chua, and Kanishka Rao. 2016. Predicting Pronunciations with Syllabification and Stress with Recurrent Neural Networks. In *INTERSPEECH*, pages 2841–2845.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2015. Morphological inflection generation using character sequence to sequence learning. *arXiv preprint arXiv:1512.06110*.

Kyle Gorman, Lucas FE Ashby, Aaron Goyzueta, Arya D 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.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Cassandra L Jacobs and Fred Mailhot. 2019. Encoder-decoder models for latent phonological representations of words. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 206–217.

Preethi Jyothi and Mark Hasegawa-Johnson. 2017. Low-resource grapheme-to-phoneme conversion using recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5030–5034. IEEE.

Ronald M Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Kimmo Koskenniemi. 1983. Two-Level Model for Morphological Analysis. In *IJCAI*, volume 83, pages 683–685.

Karen Livescu, Preethi Jyothi, and Eric Fosler-Lussier. 2016. Articulatory feature-based pronunciation modeling. *Computer Speech & Language*, 36:212–232.

Minh-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.

Sameer Maskey, Alan Black, and Laura Tomokiya. 2004. Boostrapping phonetic lexicons for new languages. In *Eighth International Conference on Spoken Language Processing*.

Christoph Molnar. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Garrett Nicolai, Kyle Gorman, and Ryan Cotterell. 2020. Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively Multilingual Neural Grapheme-to-Phoneme Conversion. *EMNLP 2017*, page 19.

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

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.

Saumya Sahai and Dravyansh Sharma. 2021. Predicting and explaining french grammatical gender. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 90–96.

Dravyansh Sharma. 2018. On Training and Evaluation of Grapheme-to-Phoneme Mappings with Limited Data. *Proc. Interspeech 2018*, pages 2858–2862.

Dravyansh Sharma, Melissa Wilson, and Antoine Bruguier. 2019. Better Morphology Prediction for Better Speech Systems. In *INTERSPEECH*, pages 3535–3539.

Pavel Sofroniev and Çağrı Çöltekin. 2018. Phonetic vector representations for sound sequence alignment. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 111–116.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Jason Taylor and Korin Richmond. 2020. Enhancing Sequence-to-Sequence Text-to-Speech with Morphology. *Submitted to IEEE ICASSP*.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Model	Inputs	en	de	es	ru	avg. rel. gain
BiLSTM	(b/+c/+l)	(31.0/30.5/25.2)	(17.7/15.5/12.3)	(8.1/7.9/6.7)	(18.4/15.6/15.9)	(-/+7.9%/+20.0%)
BiLSTM+Attn	(b/+c/+l)	(29.0/27.1/21.3)	(12.0/11.6/11.6)	(4.9/2.6/2.4)	(14.1/13.6/13.1)	(-/+15.1%/+22.0%)

Table 4: Number of total Wiktionary entries, and inflected entries with pronunciation and morphology annotations, for the languages considered.

Appendix

A On size of data

We record the size of data scraped from Wiktionary in Table 5. There is marked inconsistency in the number of annotated inflected words where the pronunciation transcription is available, as a fraction of the total vocabulary, for the languages considered.

In the main paper, we have discussed results on the publicly available Wiktionary dataset. We perform more experiments on a larger dataset (10^5-10^6 examples of annotated inflections per language) using the same data format and methodology for (American) English, German, Spanish and Russian (Table 4). We get very similar observations in this regime as well in terms of relative gains in model performances using our techniques, but these results are likely more representative of word error rates for the whole languages.

Language	Total senses	Annotated inflections
en	1.25M	7543
es	0.93M	28495
fi	0.24M	3663
fr	0.46M	24062
hu	77.7K	31486
pt	0.39M	2647
ru	0.47M	20558

Table 5: Number of total Wiktionary entries, and inflected entries with pronunciation and morphology annotations, for the languages considered.

B Error analysis

Neural sequence to sequence models, while highly accurate on average, make “silly” mistakes like omitting or inserting a phoneme which are hard to explain. With that caveat in place, there are still reasonable patterns to be gleaned when comparing the outputs of the various neural models discussed here. BiLSTM+Attn model seems to not only be making fewer of these “silly” mistakes, but also appears to be better at learning the genuinely more challenging predictions. For example, the French word *pédagogiques* (‘pedagogical’, plural) /pe.da.gɔ.ʒik/ is pronounced correctly by BiLSTM+Attn, but as /pe.da.ʒɔ.ʒik/ by BiLSTM. Similarly BiLSTM+Attn predicts /ˈdʒæmɪŋ/, while Transformer network says /ˈdʒamɪŋ/ for *jamming* (en). We note that errors for Spanish often involve incorrect stress assignment since the grapheme-to-phoneme mapping is highly consistent.

Adding morphological class information seems to reduce the error in endings for morphologically rich languages, which can be an important source of error if there is relative scarcity of transcriptions available for the inflected words. For example, for our BiLSTM+Attn model, the pronunciation for фуррем (ru, ‘furry’ instrumental singular noun) is fixed from /ˈfurʲːem/ to /ˈfurʲːɪm/, and *koronavírusról* (hu, ‘coronavirus’ delative singular) gets corrected from /ˈkoronɒviːruʃoːl/ to /ˈkoronɒviːruʃroːl/. On the other hand, adding lemma pronunciation usually helps with pronouncing the root morpheme correctly. Without the lemma injection, our BiLSTM+Attn model mispronounces *debriefing* (en) as /dɪˈbɹiːfɪŋ/ and *sentences* (en) as /sɛnˈtɛnsɪz/. Based on these observations, it sounds interesting to try to inject both categorical and lemma information simultaneously.