Interpreting Character Embeddings With Perceptual Representations: The Case of Shape, Sound, and Color

Sidsel Boldsen and **Manex Agirrezabal** and **Nora Hollenstein**
Centre for Language Technology
Department of Nordic Studies and Linguistics
University of Copenhagen
{sbol,manex.aguirrezabal,nora.hollenstein}@hum.ku.dk

Abstract

Character-level information is included in many NLP models, but evaluating the information encoded in character representations is an open issue. We leverage perceptual representations in the form of shape, sound, and color embeddings and perform a representational similarity analysis to evaluate their correlation with textual representations in five languages. This cross-lingual analysis shows that textual character representations correlate strongly with sound representations for languages using an alphabetic script, while shape correlates with featural scripts. We further develop a set of probing classifiers to intrinsically evaluate what phonological information is encoded in character embeddings. Our results suggest that information on features such as *voicing* are embedded in both LSTM and transformer-based representations.

1 Introduction

On the one hand, writing is an essential form of human communication. Writing systems and orthographies differ across languages and impact our reading behavior. Psycholinguists have extensively studied the effect of orthographic depth, i.e., the transparency of grapheme-to-phoneme mappings, on reading acquisition as well as skilled reading (Seymour et al., 2003).

On the other hand, the wide range of cross-linguistic diversity is still a major challenge for natural language processing (NLP) and for the study of language more generally (Mielke et al., 2019; Gutierrez-Vasques and Mijangos, 2020), especially on sub-word levels (Gutierrez-Vasques et al., 2021). This increases the importance of cross-lingual analyses of character-level language models (LMs), because anglocentrism in linguistic research is not only prevalent in NLP, but also in (reading and) orthography research (Share, 2008).

Character-based language models have gained significant attention in recent years in languages with Latin scripts, since they contain meaningful information on various linguistic levels and enhance the robustness of models. Oh et al. (2021) suggest that character LMs provide a more human-like account of sentence processing, which assumes a larger role of morphology, phonotactics, and orthographic complexity than was previously thought. Moreover, including character and sub-character information in LMs for Asian scripts is a standard practice. Despite of this recent attention, work focusing on getting a deeper understanding of character representation is scarce (Kann and Monsalve-Mercado, 2021), in particular regarding the comparison between languages and different types of scripts.

The goal of this work is to improve our understanding of learned character representations, for better interpretability of the models. Like other neural network based models, character-level LMs can be seen as black-box methods and reveal limited insights about the causes for their predictions (Gilpin et al., 2018). We investigate the information encoded in character embeddings by comparing them to perceptual representations. Such representations we design by mimicking features of human language processing, from reading, writing and speaking, by the creation of embeddings based on the *shape* of characters, the *sound* (phonological features derived from grapheme-to-phoneme mappings) and *color* (elicited in the form of grapheme-color mappings from synesthetes).

Contributions We train models to learn three types of character embeddings: a positive pointwise mutual information (PPMI) vectorization, a recurrent model, and a transformer model. As an intrinsic evaluation method, we conduct a representational similarity analysis (RSA) between the distances of textual character representations and the perceptual representations in the form of shape, sound, and color embeddings. Furthermore, to pro-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6819 - 6836

vide more interpretable evaluation methods for character embeddings, we propose a novel probing task of predicting phonological features. Crucially, we address the cross-linguistic challenges that arise with character-level modeling by taking into account languages of varying scripts and orthographic depths. We argue that character-level black-box models can only be understood through cross-linguistic approaches and not on individual languages. We perform analyses of five languages: Dutch, English, Japanese, Korean, and Spanish. We discuss the compelling patterns of significant correlations and show the effectiveness of the probing classifiers even in a zero-shot scenario. The implementation and character representations are available online[1].

2 Related Work

Character-level information in LMs. Including character-level information in LMs of languages with Latin scripts has become a common practice in NLP in recent years. This has been the case for different tasks, such as language modeling (Kim et al., 2016; Al-Rfou et al., 2019), part-of-speech tagging (Ling et al., 2015), morphological inflection (Faruqui et al., 2016; Kann and Schütze, 2016; Kann et al., 2020), named entity recognition (Lample et al., 2016), machine translation (Sennrich et al., 2016; Ngo et al., 2019). Character-level information can enhance the models by providing background knowledge in the form of the underlying structures of words in a language (Adouane et al., 2018). Ma et al. (2020) showed how combining character- and word-level information in pretrained LMs improves not only the performance but also the robustness of the model.

For certain languages, it is standard practice to include sub-token information in LMs, which happens naturally due to the compositional structure of their orthographies. This is the case for East Asian languages such as Korean and Japanese (e.g., Misawa et al. 2017; Chen et al. 2015). Korean LMs are often trained on *Jamos* (i.e., letters, as opposed to syllables), the smallest unit of the Korean script (Ahn et al., 2017; Park et al., 2018). This reduces the vocabulary size and injects syntactic and semantic information to the model that is difficult to access with conventional character- or token-level units (Stratos, 2017). Recently, Lee et al. (2020b) showed that a Korean BERT model using sub-character information requires less training data than previous models. Similarly, Japanese LMs also benefit from sub-character information (Nguyen et al., 2017).

Evaluating character embeddings. Character-based language models are most often evaluated on downstream NLP tasks or on next character or word prediction (e.g., Takase et al. 2019; Tay et al. 2021; Clark et al. 2021). Additionally, they can be evaluated on word-level intrinsic evaluation tasks such as word analogy or similarity (e.g., Li et al. 2015). While work on intrisic evaluation of character embeddings is scarce (Kann and Monsalve-Mercado, 2021), the evaluation of neural models trained on phonemes have received more attention, focusing on what phonological knowledge is embedded within (Silfverberg et al., 2018; Kolachina and Magyar, 2019; Mayer and Nelson, 2020; Mayer, 2020; Silfverberg et al., 2021). Mayer (2020) and Mayer and Nelson (2020) use characters as an approximation of phonemes in the case of Samoa and Finnish, respectively, as graphemes are closely connected to phonemes in these orthographies.

The methods we leverage in this paper, previously applied for evaluating different types of representations, are *representational similarity analysis* (RSA) and probing classifiers. The former was first proposed by Kriegeskorte et al. (2008) for comparing brain activity vectors in heterogeneous representational spaces, but has also been applied in NLP as an interpretability metric as it allows us to study the relation between language representations (Abnar et al., 2019; Abdou et al., 2019; Chrupała and Alishahi, 2019). RSA enables a transparent comparison between the representational geometries of different models and modalities (Søgaard, 2021).

Contrarily, *probing classifiers* learn to classify output representations in supervised settings (Ettinger et al., 2016). The intuition behind probing is that if a classifier can be learned to accurately predict certain linguistic properties from the representations of a neural model, then this model has "learned" this property. Typically, lightly parametrized classifiers (like logistic regression) are applied, however, the exact trade-off between accuracy and complexity of a probe is an open question (Belinkov, 2021). In recent years, NLP studies have used probing classifiers to investigate

[1] https://github.com/syssel/Interpreting-character-embeddings

whether LMs encode linguistic properties including morphological features (such as *person* and *number*, Torroba Hennigen et al. (2020)) and word sense (Coenen et al., 2019). However, we apply probing classifiers for the first time to character representations.

Impact of different orthographies on linguistics and human language learning. Orthographic depth, i.e., the transparency of grapheme-phoneme correspondences in written language (Frost et al., 1987; Katz and Frost, 1992), is a well-studied factor influencing reading acquisition and skilled reading behavior (Seymour et al., 2003; Landerl et al., 2013; Richlan, 2020). For instance, English is considered to be a *deep orthography*, as there are often multiple different pronunciations for the same spelling patterns (e.g., <gh> in *tough* and *though*). This contrasts *shallow orthographies* with more reliable grapheme-phoneme correspondences, such as Spanish. The consistency and complexity with which print reflects speech is one of the prime factors of cross-linguistic differences in reading fluency (Ziegler et al., 2010; Schmalz et al., 2015). It is the starting point for any discussion that centers on reading development across languages (Papadopoulos et al., 2021). Since the orthography has such a high impact on human reading behavior, its effect should also be considered more carefully in the development of NLP models.

Impact of different orthographies on NLP models. While orthographic depth has been discussed at length in reading research and psychology, it has rarely been addressed in NLP. This partly due to the prevalent anglocentrism and missing resources (Bender, 2018). Some research has gone into studying the differences between languages when it comes to train computational LMs (Mielke et al., 2019), showing the impact of the vocabulary size and sentence length, but there is lack of NLP research analyzing or taking into account the varying orthographies across languages. Two notable exceptions are the recent methods proposed by Marjou (2021) and Sproat and Gutkin (2021), who use neural networks to estimate the transparency of orthographies and degree of logography, respectively. Moreover, Gorman et al. (2020) conducted a shared task on grapheme-to-phoneme prediction. Their results show an urgency for improving these systems and the pronunciation dictionaries used to train them across languages and scripts.

3 Character Representations

We train three types of character embeddings based on textual input: count-based PPMI embeddings, and embeddings learned by LSTM and transformer language model.

3.1 Character Language Models

We use the Wiki40B multilingual dataset (Guo et al., 2020) to train the character models. For each of the five languages, English (en), Dutch (nl), Spanish (es), Korean (ko), and Japanese (ja), we extract training sets of 3 million characters. See Appendix A for details on preprocessing. The first three languages all use variants of the *Latin* script, while *Hangul* (Korean) and *Hiragana* (one of three scripts used in Japanese) are *syllabic* scripts, in which most graphemes denote entire syllables. We preprocess Korean *Hangul* characters, decomposing them into constituent *Jamos*, each corresponding roughly to a single phoneme. For Japanese, we convert *Kanji* symbols to *Hiragana* and train the language model on *Hiragana* and *Katakana* characters. The representational similarity analyses are then only performed on *Hiragana*. Figure 1 shows 2-dimensional plots of the learned textual character representations.

Count-based PPMI embeddings. We generate vectorized character representations in a purely count-based manner with a positive pointwise mutual information (PPMI) weighting. While the importance of positional information is less obvious for modelling word semantics, it is crucial for modelling the distribution of sounds. Following the approach by Mayer (2020), we let our PPMI weighting diverge from traditional bag-of-words models by distinguishing contexts by their relative position to a target. Thus, embeddings will have independent values for the contexts `AB_`, `_AB`, and `A_B`, counting the number of times a target follows, precedes, and mediates a string `AB`. Using bigram contexts, the resulting embeddings have a dimension of $3 \cdot c^2$, where c is the number of characters in a given language, and 3 indicating the number of possible relative positions.

LSTM. We train a recurrent language model consisting of two unidirectional long-short term memory (LSTM) layers. It receives sequences of 40 characters as input at each time step and is trained for next character prediction. The model is trained with an Adam optimizer (Kingma and Ba, 2015),

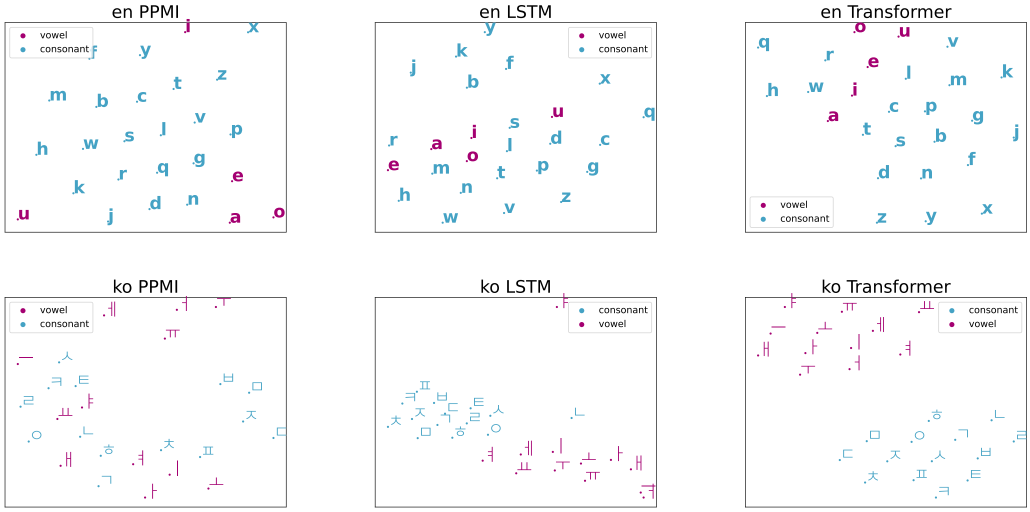

Figure 1: tSNE cluster plots of the character distances from the three types of character language models for English and Korean (see Appendix Figure 5 for the plots for Dutch, Spanish and Japanese).

Figure 2: Example of letter-color associations from single subjects.

an initial learning rate of 0.01, and a batch size of 128. We extract the hidden representations of 128 dimensions as the character embeddings. See Appendix C.1 for training specifications and Appendix C.2 for perplexity metrics. We additionally experimented with bidirectional LSTMs (see Table 1) and 1-layer LSTMs without any substantial changes in the results (Appendix C).

Transformer. Similarly, we also train a transformer character model on the same data (Vaswani et al., 2017). The input layer consists of character and positional embeddings, followed by a single transformer block with 2 heads and a hidden layer size of 128. We follow the same training procedure as for the LSTM and extract the representations of the hidden layer as the character embeddings. Again, see Appendix C for additional details, model modifications, and perplexity metrics.

3.2 Perceptual Representations

Sound. The first perceptual representation that we consider is *sound*. To retrieve this representation, we map characters to a phonological distinctive feature space. This method has previously been applied to phonemes as a means of generalisation compared to sparse representations (Rumelhart and McClelland, 1986; Mirea and Bicknell, 2019), and to evaluate the knowledge embedded representations learned from neural networks (Silfverberg et al., 2018; Kolachina and Magyar, 2019).

As sound and speech are only indirectly reflected in writing, we approximate sound representations of characters using grapheme-to-phoneme alignment: For all languages, we extract data from the WikiPron pronunciation dictionary (Lee et al., 2020a) and use the `m2m-aligner` (Jiampojamarn et al., 2007) to align graphemes with phonemes in an unsupervised manner. Having alignments from the WikiPron data, we chose the most frequent phoneme mapping to represent the sound of each character (resulting mappings are listed in the Appendix D) We also considered extracting the most frequent phoneme mapping only from word-initial positions. The intuition behind this approach was to retrieve representations as close to phonemic as possible, as sounds in the initial position are expected to be less prone to phenomena such as reduction and assimilation reflected in the WikiPron data (e.g., reduction of English "o" to ə). However, the word-initial position is also subject to phonotactic restrictions: For example, in Korean only including consonants occurring word-initially heavily reduces the inventory considered.

Having phonemes mapped to characters, we are able to associate it with a set of phonological distinctive features, which we use to form our final sound representation: Using the `ipapy`[2] toolkit, we retrieve International Phonetic Alphabet (IPA) descriptions of the phoneme mappings from which we create a sparse vector that describes what phonological features (e.g., consonant manner of articulation, ±plosive, or vowel height, ±front) are active. For every language, this provides us with a sound embedding table, $S^{|V|\times|F|}$, where V is the set of characters and F is the set of distinctive features:

$$S_{i,j} = \begin{cases} 1 & \text{if } F_j \in \text{phonmap}(V_i). \\ 0 & \text{otherwise.} \end{cases}$$

Color. Inspired by Kann and Monsalve-Mercado (2021), we compute color character representations from synesthesia data. Grapheme-color synesthesia is a neurological phenomenon in which viewing a grapheme elicits an automatic, involuntary, and consistent sensation of color (Eagleman et al., 2007). Color-to-letter associations in synesthesia allow to examine the relationships between visual, acoustic, and semantic aspects of language. Recent research in this area has found cross-linguistic similarities in synesthesia, suggesting that some influences on grapheme-color associations in synesthesia might be universal and highlighting the importance of multilingual analyses (Root et al., 2018). Figure 2 shows example grapheme-color associations from individual subjects for each of our studied languages. It emphasizes the preference for red color tones for the first letter of the alphabet irrespective of the language (Root et al., 2018).

We use the cross-linguistic synesthesia data collected by Root et al. 2018 (see Appendix B for the dataset statistics). In order to extract color representations we compute the Euclidean distances between the 3-dimensional CIELuv color coding scheme for all character combinations. We average the distances across all participants of the same language. The resulting vector representations reflect the finding of Root et al. (2018) that the first grapheme in any language is unusually distinct (see Figure 4 in Appendix) .

[2] https://github.com/pettarin/ipapy

Shape. Lastly, we also create simple character representations based on their shape. Previous works (Brang et al., 2011; Watson et al., 2012) have relied on Gibson (1969) or Courrieu et al. (2004) to build shape-related embeddings from human similarity judgements. However, we create shape embeddings directly from their visual expressions. We create an image for each printed character as shown in Figure 6 in the appendix. For each script, all images have the same width and height (the largest width among all characters incremented with 10 pixels, and the same for the height, which results approximately in 35×45 pixels) and all characters are drawn at position `{5,5}`. We use the font *Arial Unicode MS* with size 28. From these images, we create shape representations by reading the images as gray scale images row-wise from top to bottom and flattening the matrix into vectors.

4 Representational Similarity Analysis

In order to analyze the relation between the learned character representations and the three perceptual representations – sound, shape, and color – we first compute the pairwise distances between characters of a single model/representation type to analyze how similar the model's representations for each character are to each other[3]. For each pair of experimental conditions, the spatial correlation is calculated between the distances of all characters of a language. Figure 3 shows the Pearson correlations between the character distances of all embedding types. The figure also includes a baseline, where the correlation between random distances and the distances of the respective character representations is computed. We correct the significance results by applying the Bonferroni correction for multiple comparisons.

As expected, the textual character representations show high correlation amongst each other for all five languages. The correlations between the textual embeddings and the perceptual representations show that even though the first are purely trained on written language, they still learn to encode certain inherent characteristics of human language processing and production.

As a general pattern, the textual character representations correlate strongly with sound representations, moderately with color representations, and not at all with the shape representations (with the

[3] We use cosine distance for all textual, sound and shape representations; and Euclidean distance for color.

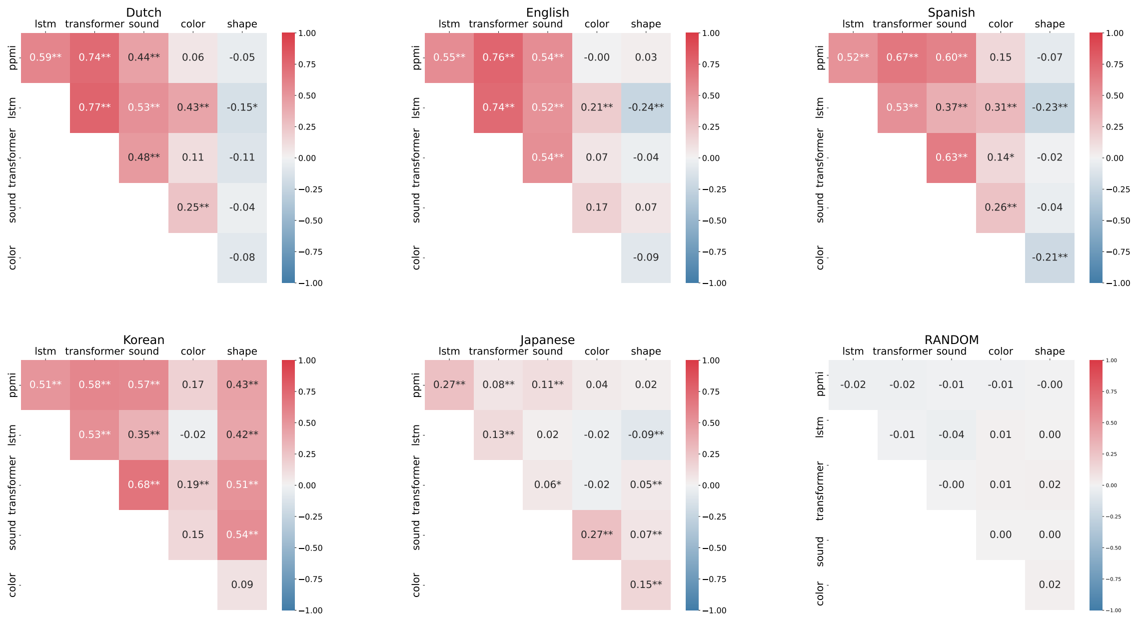

Figure 3: Pearson correlation between all representations types for all five languages and for the random baseline (bottom right). A * marks a significant correlation ($p < 0.01$), ** marks a significant correlation under the Bonferroni correction ($p < 0.003$).

exception of Korean, discussed below). Japanese character embeddings behave differently. For instance, the correlation with the sound representations is weaker than for the other languages, which might be due to the syllabic nature of the Japanese script. In the following, we discuss the results for each of the perceptual embedding types in detail.

4.1 Sound

The PPMI character embeddings show the highest correlation with sound representations, followed closely by transformer embeddings. This is notable in the three languages with Latin scripts (en, es, nl). To explain this finding, we speculate that the context and learning direction available to the LMs provide phonetic information. While the PPMI embeddings have access to contextual information in both directions, the unidirectional LSTM and transformer learn from left-to-right only. Therefore, as an addition, we trained a bidirectional LSTM (hidden dimension = 256) to show that the addition of right-to-left information improves the correlation to the sound representations. The results are shown in Table 1. Moreover, comparing the results across Latin script, we note that Spanish character embeddings from all models achieve higher correlations than Dutch and English. The shallow orthography of the Spanish language explains this finding. This is also the case for Korean.

	en	es	nl
PPMI	**0.54**	0.60	0.44
LSTM	0.52	0.37	**0.53**
biLSTM	0.48	0.34	0.42
Transformer	**0.54**	**0.63**	0.48

Table 1: Correlations between sound representations and character embeddings.

	syllables	Jamos
Sound	0.04	0.68
Color	–	0.19
Shape	0.03	0.51

Table 2: Pearson correlation coefficients for Korean transformer character embeddings based on Hangul syllables vs. Jamos. As the synesthesia data only includes Jamos, we exclude the syllable correlation for color.

4.2 Color

Our findings on the correlation between English character embeddings and synesthesia data are in line with Kann and Monsalve-Mercado (2021), who find that LSTMs agree with human letter-color perceptions more than transformers on a dataset with more participants (0.08 for LSTM-LM and 0.0 for transformer-LM). Moreover, we reach the same conclusion for the other alphabetic scripts, Dutch and Spanish, while for Korean and Japanese there is no clear pattern evident from the correlation coefficients. This might be due to the smaller number of synesthete participants in the dataset.

4.3 Shape

The character embeddings of non-featural Latin scripts show low (or even negative) correlation to the shape embedding. However, due to their featural writing systems (Sampson, 1985; Marjou, 2021), Japanese and especially Korean embeddings correlate significantly with shape. The fact that the Korean consonant graphemes were designed to resemble the place of articulation (Lee, 2021; Gale, 1912), can explain the high correlations between character and shape embeddings for this language.

This is also shown in the positive correlation between sound and shape representations, which is absent for the other languages. To analyze this further, we compare our initial results with transformer character representations computed based on *Jamos* (e.g., individual phonemes such as "ㄱ"), to character representations of full *Hangul* characters (e.g., syllables such as "공"). Table 2 shows higher correlations for characters decomposed into Jamos. The correlation between sound and shape is also lower for full syllables (0.31).

In this light, the result is unsurprising and can be interpreted as an effective proof-of-concept of using a correlation analysis between textual and perceptual representations. More genuine shape representations, for example learned by a convolutional neural network, could be applied to reveal more accurate correlation patterns for Latin scripts.

5 Probing Task

Except for Japanese, the results show that the neural embeddings correlate the most with the perceptual sound representations. To get a closer look at the information that may be encoded in the dense embeddings, we design a probing task in which classifiers are trained to predict whether certain distinctive features are present given character embeddings as input.

5.1 Classifier Setup

For each distinctive feature, we train a binary Logistic Regression to predict whether the the feature is present (1), or not (0). The labels are given by the sound representations as explained in Section 3.2. As the number of samples is small (limited to the number of characters in a language), we do this in a leave-one-out manner, training a classifier for each character, while using the rest for training. In both test and training, for features that only concern consonants (e.g., *manner of articulation* and *voicing*), we exclude vowels, and similarly, for features that only concern vowels (e.g., *vowel height* and *vowel rounding*), we exclude consonants.

The performance of the probes are evaluated for each distinctive feature using F1 scores and by comparison with two baseline strategies, namely, (a) to predict labels uniformly at random, and (b) to always predict the most frequent label according to the training distribution. The former is given as the average across 1000 runs.

5.2 Zero-Shot Classifiers

For some features, choosing the most frequent label is a good strategy and will yield good results. To further challenge the knowledge learned by the embeddings and distinguish the classifiers from the strategy of choosing the most frequent baseline, we create a zero-shot setup in which the classifiers will have to be able to transfer knowledge between features in order to excel in the task. In particular, we test 1) if a classifier trained to predict whether a consonant is voiced is able to identify vowels and 2) if labial consonants are retrieved by a classifier trained to predict vowel rounding. While the intuition behind 1) relates to the sonority sequencing principle (Clements, 1990), which states that the nucleus of a syllable (vowels in the majority of the cases) represents a sonority peak, the intuition behind 2) is more experimental, relying on a global feature such as 'rounding'.

5.3 Results and Discussion

The results for the probing classifiers are found in Table 3. Generally, both LSTM and transformer embeddings outperform both the most-frequent and random baselines, with the transformer beating the LSTM by a small margin. This should, however, be taken with a grain of salt considering the limited number of examples.

Considering the global features, *vowel* and *consonant*, classifiers are able to learn this distinction using both LSTM and transformer character embeddings. In particular, consonants are identified with high certainty. This is, however, the majority group (ref. the most frequent strategy). The F1 scores for vowel prediction are considerably lower. However, in this case they cannot be explained by neither a most-frequent strategy nor a random baseline, which indicates that a global vowel/consonant distinction is captured in the embeddings.

	Model	global type *consonant*	*vowel*	consonant voicing *voiced*	*voiceless*	vowel rounding *rounded*	*unrounded*
en	LSTM	0.95	0.80	0.55	0.44	-	-
	Transformer	**0.97**	**0.92**	**0.75**	**0.63**	-	-
	Random	0.60	0.32	0.52	0.48	-	-
	Most-frequent	0.87	0.00	0.71	0.00	-	-
es	LSTM	0.96	0.75	0.40	0.50	0.00	**0.75**
	Transformer	**0.98**	**0.89**	**0.72**	**0.63**	0.00	0.33
	Random	0.61	0.27	0.49	0.49	**0.43**	0.49
	Most-frequent	0.90	0.00	0.00	0.00	0.00	0.00
ko	LSTM	**1.00**	**1.00**	**0.50**	0.65	0.00	0.83
	Transformer	**1.00**	**1.00**	0.43	0.47	0.00	0.86
	Random	0.53	0.46	0.46	0.53	**0.27**	0.62
	Most-frequent	0.73	0.00	0.00	**0.71**	0.00	**0.89**
nl	LSTM	**0.97**	**0.92**	**0.80**	**0.62**	0.00	**0.83**
	Transformer	**0.97**	**0.92**	0.75	0.57	0.00	**0.83**
	Random	0.59	0.35	0.53	0.45	**0.35**	0.57
	Most-frequent	0.84	0.00	0.73	0.00	0.00	**0.83**

Table 3: F1 score for classifiers predicting distinctive features with character embeddings (LSTM, Transformer) as input. Two baselines are included: Random (predicting labels uniformly at random) and Most-frequent (always predicting the most frequent label). Since English only has one rounded vowel (the character 'o' mapped to IPA 'ɒ'), the result for this classifier is not included. Results for predicting all distinctive features are found in Appendix E.

	Model	consonant voicing:*voiced* $\rightarrow$ global type:*vowel*
en	LSTM	0.80
	Transformer	0.50
	Random	0.64
	Most-frequent	1.00
es	LSTM	0.89
	Transformer	0.89
	Random	0.64
	Most-frequent	0.00
ko	LSTM	0.09
	Transformer	0.83
	Random	0.66
	Most-frequent	0.00
nl	LSTM	1.00
	Transformer	0.83
	Random	0.64
	Most-frequent	1.00

Table 4: F1 score for predicting vowels using a classifier trained to predict whether a consonant is voiced. Two baselines are included: Random (predicting labels uniformly at random) and Most-frequent (predicting the most-frequent label, w.r.t. the label distribution in the original task).

The findings for the voiced/voiceless consonant distinction are similar. But here the groups are more balanced, which provides the most-frequent strategy with less of an advantage and in turn the F1 scores are generally lower. For Korean, the scores are lower compared to the other languages. As the feature of consonant voicing correlates with manner in Korean (with all plosives, affricates and fricatives being voiceless, and plosives being the majority class), the task captured by the classifier may be distorted. The fact that the classifier may not be able to pick up features of voicing from the Korean embeddings are reflected in the zero-shot experiment.

The results for the first zero-shot experiment for predicting vowels using the classifier for identifying voiced consonants are found in Table 4. Here, the results for Korean are worse than the random baseline. While the results for English and Dutch can be explained by the most-frequent strategy, the result for Spanish indicates that features of voicing or sonority are encoded in the embeddings, amplifying the initial results from the probing classifier experiment.

Turning from consonant to vowel features, the inventory of vowels is considerably smaller, leaving a small number of training examples with few positive examples. Thus, the results of the probing classifiers are associated with uncertainty. For the zero-shot task of retrieving consonants with labial features from a classifier trained to predict vowel rounding, we focus our analysis on Spanish LSTM embeddings as they showed the most promising results for predicting *rounding* in the regular probing

Language	F1	True positive	False positive	False negative
es	0.47	b f v w	g k q x y z ñ	p m

Table 5: Results from the zero-shot task to predict 'rounded' consonants using the Spanish LSTM embeddings. Using a classifier to predict the vowel rounding of consonants, the following consonants are retrieved. F1 score indicates the ability to identify consonants with a labial place of articulation.

task. However, as can be seen in Table 5, while the classifier for Spanish has a high recall its precision lacks behind and retrieves many false positives.

Overall, we believe that the results are promising and a good indication on how character representations can capture features related to phonology. This especially in light of the results from the first zero-shot task, that suggested that classifiers are able to transfer knowledge of sonority from embeddings of consonants to unseen vowels.

6 Conclusion

In this work, we attempted to understand the information encoded in character-level representations. We obtained two main types of embeddings: text-based embeddings and perceptual embeddings. While the first type of representations (PPMI, LSTM, and transformer) were trained from raw text data, perceptual representations were obtained from sources mimicking human language, i.e., pronunciation dictionaries, synesthesia data and shape visualizations. We have performed representational similarity analyses between these types of embeddings for five different languages. Besides, we defined and trained models to predict certain phonological distinctive features in order to interpret the embeddings.

We found interesting patterns in the representational similarity analysis as a simple first approach for intrinsic character embedding evaluation. While clearly outperforming a random baseline in most cases, the strength of the correlations vary between scripts. For instance, the strong correlation between Korean character embeddings and shape representations provides positive evidence of the suitability of this approach. Further research is required to dissect the differences between character LMs: While the LSTM embeddings showed stronger correlation with color, the transformer embeddings were superior when compared to sound representations. The inclusion of additional languages and scripts will be helpful to identify more generalizable insights.

These perceptual representations could be used as pre-trained representations. It might be the case that they contribute differently for different tasks. For instance, sound representations would be expected to be useful for tasks revolving around phonology, such as grapheme-to-phoneme conversion, or shape representations could be relevant for predicting orthographic errors.

The phonological probing tasks show promising results, especially with respect to interpretability. Besides, this methodology is applicable to any language with sufficient raw data and a pronunciation dictionary, and could potentially shed light in measuring the phonological difficulty of certain languages. In future work, we will focus on the development of more sophisticated probes, for instance, multitask networks with shared layers across tasks. Moreover, the labels of the probing task were given from using the sound embeddings retrieved from the most frequent phoneme mapping. Had we focused the analysis on contextual character embeddings instead, that would allow us to distance ourselves from this paradigm as we would be able to analyse character and sound embeddings in the context they occur in.

Finally, we stress the need for further intrinsic evaluation methods for character representations. The high impact of orthography on human language learning is an adamant argument to consider the cross-linguistic diversity of writing systems more carefully in the development of NLP models.

Acknowledgements

We thank the anonymous reviewers from ACL Rolling Review for their valuable comments. The first author is supported by the project *Script and Text in Time and Space* funded by the Velux Foundations.

References

Mostafa Abdou, Artur Kulmizev, Felix Hill, Daniel M Low, and Anders Søgaard. 2019. Higher-order comparisons of sentence encoder representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5838–5845.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Wafia Adouane, Simon Dobnik, Jean-Philippe Bernardy, and Nasredine Semmar. 2018. A comparison of character neural language model and bootstrapping for language identification in multilingual noisy texts. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 22–31.

SungMahn Ahn, Yeojin Chung, Jaejoon Lee, and Jiheon Yang. 2017. Korean sentence generation using phoneme-level LSTM language model. *Journal of Intelligence and Information Systems*, 23(2):71–88.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166.

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–13.

Emily M Bender. 2018. How to make ends meet: Why general purpose NLU needs linguistics. In *Talk presented at the Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP (RELNLP) at ACL*.

David Brang, Romke Rouw, Vilayanur S Ramachandran, and Seana Coulson. 2011. Similarly shaped letters evoke similar colors in grapheme–color synesthesia. *Neuropsychologia*, 49(5):1355–1358. Publisher: Elsevier.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint learning of character and word embeddings. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2021. CANINE: Pre-training an efficient tokenization-free encoder for language representation. *arXiv preprint arXiv:2103.06874*.

George N Clements. 1990. The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, 1:283–333.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. In *NeurIPS*.

Pierre Courrieu, Fernand Farioli, and Jonathan Grainger. 2004. Inverse discrimination time as a perceptual distance for alphabetic characters. *Visual Cognition*, 11(7):901–919. Publisher: Taylor & Francis.

David M Eagleman, Arielle D Kagan, Stephanie S Nelson, Deepak Sagaram, and Anand K Sarma. 2007. A standardized test battery for the study of synesthesia. *Journal of neuroscience methods*, 159(1):139–145.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological Inflection Generation Using Character Sequence to Sequence Learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643.

Ram Frost, Leonard Katz, and Shlomo Bentin. 1987. Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1):104.

James S Gale. 1912. The Korean Alphabet. *Transactions of the Korea Branch of the Royal Asiatic Society*, 4(1).

Eleanor Jack Gibson. 1969. Principles of perceptual learning and development. Publisher: Appleton-Century-Crofts.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

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.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

Ximena Gutierrez-Vasques and Victor Mijangos. 2020. Productivity and predictability for measuring morphological complexity. *Entropy*, 22(1):48.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York. Association for Computational Linguistics.

Katharina Kann, Samuel R Bowman, and Kyunghyun Cho. 2020. Learning to learn morphological inflection for resource-poor languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8058–8065. Issue: 05.

Katharina Kann and Mauro M. Monsalve-Mercado. 2021. Coloring the black box: What synesthesia tells us about character embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2673–2685, Online. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70.

Leonard Katz and Ram Frost. 1992. The reading process is different for different orthographies: The Orthographic Depth Hypothesis. In *Advances in Psychology*, volume 94, pages 67–84. Elsevier.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 2741–2749. AAAI press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about phonology? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 160–169, Florence, Italy. Association for Computational Linguistics.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Karin Landerl, Franck Ramus, Kristina Moll, Heikki Lyytinen, Paavo HT Leppänen, Kaisa Lohvansuu, Michael O'Donovan, Julie Williams, Jürgen Bartling, Jennifer Bruder, et al. 2013. Predictors of developmental dyslexia in european orthographies with varying complexity. *Journal of Child Psychology and Psychiatry*, 54(6):686–694.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020a. 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.

Sang-Oak Lee. 2021. Graphical ingenuity in the Korean writing system: with new reference to calligraphy. In *The Korean Alphabet*, pages 107–116. University of Hawaii Press.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020b. Kr-BERT: A small-scale Korean-specific language model. *arXiv preprint arXiv:2008.03979*.

Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. 2015. Component-enhanced Chinese character embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 829–834.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luis Marujo, and Tiago Luís. 2015. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xavier Marjou. 2021. OTEANN: Estimating the transparency of orthographies with an artificial neural network. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 1–9, Online. Association for Computational Linguistics.

Connor Mayer. 2020. An algorithm for learning phonological classes from distributional similarity. *Phonology*, 37(1):91–131.

Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics*, 3(1):149–159.

Sabrina J Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989.

Nicole Mirea and Klinton Bicknell. 2019. Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1595–1605, Florence, Italy. Association for Computational Linguistics.

Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2017. Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition. In *Proceedings of the first workshop on subword and character level models in NLP*, pages 97–102.

Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. 2019. How transformer revitalizes character-based neural machine translation: An investigation on Japanese-Vietnamese translation systems. *arXiv preprint arXiv:1910.02238*.

Viet Nguyen, Julian Brooke, and Timothy Baldwin. 2017. Sub-character neural language modelling in Japanese. In *Proceedings of the first workshop on subword and character level models in NLP*, pages 148–153.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Timothy C Papadopoulos, Valéria Csépe, Mikko Aro, Marketa Caravolas, Irene-Anna Diakidoy, and Thierry Olive. 2021. Methodological issues in literacy research across languages: Evidence from alphabetic orthographies. *Reading Research Quarterly*, 56:S351–S370.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. Subword-level word vector representations for Korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2429–2438.

Fabio Richlan. 2020. The functional neuroanatomy of developmental dyslexia across languages and writing systems. *Frontiers in Psychology*, 11:155.

Nicholas B Root, Romke Rouw, Michiko Asano, Chai-Youn Kim, Helena Melero, Kazuhiko Yokosawa, and Vilayanur S Ramachandran. 2018. Why is the synesthete's "A" red? Using a five-language dataset to disentangle the effects of shape, sound, semantics, and ordinality on inducer–concurrent relationships in grapheme-color synesthesia. *Cortex*, 99:375–389.

D. E. Rumelhart and J. L. McClelland. 1986. *On Learning the Past Tenses of English Verbs*, page 216–271. MIT Press, Cambridge, MA, USA.

Geoffrey. Sampson. 1985. *Writing systems: a linguistic introduction*. Hutchinson, London.

Geoffrey Sampson. 1990. The writing systems of the world. *Journal of Linguistics*, 26(1):275–276.

Xenia Schmalz, Eva Marinus, Max Coltheart, and Anne Castles. 2015. Getting to the bottom of orthographic depth. *Psychonomic bulletin & review*, 22(6):1614–1629.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Philip HK Seymour, Mikko Aro, Jane M Erskine, and Collaboration with COST Action A8 Network. 2003. Foundation literacy acquisition in european orthographies. *British Journal of psychology*, 94(2):143–174.

David L Share. 2008. On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(4):584.

Miikka Silfverberg, Francis Tyers, Garrett Nicolai, and Mans Hulden. 2021. Do RNN states encode abstract phonological alternations? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5501–5513, Online. Association for Computational Linguistics.

Miikka P. Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Anders Søgaard. 2021. Explainable natural language processing. *Synthesis Lectures on Human Language Technologies*, 14(3):1–123.

Richard Sproat and Alexander Gutkin. 2021. The Taxonomy of Writing Systems: How to Measure How Logographic a System Is. *Computational Linguistics*, 47(3):477–528.

Karl Stratos. 2017. A Sub-Character Architecture for Korean Language Processing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 721–726.

Sho Takase, Jun Suzuki, and Masaaki Nagata. 2019. Character n-gram embeddings to improve RNN language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5074–5082.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Pnformation processing Systems*, pages 5998–6008.

Marcus R Watson, Kathleen A Akins, and James T Enns. 2012. Second-order mappings in grapheme–color synesthesia. *Psychonomic Bulletin & Review*, 19(2):211–217. Publisher: Springer.

Johannes C Ziegler, Daisy Bertrand, Dénes Tóth, Valéria Csépe, Alexandra Reis, Luís Faísca, Nina Saine, Heikki Lyytinen, Anniek Vaessen, and Leo Blomert. 2010. Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, 21(4):551–559.

A Preprocessing

We download the Wiki40B dataset for each of the five languages (English, Dutch, Japanese, Korean, and Spanish) from TensorFlow Hub[4]. We lowercase all letters. For English and Dutch, we consider the 26 standard letters of the alphabet, digits and punctuation marks. For Spanish, we additionally add *ñ* and and remove diacritics from vowels. For Korean, we consider all Hangul characters, digits and punctuation marks. Since Hangul is a featural writing system (Sampson, 1990), we split the compound symbols into phoneme-like constituents called *Jamos*[5]. For Japanese, we convert Kanji characters to Hiragana[6] to reduce the large vocabulary size to a syllabic alphabet. The language model is then trained on Hiragana and Katakana characters. However, for subsequent analyses we focus only on Hiragana. For all languages, we replace any other special characters with the symbol €.

B Datasets

This section provides further information about the datasets used to extract the perceptual representations.

B.1 Synesthesia Dataset

As described in the main paper, we use the synesthesia data collected by Root et al. 2018. The data is available upon request by the first author. Table 6 shows the number of characters and participants included for each language in the dataset.

Language	# Chars	# Participants
English	26	47
Dutch	26	110
Japanese	46	27
Korean	24	13
Spanish	26	32

Table 6: Synesthesia dataset details showing the number of characters included for each language and the number of synesthetes participating in the study.

In Figure 4 the characters are plotted by the distances between their corresponding colors. Based on this dataset, Root et al. 2018 showed how some influences on grapheme-color associations in synesthesia might be universal across languages. Their results suggest that grapheme-color associations follow an ordinal explanation, meaning that the unusually distinct first grapheme of a synesthete's alphabet tends to be associated with the unusually distinct color red. In line with their findings, the clusters show the greatest distance between the associated colors of the first grapheme of the alphabets (i.e., "a" in English and Spanish and "ㄱ" in Korean).

B.2 Shape Dataset

Please find in Figure 6 some examples of character figures that were used to build shape representations. We besides include in figure 7 three dendrograms calculated from the shape representations. For Spanish, English, and Dutch, we only calculated one dendrogram, as the only difference is that the Spanish alphabet contains the "ñ" letter. For Japanese, we show a random subset (50%) of the *Hiragana* alphabet, as it did not fit properly in our plots.

C Models

C.1 Training Procedure

For the LSTM, biLSTM and transformer models, the number of epochs is set to 100, but the models are trained with early stopping and training is ended after 3 epochs without improvement on the validation loss. The best model is saved and used to extract the character embeddings. For reproducibility purposes, we set a single random seed.

C.2 Perplexity

Table 7 reports the performance of the models in terms of per-character perplexity (PPL), defined as the base-2 exponentiation of the cross-entropy.

	Model	PPL
en	LSTM	170.99
	Transformer	80.22
es	LSTM	83.81
	Transformer	56.22
nl	LSTM	68.28
	Transformer	45.92
ko	LSTM	106.91
	Transformer	76.45
ja	LSTM	268.84
	Transformer	625.29

Table 7: Model perplexities on validation set.

C.3 Additional Experiments

We experimented with adding additional layers in the LSTM models. The results show slight differ-

[4] https://www.tensorflow.org/datasets/catalog/wiki40b
[5] https://pypi.org/project/jamotools/
[6] https://pypi.org/project/pykakasi/

ences in the Pearson correlation coefficients, but the general trends remains the same.

We also experimented with more heads (4 instead of 2) in the transformer models as well as taking the representations from the embeddings layers instead of the hidden layer. However, the results did not yield significant changes.

D Grapheme-to-Phoneme Alignments

Resulting grapheme-to-phoneme alignments from the WikiPron dataset retrieved by choosing the most frequent phone mapping of a character based on an unsupervised alignment of the data.

D.1 Dutch

a:ɑ, b:b, c:k, d:d, e:ə, f:f , g:ɣ, h:ɦ, i:ɪ, j:ɛi̯ , k:k , l:l , m:m , n:n , o:ɔ, p:p , r:r , s:s , t:t , u:ʏ, v:v , w:ʋ, x:ks, z:z , q:k , y:i

D.2 English

a:ə, b:b , c:k , d:d , e:ɛ, f:f , g:g , h:h , i:, j:ʤ, k:k , l:l , m:m , n:n , o:ə, p:p, q:k , r:ɹ, s:s , t:t , u:ʌ, v:v , w:w , x:ks , y:i , z:z

D.3 Korean

ㄱ:k̚, ㄲ:k͈, ㄴ:n, ㄷ:d, ㄸ:t͈, ㄹ:ɭ, ㅁ:m, ㅂ:p, ㅃ:p͈, ㅅ:s^h, ㅆ:s͈, ㅇ:■, ㅈ:dʑ, ㅉ:tɕ, ㅊ:tɕ^h, ㅋ:k^h, ㅌ:t^h, ㅍ:p^h, ㅎ:ɦ, ㅀ:ɭ, ㅄ:p̚, ㅏ:a̠, ㅐ:eː, ㅑ:a̠, ㅒ:ɛː, ㅓ:ʌ̹, ㅔ:e̞, ㅕ:ʌ̹, ㅖ:e̞, ㅗ:o̞, ㅘ:a̠, ㅙ:ɛː, ㅚ:e̞, ㅛ:o, ㅜ:u, ㅝ:ʌ̹, ㅞ:e̞, ㅟ:i, ㅠ:u, ㅡ:ɯ, ㅢ:i, ㅣ:i, ㄳ:k̚, ㄵ:n, ㄶ:n, :ɭ, ㄺ:k̚, ㄻ:m, ㄼ:ɭ, ㄽ:ɭ, ㄾ:ɭ, ㄿ:p̚, :m

D.4 Spanish

a:a , b:b , c:k , d:d , e:e , f:f , g:g , h:x , i:i , j:x , k:k , l:l , m:m , n:n , o:o , p:p , q:k , r:ɾ, s:s , t:t , u:u , v:b , w:w , x:ks , y:ɟ͡ʝ, z:θ

D.5 Japanese

For the Japanese alignments, we refer to the Japanese log file in the `ipa_embeddings` folder in the code repository.

E Probing Task

The results of all the probing task on all distinctive features are found in Table 8.

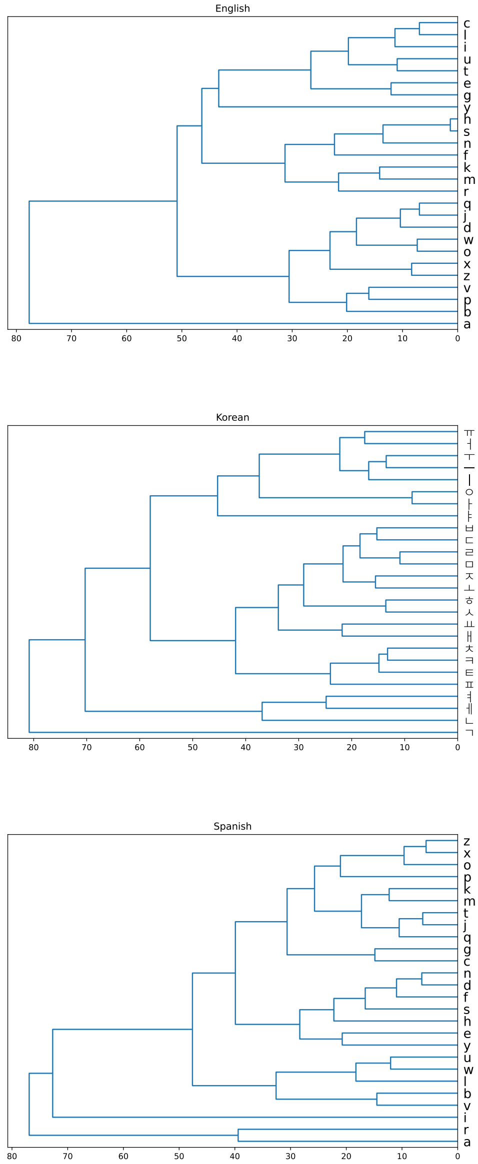

Figure 4: Dendrograms of the distances between colors assigned to each character for English, Korean and Spanish. The leaves are sorted so that the minimum distance between its direct descendants is plotted first.

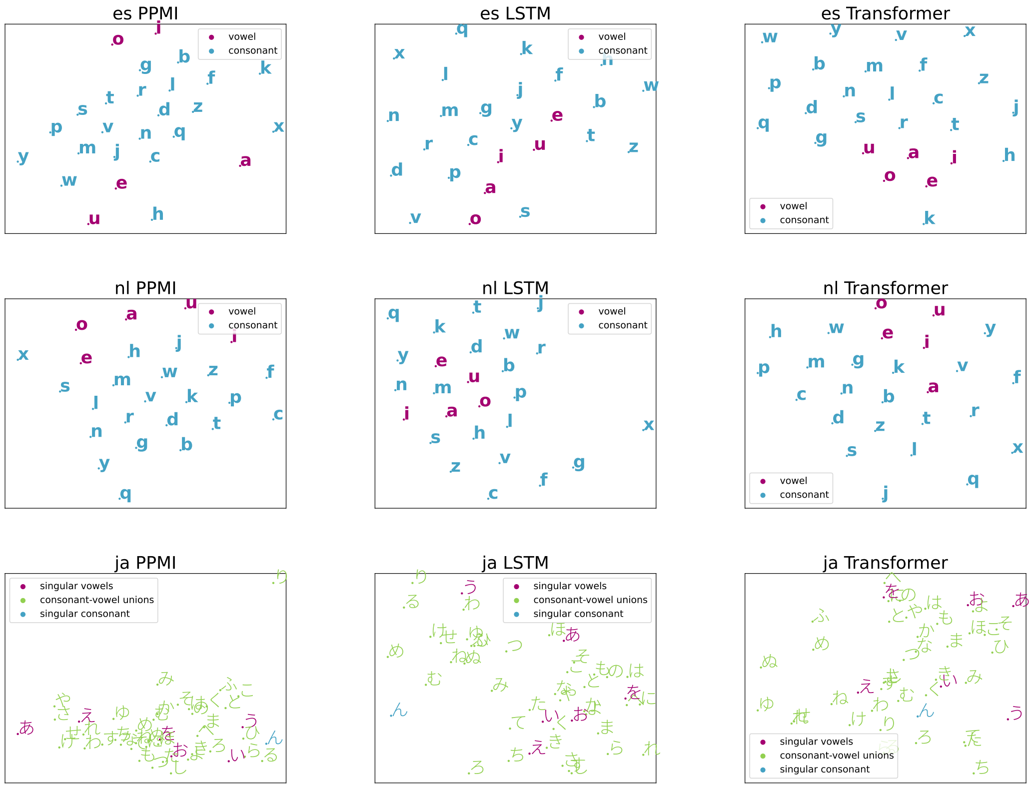

Figure 5: tSNE cluster plots of the three types of character models for Spanish, Dutch and Japanese.

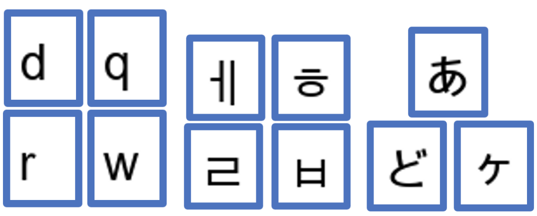

Figure 6: Example images from which we extract character shape representations from the Latin alphabets, the Korean *Hangul* alphabet and the Japanese *Hiragana* alphabet.

	Model	global type		consonant voicing		consonant place					
		consonant	*vowel*	*voiced*	*voiceless*	*alveolar*	*alveolo-palatal*	*bilabial*	*labio-dental*	*palatal*	*velar*
en	LSTM	0.95	0.80	0.55	0.44	0.53	-	0.00	0.00	-	0.29
	Transformer	0.97	0.92	0.75	0.63	0.38	-	0.00	0.00	-	0.00
	Random	0.60	0.32	0.52	0.48	0.47	-	0.22	0.16	-	0.33
	Most-frequent	0.87	0.00	0.71	0.00	0.00	-	0.00	0.00	-	0.00
es	LSTM	0.96	0.75	0.40	0.50	0.33	-	0.40	-	0.00	0.46
	Transformer	0.98	0.89	0.72	0.63	0.62	-	0.00	-	0.00	0.00
	Random	0.61	0.27	0.49	0.49	0.38	-	0.26	-	0.15	0.39
	Most-frequent	0.90	0.00	0.00	0.00	0.00	-	0.00	-	0.00	0.00
ko	LSTM	1.00	1.00	0.50	0.65	0.47	0.00	0.00	-	-	0.00
	Transformer	1.00	1.00	0.43	0.47	0.29	0.00	0.17	-	-	0.00
	Random	0.53	0.46	0.46	0.53	0.41	0.16	0.32	-	-	0.29
	Most-frequent	0.73	0.00	0.00	0.71	0.00	0.00	0.00	-	-	0.00
nl	LSTM	0.97	0.92	0.80	0.62	0.67	-	0.00	0.00	-	0.33
	Transformer	0.97	0.92	0.75	0.57	0.57	-	0.00	0.00	-	0.57
	Random	0.59	0.35	0.53	0.45	0.45	-	0.24	0.24	-	0.34
	Most-frequent	0.84	0.00	0.73	0.00	0.00	-	0.00	0.00	-	0.00

	Model	consonant manner				
		approximant	*nasal*	*non-sibilant-fricative*	*plosive*	*sibilant-fricative*
en	LSTM	0.00	0.00	0.00	0.80	0.00
	Transformer	0.00	0.00	0.00	0.74	0.00
	Random	0.17	0.18	0.22	0.50	0.28
	Most-frequent	0.00	0.00	0.00	0.00	0.00
es	LSTM	-	0.00	0.00	0.50	0.00
	Transformer	-	0.00	0.00	0.33	0.00
	Random	-	0.21	0.26	0.47	0.15
	Most-frequent	-	0.00	0.00	0.00	0.00
ko	LSTM	-	0.00	-	0.50	0.00
	Transformer	-	0.40	-	0.67	0.00
	Random	-	0.29	-	0.51	0.22
	Most-frequent	-	0.00	-	0.00	0.00
nl	LSTM	-	0.00	0.00	0.38	0.00
	Transformer	-	0.00	0.00	0.40	0.00
	Random	-	0.17	0.29	0.45	0.25
	Most-frequent	-	0.00	0.00	0.00	0.00

	Model	vowel height				vowel backness		vowel rounding	
		close	*close-mid*	*mid*	*open-mid*	*front*	*back*	*rounded*	*unrounded*
en	LSTM	-	-	0.00	0.00	0.00	-	-	-
	Transformer	-	-	0.00	0.00	0.00	-	-	-
	Random	-	-	0.37	0.37	0.39	-	-	-
	Most-frequent	-	-	0.00	0.00	0.00	-	-	-
es	LSTM	0.00	0.00	-	-	0.75	0.00	0.00	0.75
	Transformer	0.00	0.00	-	-	0.33	0.00	0.00	0.33
	Random	0.41	0.42	-	-	0.50	0.42	0.43	0.49
	Most-frequent	0.00	0.00	-	-	0.00	0.00	0.00	0.00
ko	LSTM	0.00	-	0.00	0.20	0.56	0.25	0.00	0.83
	Transformer	0.00	-	0.00	0.00	0.60	0.13	0.00	0.86
	Random	0.36	-	0.32	0.36	0.48	0.43	0.27	0.62
	Most-frequent	0.00	-	0.00	0.00	0.00	0.00	0.00	0.89
nl	LSTM	0.00	-	-	0.00	0.00	0.00	0.00	0.83
	Transformer	0.00	-	-	0.00	0.00	0.00	0.00	0.83
	Random	0.34	-	-	0.35	0.34	0.35	0.35	0.57
	Most-frequent	0.00	-	-	0.00	0.00	0.00	0.00	0.83

Table 8: F1 score for classifiers predicting distinctive features with character embeddings (LSTM, Transformer) as input. Two baselines are included: Random (predicting labels uniformly at random) and Most-frequent (always predicting the most frequent label). A language/feature combination with "-" indicates that no classifier was trained due to the lack of examples.

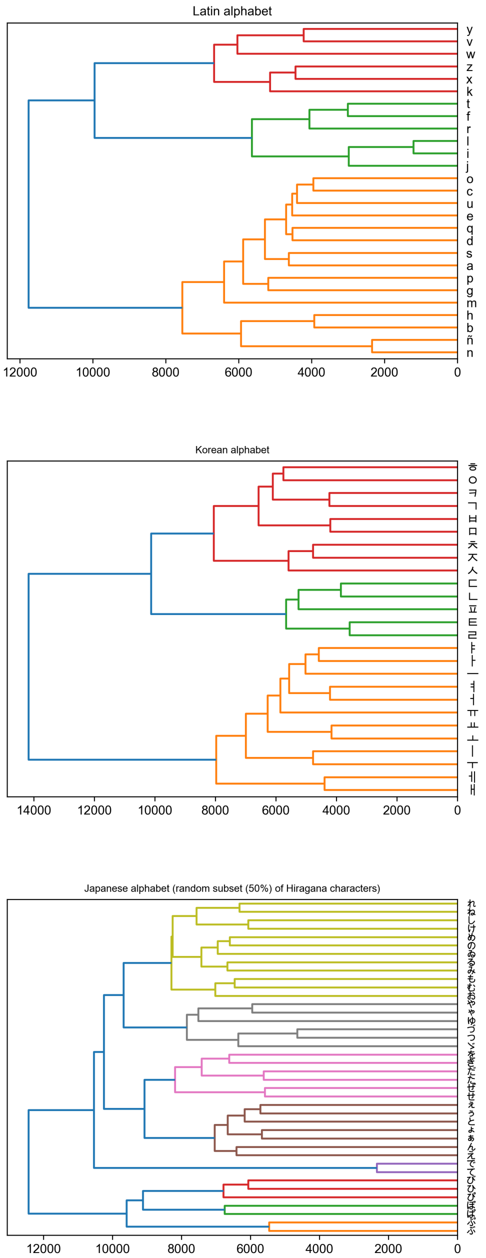

Figure 7: Dendrograms of the distances between shape representations for the Latin alphabet (including the Spanish ñ letter), Korean *Hangul* alphabet and a subset of the Japanese *Hiragana* alphabet.