Can phylogenetic and areal information improve multilingual neural grapheme-to-phoneme conversion?

Chundra A. Cathcart
Department of Comparative Language Science
University of Zurich
Plattenstrasse 54
CH-8032 Zürich
chundra.cathcart@uzh.ch

Abstract

This paper seeks to assess whether phylogenetic and areal information can improve results in a multilingual neural grapheme-to-phoneme conversion task. We find that models which incorporate embeddings corresponding to languages' identity as well as their genetic and geographic profile underperform a baseline that incorporates a language embedding alone, but this result is largely restricted to languages with Latin orthography. Conversely, phylogenetic and geographic embeddings improve performance for languages from South and Southeast Asia. We discuss possible reasons for this discrepancy as well a steps to be taken in future directions which make use of extra-linguistic information of this sort in NLP tasks such as g2p.

1 Introduction

Neural multilingual tasks in natural language processing such as grapheme-to-phoneme conversion show good performance due to their ability to capture global cross-linguistic tendencies as well as language-specific trends. Embeddings corresponding to individual languages found in training data have been shown to improve multilingual neural grapheme-to-phoneme conversion (g2p). The potential for improving g2p using other linguistic or extra-linguistic features remains relatively unexplored. This paper seeks to assess the degree to which the inclusion of phylogenetic and areal information can improve multilingual neural g2p. By explicitly encoding our assumptions into the model architecture, we seek to isolate the effect of each of the two phenomena, geographic proximity and genetic relatedness, on the task we perform.

We use a state-of-the-art neural encoder-decoder with a hard attention mechanism that incorporates a language model on the target side; a good model of phonotactics is therefore critical to high performance. We believe that grapheme-to-phoneme mappings are purely a byproduct of sociocultural and geopolitical contingencies, better captured by areal proximity between languages; furthermore, we take into account evidence from the literature on historical linguistic and evolutionary linguistics demonstrating that phonotactic patterns across languages carry phylogenetic signal. For this reason, we take the view that areal and phylogenetic information is best used in the neural g2p setting by feeding areal information along with language-level embeddings to the encoder, and feeding phylogenetic information along with language-level embeddings to the decoder.

We find that the inclusion of this phylogenetic and areal information results in an overall *decrease* in model performance over a baseline that includes information regarding the language to which a grapheme-phoneme sequence pair belongs. However, we find that this decrease in performance is virtually limited to data points with Latin orthography, and that the inclusion of phylogenetic and areal data tends to improve results for lower-resource languages from the same genetic group and/or area, namely South and Southeast Asian languages. We discuss possible reasons for this discrepancy, as well as ways to extend insights from this study in order to improve g2p in low-resource languages.

2 Background

Attention-based long short-term memory (LSTM, Hochreiter and Schmidhuber 1997)

encoder-decoder models achieve state-of-the-art results in monolingual and multilingual approaches g2p and transliteration, largely outperforming non-neural methods (Peters et al., 2017; Kunchukuttan et al., 2018). In the context of low-resource g2p, multilingual methods aim to leverage similarity across languages in order to combine useful global, cross-linguistic trends with language specific idiosyncrasies in order to improve the accuracy of conversion. The approach of Peters et al. (2017) achieves state-of-the-art results by concatenating a language embedding to the input, a practice which improves model performance over models with no language embedding.

Beyond the level of language, the role of other information in neural g2p remains an open question. Peters et al. (2017) found that language embeddings learned by their g2p model showed only limited phylogenetic signal, in contrast to language vectors learned from larger corpora (Östling and Tiedemann, 2017). Similarly, Gutkin and Sproat (2017) find limited and ultimately inconclusive support for the idea that areal and phylogenetic features improve multilingual speech synthesis. In non-neural approaches to g2p, Kim and Snyder (2012) found that the inclusion of phylogenetic and areal information did not greatly improve model performance, compared to linguistic features such as phonemic context features.

It is not surprising that typological features related to sound patterns improve g2p, given the usefulness of typological features in a large number of NLP applications (Søgaard and Wulff, 2012; Rama and Kolachina, 2012; Zhang and Barzilay, 2015; Agić, 2017). Given that areal, genetic and typological similarity are correlated, it is surprising that this information has not improved results, since a rich genetic and areal representation can serve as a proxy to a language's typological profile. Furthermore, genetic and areal information may give us a window to the diachronic processes which have shaped the grapheme-to-phoneme correspondences observed today, and this information may prove valuable in capturing patterns across data.

One possibility as to why areal and phylogenetic features did not lead to major improvements in the works cited is that insufficiently granular representations of areality and phylogeny were used. Kim and Snyder (2012) use language family features from WALS (Dryer and Haspelmath, 2013), which consist of a large family (e.g., Indo-European) and small family (e.g., Germanic) in order to encode family- and region-specific information. Similarly, Gutkin and Sproat (2017) use only longitude and latitude to represent the areal position of each language in their sample. In contrast, resources like URIEL (Littell et al., 2017) provide a much more fine-grained representation of languages' genetic and areal profiles, making use of highly articulated genetic information from Glottolog (Hammarström et al., 2017) and measuring each language's distance to several hundred fixed points on the globe, which helps to capture information regarding the region over which the language is spoken.

3 Rationale

Our goal is to assess the benefit of using phylogenetic and areal information in a multilingual g2p task. We use the hard nonmonotonic attention mechanism of Wu et al. (2018) for the multilingual experiments in this study. This model outperforms existing soft attention mechanisms (e.g., Bahdanau et al. 2014 and Luong et al. 2015, employed by Peters et al. 2017) on g2p and related tasks. Wu and Cotterell (2019) extend this architecture to include strict monotonicity constraints, which brings about improved performance on English g2p.

We refrain from using any sort of monotonicity constraint because the dataset of Deri and Knight (2016), which we use, contains writing systems where the grapheme-to-phoneme mapping is monotonically decreasing as well as non-monotonic. The former case, represented by right-to-left writing systems like Arabic and Hebrew, can be dealt with via the simple preprocessing step of reversing the input sequence. The latter case is represented by South Asian writing systems like Hindi, where the grapheme representing the vowel in the phonemic sequence /kɪ/ proceeds the grapheme corresponding to the consonant. In theory, a monotonic alignment can learn a

mapping between the pair of graphemes and the pair of phonemes, but may not learn which individual grapheme is responsible for generating which phoneme, potentially leading to an non-parsimonious encoding of information if not decreases in accuracy on test data.

The attention mechanism we employ has the added benefit of incorporating a target-side language model; this language model is crucial to good decoder performance. This architectural feature means that not all additional features concatenated to the linguistic input must be fed solely to the LSTM encoder; they can be fed directly to the decoder as well. We suspect that this ability to de-couple sources of information will lead to better use of the features we seek to encode.

Ultimately, we do not think that it makes sense to feed both areal and phylogenetic information to the model encoder. The task of the encoder is to learn a latent representation that can be used to generate the output conditional on the input. The mapping of orthographic symbols to phonemes, in our opinion, is purely a product of geopolitical, sociocultural and religious history (Baddeley and Voeste, 2012). Languages' geopolitical and sociocultural histories are often strongly correlated with their phylogenetic positions; however, phylogeny is at best only distally related to the processes that shape languages' orthographies. Areal position is distally related as well, but less so; in most cases, the phylogenetic spread of languages at a large scale pre-dates the diffusion of writing systems between them (Falk, 1993; Salomon, 1998). For this reason, we believe that the task carried out by the encoder, i.e., learning a latent representation for each grapheme in a language, should encode areal information and not genetic information, the latter being more appropriate for inclusion in the language model of the decoder. Recent work in historical linguistics suggests that high-definition phonotactic information carries a cross-linguistic phylogenetic signal (Macklin-Cordes and Round, 2015), though this result is based on relatively consistently transcribed data sets from languages from a single region of the world; if this is is the case, then it would follow that feeding a phylogenetic embedding to the neural decoder could potentially improve model performance under this architecture.

4 Data

We use the cleaned multilingual data set of Deri and Knight (2016). To maximize the potential for multilingual knowledge transfer, we exclude ideograph-based orthographies, and additionally exclude orthographic systems that correspond to only one language in the data set, unless the system shares grapheme naming conventions with other systems (e.g., virtually all South Asian writing systems with the exception of Sinhala). In general, previous authors report poor performance in multilingual g2p for languages with unique writing systems such as Georgian or Armenian. The exclusion of this data means that our results are not directly comparable to those of Peters et al. (2017); however, our primary goal is to assess the relative performance of models which incorporate phylogenetic and areal information against those which do not. We exclude all languages with fewer than 1000 observations from the training set but not the test set, with the aim of predicting outputs from unseen languages as well as held-out data from observed languages.

For the output of the decoder, we use the cleaned phonemic transcriptions of Deri and Knight (2016). For the encoder input, we use the corresponding orthographic sequences. Following Deri and Knight (2016), we collect the Unicode names of each orthographic symbol in our data set. We convert these names to a binary representation for each symbol in which a dimension is valued 1 if the corresponding element appears in the symbol's Unicode name (we exclude the element corresponding to the name of the writing system, e.g., `DEVANAGARI`). This allows us to capture similarities between symbols that have non-identical Unicode names that share elements, e.g., `THAI CHARACTER KHO KHAI` and `LAO LETTER KHO SUNG`, which correspond to the same phoneme.

We collect geographic and phylogenetic information for each language from the URIEL database (Littell et al., 2017). Geographic vectors consist of distances for each language from 299 fixed points on the Earth's sur-

face. Phylogenetic vectors contain dimensions which represent whether or not a language is present in language family or subfamily, according to Glottolog (Hammarström et al., 2017). We drop dimensions where no language is present, yielding 301 dimensions.

The data set we use contains 191461 data points on which to train the model from 388 languages. The training and test data contain 1109 unique orthographic symbols with 416 unique elements appearing across their Unicode names and 508 unique phonemic symbols. We train only on data points where the input sequence has 30 time steps or fewer.

5 Model

We employ the encoder-decoder with hard nonmonotonic attention described by Wu et al. (2018). For each input x, a latent representation $\boldsymbol{h}_j^{\text{enc}} \in \mathbb{R}^{2D}$ is learned for each time step $j \in \{1, ..., |x|\}$ via a bidirectional LSTM on the basis of the input symbol at time step j. For each output y, a latent representation $\boldsymbol{h}_i^{\text{dec}} \in \mathbb{R}^{D}$ is learned via a forward LSTM for each time step $i \in \{1, ..., |y|\}$ on the basis of the output symbol at time step $i-1$. The probability that the output is aligned with the jth input symbol at time i is equal to $\text{softmax}(\boldsymbol{h}_i^{\text{dec}\top} \boldsymbol{T} \boldsymbol{h}_j^{\text{enc}})$, where $\boldsymbol{T} \in \mathbb{R}^{D \times 2D}$ is a learned parameter. The emission probability of the output symbol at time i given such an alignment is equal to $\text{softmax}(\boldsymbol{W} \tanh(\boldsymbol{S}[\boldsymbol{h}_i^{\text{dec}}; \boldsymbol{h}_j^{\text{enc}}]))$, and is hence also dependent on the previous output symbols ($\boldsymbol{W} \in \mathbb{R}^{\Sigma_y \times 3D}$ and $\boldsymbol{S} \in \mathbb{R}^{3D \times 3D}$ are learned parameters). The model architecture allows us to marginalize over alignment probabilities in a straightforward manner.

Peters et al. (2017) incorporate language-specific information by simply prepending a token representing the language ID of each sequence fed to the decoder. This stands in contrast to approaches to embeddings and latent variable models which concatenate a representation to each token of each sequence fed to a recurrent neural network (Östling and Tiedemann, 2017; Kim et al., 2018). We choose the latter approach; perhaps naïvely, it strikes us that if the language token is appended only to the right or left of an input sequence, the forget gates of the LSTM could potentially prevent information about the language from being propagated to latent representations of the encoder at faraway time steps, unless a large negative bias is added. Concatenating this representation at every time step ensures that this information is passed to the LSTM at every time step. It is worth noting that this concatenation allows the representation learned by the LSTM's kernel weights on the basis of the input to vary across languages, but the recurrent weights and biases are held constant across languages. It may be desirable in multilingual tasks to allow these parameters to vary as well.

For all of our experiments, the output of the encoder-decoder consists of a sequence of phonemes in a one-hot encoding. The encoder input consists of a sequence of orthographic symbols in a one-hot encoding; to each one-hot vector we concatenate a binary vector representing the symbol's Unicode name. The decoder input consists of phonemes from the previous time step in a one-hot encoding. In our different experiments, different representations of extra-linguistic features are concatenated to the encoder and decoder inputs (as described above), summarized in the table below. These representations consist of 8-dimensional dense embeddings learned for each input-output pair's corresponding language ID, geographic vector, and phylogenetic vector.

	Encoder input	Decoder input
Lang	LANG	LANG
GeoGen	GEO	GEN
LangGeoGen	LANG $\oplus$ GEO	LANG $\oplus$ GEN
LangGeo	LANG $\oplus$ GEO	LANG $\oplus$ GEO
LangGen	LANG $\oplus$ GEN	LANG $\oplus$ GEN

This experimental setup allows us not only to assess the value of phylogenetic and areal information when coupled with a data point's language ID, but whether or not this information can serve as a unique identifier independent of language ID, as well as the value of each type of feature, phylogenetic and areal, on its own.

We employ a 64-dimensional hidden layer for the LSTM. For each experiment, we train the model on 90% of the training data for 15 epochs using a batch size of 256, and validate

the model on the remaining 10%. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of .001.

6 Results and Discussion

We evaluate the accuracy of our results according to the phoneme error rate (PER), which we define to be the Levenshtein distance between the predicted and target phoneme sequence divided by the length of the longer sequence, and the word error rate (WER), which we define as the proportion of predicted sequences that show any difference from the target. We average these values across data points.

Observing the overall results in Table 1, it is clear that on the whole, the inclusion of phylogenetic and geographic information in the encoder-decoder does not improve results, but in fact leads to a worsening of performance over the baseline which incorporates only a language embedding. A model including phylogenetic and geographic information fares worse in predicting phoneme sequences for unseen languages than a model in which only a language embedding is used, even though the embedding learned is somewhat meaningless on its own, as no training data are associated with it. This is similar to the finding of Peters et al. (2017), who reason that this may be due to the model learning negative associations between unusual phoneme sequences and languages in the training data, resulting in the model generating more standard phoneme sequences for unseen languages.

It comes as somewhat of a surprise that the inclusion of phylogenetic and areal information leads to an overall decrease in performance. It seems to be the case that the model has erroneously applied phonotactic information from related languages as well as spelling rules from proximate languages to contexts where this information is not appropriate. The model which employs language embeddings alone is less likely to be corrupted by this extra information. It is worth noting that the dimensions of the language, phylogenetic, and geographic embeddings are equal, and it is perhaps the case that in making them equal, we have assigned to much weight to the latter features, allowing them to overly influence model performance.

Figure 1 shows the language-level PER plotted by training sample size for each model used in our experiments. The overall higher PER for the models GeoGen (middle panel) and LangGen (rightmost panel), which make use of phylogenetic information, but without the aid of language embeddings or geographic information, respectively. However, the overall patterns otherwise do not shed much light on the relationship between training sample size and error rate across the different models; rather, they show that some low-frequency languages in the training data have orthographies that the remaining models have little difficulty capturing; this likely includes low-resource Austronesian languages with relatively transparent Latin writing systems.

Aggregating the PER for each model by writing system, we see in Table 2 that the model LangGeoGen outperforms the Lang model, if at times only marginally, on all non-Latin orthographies except for Devanagari, Lao, and Myanmar scripts. This result is striking, and lends support to the idea that the LangGeoGen performs so poorly on Latin data because it learns too much from this high-resource data set, extending information from observed data points to data points where it does not belong. In contrast, for lower-resource languages that show similar phonotactics due to genetic relatedness and grapheme-phoneme correspondence due to areal proximity (in South and Southeast Asia, symbols in different writing systems often have similar names), the model seems to have done a better job of leveraging information across training data sets.

Ultimately, the impact brought about by including phylogenetic and areal information in our model architecture remains difficult to assess. It is certainly the case that geographic and phylogenetic information alone cannot serve jointly as a unique language-specific identifier independent of language ID. For Latin orthographies, the inclusion of this information had detrimental effects, possibly because higher-resource languages dominate the landscape of the training data, providing evidence for patterns that are then applied to test data points where they are not appropriate. However, this was generally not the case for non-Latin orthographies, where the inclu-

	PER			WER		
Lang	**0.032**	**0.032**	**0.045**	**0.746**	**0.739**	**0.888**
LangGeoGen	0.048	0.047	0.056	0.768	0.759	0.960
GeoGen	0.033	**0.032**	0.053	0.907	0.905	0.941
LangGeo	0.033	**0.032**	0.046	0.771	0.766	0.889
LangGen	0.048	0.047	0.059	0.913	0.910	0.965

Table 1: PER and WER for each model, overall values and values for observed versus unseen languages, respectively

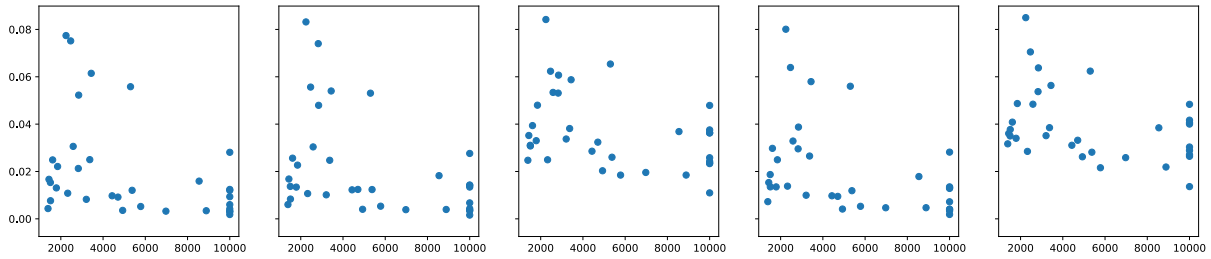

Figure 1: PER plotted by size of training data set for each language, for Lang, LangGeoGen, GeoGen, LangGeo, and LangGen models

	Lang	LangGeoGen	GeoGen	LangGeo	LangGen
LATIN	**0.030**	0.032	0.047	0.032	0.046
CYRILLIC	0.033	**0.030**	0.051	0.032	0.052
ARABIC	**0.059**	**0.059**	0.071	0.061	0.069
BENGALI	0.031	**0.029**	0.055	0.030	0.060
GUJARATI	0.060	**0.059**	0.067	0.050	0.070
TIBETAN	0.080	**0.076**	0.074	0.076	0.073
DEVANAGARI	**0.018**	0.022	0.039	0.022	0.041
KANNADA	0.044	**0.042**	0.067	0.041	0.048
KHMER	0.082	**0.076**	0.079	0.082	0.082
LAO	**0.021**	0.024	0.028	0.026	0.030
MALAYALAM	0.063	**0.049**	0.066	0.058	0.062
MYANMAR	**0.027**	0.028	0.039	0.030	0.048
GURMUKHI	0.056	**0.053**	0.072	0.060	0.074
TAMIL	0.072	0.071	0.063	**0.051**	0.067
TELUGU	0.075	0.070	0.072	**0.069**	0.072
THAI	0.075	**0.056**	0.062	0.064	0.070

Table 2: Average PER for each model aggregated by orthography

sion of this information led to improved performance.

Impressionistically, we also note that narrowness/broadness of IPA transcriptions is not consistent across the entire data set, undoubtedly an artifact of idiosyncratic variation between transcribers. There seems to be a narrowness bias in favor of Western Europe, even in the cleaned data set; data points from this region tend to have more detailed transcriptions which note for instance whether coronals such as /d/ are dental or apical. This undoubtedly has the effect of heightening and exaggerating differences among closely related speech varieties from Romance or Germanic. This issue could be addressed by splitting the training data into different sets (possibly excluding languages the cover a large region, which we did not do), or allowing a more flexible model parameterization which allows the influence of phylogeny and geography to vary across orthographic system or macroarea in a way that is sensitive to regions where transcriber attention to phonetic detail appears to be higher.

7 Future directions

This paper sought to address the effect of including phylogenetic and geographic information in g2p, reasoning that inconclusive and negative results from previous work on related tasks were perhaps due to the use of an insufficiently granular representation of these features. Our results are also somewhat inconclusive, but suggest that there may be some value to using this information for tasks involving lower-resource languages that are genetically and/or geographically close together, such as the languages of South and Southeast Asia. The negative result achieved for Latin orthographies suggests that some caution should be employed in using phylogenetic and areal information in NLP tasks for data sets of varying coverage, but if a flexible means of including this information without allowing it to drown out other valuable information can be employed, then this may serve as a promising research direction. It is also worth noting that we did not use a maximally granular phylogenetic representation, as the Glottolog information employed by URIEL does not employ any notion of chronological divergence between languages, as can be estimated using computational phylogenetic methods. As these methods are applied to more language families of the world, we can gain and incorporate a more nuanced understanding of phylogenetic distances between languages within the same family.

An outstanding issue is that we made use of a nonmonotonic hard attention mechanism, despite the fact that monotonicity appears to serve as a useful inductive bias with the potential to improve g2p results for certain languages, but not for others (e.g., languages with Devanagari and related writing systems). Wu and Cotterell (2019) achieve monotonicity via a neural model of transition probabilities from previous alignments to current alignments; in this matrix of neural transition probabilities between alignments, cells corresponding to any transition that would break strict monotonicity are multiplied by structural zeros, effectively preventing backwards jumps in alignment. Rather than multiply these cells by zero, it is in theory possible to learn a parameter for each writing system or even language in the data set which, when sigmoid-transformed, would multiply these cells by a value between 0 and 1, effectively allowing strictness of monotonicity to vary across the data set. Ultimately, we hope that insights from this study can lend themselves to the development of flexible approaches to g2p and related tasks which not only capture cross-linguistic variation but allow the importance of different potentially informative features to be weighted differently across languages, thus leading to improved performance.

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10.

Susan Baddeley and Anja Voeste. 2012. *Orthographies in Early Modern Europe*. Walter de Gruyter, Berlin.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv*, pages arXiv–1409.

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.

Matthew S. Dryer and Martin Haspelmath. 2013. WALS online. http://wals.info/.

Harry Falk. 1993. *Schrift im alten Indien: ein Forschungsbericht mit Anmerkungen*. Gunter Narr Verlag, Tübingen.

Alexander Gutkin and Richard Sproat. 2017. Areal and phylogenetic features for multilingual speech synthesis. *INTERSPEECH 2017*, pages 2078–2082.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.3. Max Planck Institute for the Science of Human History.

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

Yoon Kim, Sam Wiseman, and Alexander M Rush. 2018. A tutorial on deep latent variable models of natural language. *arXiv preprint arXiv:1812.06834*.

Young-Bum Kim and Benjamin Snyder. 2012. Universal grapheme-to-phoneme prediction over latin alphabets. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 332–343. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, and Pushpak Bhattacharyya. 2018. Leveraging orthographic similarity for multilingual neural transliteration. *Transactions of the Association for Computational Linguistics*, 6:303–316.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Jayden Macklin-Cordes and Erich Round. 2015. High-definition phonotactics reflect linguistic pasts. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*. University of Tübingen, online publication system.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649.

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

Taraka Rama and Prasanth Kolachina. 2012. How good are typological distances for determining genealogical relationships among languages? In *Proceedings of COLING 2012: Posters*, pages 975–984.

Richard Salomon. 1998. *Indian epigraphy: a guide to the study of inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan languages*. Oxford University Press, Oxford.

Anders Søgaard and Julie Wulff. 2012. An empirical etudy of non-lexical extensions to delexicalized transfer. In *Proceedings of COLING 2012: Posters*, pages 1181–1190.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. *arXiv preprint arXiv:1905.06319*.

Shijie Wu, Pamela Shapiro, and Ryan Cotterell. 2018. Hard non-monotonic attention for character-level transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4425–4438.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. Association for Computational Linguistics.