

PWESUITE: Phonetic Word Embeddings and Tasks They Facilitate

Vilém Zouhar^E Calvin Chang^C Chenxuan Cui^C Nathaniel Carlson^Y
 Nathaniel R. Robinson^C Mrinmaya Sachan^E David Mortensen^C

^EDepartment of Computer Science, ETH Zurich

^CLanguage Technologies Institute, Carnegie Mellon University

^YDepartment of Computer Science, Brigham Young University

{vzouhar, msachan}@ethz.ch natbcar@gmail.com

{kalvinc, cxcui, nrrobbins, dmortens}@cs.cmu.edu

Abstract

Mapping words into a fixed-dimensional vector space is the backbone of modern NLP. While most word embedding methods successfully encode semantic information, they overlook phonetic information that is crucial for many tasks. We develop three methods that use articulatory features to build phonetically informed word embeddings. To address the inconsistent evaluation of existing phonetic word embedding methods, we also contribute a task suite to fairly evaluate past, current, and future methods. We evaluate both (1) intrinsic aspects of phonetic word embeddings, such as word retrieval and correlation with sound similarity, and (2) extrinsic performance on tasks such as rhyme and cognate detection and sound analogies. We hope our task suite will promote reproducibility and inspire future phonetic embedding research.

Keywords: phonetic word embeddings, representation learning, phonology, articulatory features, evaluation

Code: github.com/zouharvi/pwesuite
 Dataset: huggingface.co/datasets/zouharvi/pwesuite-eval

1. Introduction

Word embeddings are omnipresent in modern NLP (Le and Mikolov, 2014; Pennington et al., 2014; Almeida and Xexéo, 2019, inter alia). Their main benefit lies in compressing some information into fixed-dimensional vectors. These vectors can be used as machine-learning features for NLP applications, and their study can reveal linguistic insights (Hamilton et al., 2016; Ryskina, Maria and Rabinovich, Ella and Berg-Kirkpatrick, Taylor and Mortensen, David R. and Tsvetkov, Yulia, 2020; Francis et al., 2021). Word embeddings are often trained via methods from distributional semantics (Camacho-Collados and Pilehvar, 2018) and thus bear semantic information. For example, the embedding for the word *carrot* may encode higher similarity to embeddings for other vegetables than to that of *ocean*.

Some applications may require a different type of information to be encoded. The orthography, especially in English, can obscure the pronunciation. A poem generation model, for instance, may need embeddings to reflect that *ocean* rhymes with *motion* and not with a *soybean*, even though the spelling of the words' final syllables suggest otherwise (see Figure 1). Such embeddings, called

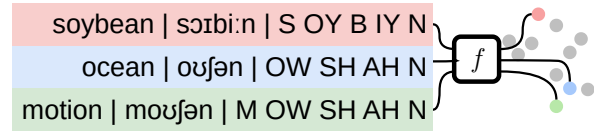


Figure 1: Embedding function f projects words in various forms (left) to a vector space (right).

phonetic word embeddings, contain phonetic information and have been of recent interest (Parish, 2017; Yang and Hirschberg, 2019; Hu et al., 2020; Sharma et al., 2021).¹ The objective is that words with similar pronunciation should be mapped to vectors near each other in embedding space. Many tasks have benefited from incorporating phonetic word embeddings, including cognate and loanword detection (Rama, 2016; Nath et al., 2022b,a), named entity recognition (Bharadwaj et al., 2016; Chaudhary et al., 2018), spelling correction (Zhang et al., 2021), and speech recognition (Fang et al., 2020). See Section 6.2 for a more detailed list of possible applications.

We introduce four phonetic word embedding methods—count-based, autoencoder, and metric and contrastive learning. Though some of these techniques are inspired by previous work, we are the first to apply them with supervision from articulatory feature vectors, a seldom-exploited form of

^ECo-first authors.

¹The technically correct term is *phonological word embeddings* but prior literature uses the term *phonetic*.

linguistic knowledge for representation learning.

More importantly, we introduce an evaluation suite for testing the performance of phonetic embeddings. The motivations for this are twofold. First, prior work is inconsistent in evaluating models. This prevents the field from observing long-term improvements in such embeddings and from making fair comparisons across different approaches. Secondly, when a practitioner is deciding which phonetic word embedding method to use, the go-to approach is to first apply the embeddings (generally fast) and then train a downstream model on those embeddings (compute and time intensive). Instead, intrinsic embedding evaluation metrics (cheap)—if shown to correlate well with extrinsic metrics—could provide useful signals in embedding method selection prior to training of downstream models (expensive). In contrast to semantic word embeddings (Bakarov, 2018), we show that intrinsic and extrinsic metrics for phonetic word embeddings generally correlate with each other. While Ghannay et al. (2016) evaluate acoustic word embeddings, we specialize in phonetic word embeddings *for text, not speech*.

Our main contribution is this evaluation suite for phonetic word embeddings, the equivalent of which does not yet exist in this subfield. We also contribute multiple methods for and a survey of existing phonetic word embeddings.

2. Survey of Phonetic Embeddings

Given an alphabet Σ and a dataset of words $\mathcal{W} \subseteq \Sigma^*$, d -dimensional word embeddings are given by a function $f : \mathcal{W} \rightarrow \mathbb{R}^d$. This function takes an element from Σ^* (set of all possible words over the alphabet Σ) and maps it to a d -dimensional vector of numbers. For many embedding functions, \mathcal{W} is a finite set of words, and the embeddings are not defined for unseen words (Mikolov et al., 2013a; Pennington et al., 2014). Other embedding functions—which we dub *open*—are able to provide an embedding for any word $x \in \Sigma^*$ (Bojanowski et al., 2017). An illustration of a *phonetic* embedding function is shown in Figure 1 (*motion* is closer to *ocean* than to *soybean*).

We use 3 distinct alphabets: characters Σ_C , IPA symbols Σ_P and ARPAbet symbols Σ_A . We use Σ when the choice is not important and refer to elements of Σ as characters or phonemes. We review some *semantic* embeddings in Section 5 and now focus on prior work in *phonetic* embeddings. From our formalism it also follows that we are interested in phonetic representations of *textual* input.

2.1. Poetic Sound Similarity

Parrish (2017) learns word embeddings captur-

ing pronunciation similarity for poetry generation for words in the CMU Pronouncing Dictionary (Group, 2014). First, each phoneme is mapped to a set of phonetic features \mathcal{F} using the function $P2F : \Sigma_A \rightarrow 2^{\mathcal{F}}$. From the sequence of sets that each sequence of phonemes maps to, bi-grams of phonetic features are created (using Cartesian product \times between sets a_i and a_{i+1}) and counted. The function COUNTVEC outputs these bi-gram counts in a vector of constant dimension. The resulting vector is then reduced using PCA to the target embedding dimension d .

$$W2F(x) = \langle P2F(x_i) | x_i \in x \rangle \quad (\text{array}) \quad (1)$$

$$F2V(a) = \text{COUNTVEC}(\bigcup_{1 \leq i \leq |a|-1} a_i \times a_{i+1}) \quad (2)$$

$$f_{\text{PAR}} = \text{PCA}_d(\{F2V(W2F(x)) | x \in \mathcal{W}\}) \quad (3)$$

The function f_{PAR} can provide embeddings even for words unseen during training. This is because the only component dependent on the training data is the PCA over the vector of bigram counts, which can also be applied to new vectors.

2.2. phoneme2vec

Fang et al. (2020) do not use hand-crafted features and learn phoneme embeddings using a more complex, deep-learning, model. They start with a gold sequence of phonemes (x_i) and a noisy sequence of phonemes (y_i). The phonemes are one-hot encoded in matrices X and Y . The gold sequence is first read by an LSTM model, yielding the initial hidden state h_0 . From this hidden state, the phonemes (\hat{y}_i) are decoded using teacher forcing (upon predicting \hat{y}_i , the model receives the correct x_i as the input). The phoneme embedding matrix V is trained jointly with the model weights and constitutes the embedding function.

$$h_0 = \text{LSTM}(XV) \quad (4)$$

$$\mathcal{L}_{\text{p2v}} = -\sum_{0 < i \leq |y|} \log \text{softmax}(\text{LSTM}(Y_{<i}V)_{y_i}) \quad (5)$$

For a fair comparison, we average these vectors which are *phoneme*-level to get word-level embeddings. In addition, in contrast to other embeddings, these phoneme embeddings are only 50-dimensional. We revisit the question of dimensionality in Section 5.5.

2.3. Phonetic Similarity Embeddings

Sharma et al. (2021) propose a vowel-weighted phonetic similarity metric to compute similarities between words. They then use it for training phonetic word embeddings which should share some properties with this similarity function. This is in contrast to the previous approaches, where the

embedding training is indirect on an auxiliary task. Given a sound similarity function S_{PSE} , they construct a matrix of similarity scores $S \in \mathbb{R}^{|\mathcal{W}| \times |\mathcal{W}|}$ such that $S_{i,j} = S_{\text{PSE}}(\mathcal{W}_i, \mathcal{W}_j)$. On this matrix, they use non-negative matrix factorization to learn the embedding matrix $V \in \mathbb{R}^{|\mathcal{W}| \times d}$ such that the following loss is minimized:

$$\mathcal{L}_{\text{PSE}} = \|S - V \cdot V^T\|^2 \quad (6)$$

Then, the i -th row of V contains the embedding for i -th word from \mathcal{W} . A critical disadvantage of this approach is that it cannot be used for embedding new words because the matrix V would need to be recomputed again. We apply the sound similarity function S_{PSE} , defined specifically for English, to all evaluation languages.

3. Our Models

We now introduce several embedding baselines. Then, we describe our articulatory distance metric and models trained with supervision therefrom.

3.1. Count-based Vectors

Perhaps the most straightforward way of creating a vector representation for a sequence of input characters or phonemes $x \in \Sigma^*$ is simply counting n-grams in this sequence. We use a term frequency-inverse document frequency (TF-IDF) vectorizer of 1-, 2-, and 3-grams (formally denoted $[x]_n$) across the input sequence of symbols (e.g. characters) with a maximum of 300 features. This vector then becomes our word embedding. For instance, the first dimension may be the TF-IDF score or occurrence count of the bigram $\langle /dɪn/, /a/ \rangle$.

$$\text{C2V}(x) = [x]_1 \cup [x]_2 \cup [x]_3 \quad (\text{features}) \quad (7)$$

$$f_{\text{count}}(x) = \text{TF-IDF}_{\text{features}}^{\text{features}}(\{\text{C2V}(x) | x \in \mathcal{W}\}) \quad (8)$$

3.2. Autoencoder

Another common approach, though less interpretable, for vector representation with fixed dimension size is an encoder-decoder autoencoder. Specifically, we use this architecture together with forced-teacher decoding and use the bottleneck vector as the phonetic word embedding. In an ideal case, the fixed-size bottleneck contains all the information to reconstruct the whole sequence from Σ^* .

$$f_{\theta}(x) = \text{LSTM}(x | \theta) \quad (\text{encoder}) \quad (9)$$

$$d_{\theta'}(x) = \text{LSTM}(x | \theta') \quad (\text{decoder}) \quad (10)$$

$$\mathcal{L}_{\text{auto.}} = \sum_{0 < i \leq |x|} -\log \text{softmax}(d_{\theta'}(f_{\theta}(x)) | x_{<i} x_i) \quad (11)$$

3.3. Phonetic Word Embeddings With Articulatory Features

3.3.1. Articulatory Features and Distance

Articulatory features (Bloomfield, 1993; Jakobson et al., 1951; Chomsky and Halle, 1968) decompose sounds into their constituent properties. Each segment can be mapped to a vector with n different features (24 for PanPhon Mortensen et al., 2016) such as whether the phoneme segment is produced with a nasal airflow or if it is produced with raised or lowered tongue tip. A segment is a group of phonetic characters (e.g., as defined by Unicode) that represent a single sound. We define $\alpha: \Sigma_P \rightarrow \{-1, 0, +1\}^{24}$ as the function which maps a phoneme segment into a vector of articulatory features. Values $+1/-1$ mean present/not present and the value 0 is used when the feature is irrelevant.

The articulatory distance, also called *feature edit distance* (Mortensen et al., 2016), is a version of Levenshtein distance with custom costs. Specifically, the substitution cost is proportional to the Hamming distance between the source and target when they are represented as articulatory feature vectors. Omitting edge-cases, it is defined as:

$$A_{i,j}(x, x') = \min \begin{cases} A_{i-1,j}(x, x') + d(x) \\ A_{i,j-1}(x, x') + i(x') \\ A_{i-1,j-1}(x, x') + s(x_i, x'_j) \end{cases} \quad (12)$$

$$A(x, x') = A_{|x|, |x'|}(x, x') \quad (13)$$

where d and i are deletion and insertion costs, which we set to constant 1. The function s is a substitution cost, defined as the number of elements (normalized) that need to be changed to render the two articulatory vectors identical:

$$s(x, x') = \frac{1}{24} \sum_{i=1}^{24} |\alpha(x)_i - \alpha(x')_i| \quad (14)$$

The articulatory distance A induces a metric space-like structure for words in Σ^* . It quantifies the phonetic similarity between a pair of words, capturing the intuition that $/pæt/$ and $/bæt/$ are phonetically closer than $/pæt/$ and $/hæt/$, for example.

3.3.2. Metric Learning

As one means of generating word embeddings, we use the last hidden state of an LSTM-based model. We use characters Σ_C , IPA symbols Σ_P (Section 2) and articulatory feature vectors as the input. We discuss these choices and especially their effect on performance and transferability in Section 5.3.

We now have a function f that produces a vector for each input word. However, it is not yet trained to produce vectors encoding phonetic information.

We, therefore, define the following differentiable loss where A is the articulatory distance.

$$\mathcal{L}_{\text{dist.}} = \frac{1}{|\mathcal{W}|} \sum_{\substack{x_a \in \mathcal{W} \\ x_b \sim \mathcal{W}}} \left(\|f_\theta(x_a) - f_\theta(x_b)\|^2 - A(x_a, x_b) \right)^2 \quad (15)$$

This forces the embeddings to be spaced in the same way as the articulatory distance (A , Section 3.3.1) would space them. Metric learning (learning a function to space output vectors similarly to some other metric) has been employed previously (Yang and Jin, 2006; Bellet et al., 2015; Kaya and Bilge, 2019) and was used to train *acoustic* embeddings by Yang and Hirschberg (2019).

3.3.3. Triplet Margin loss

While the previous approach forces the embeddings to be spaced exactly as by the articulatory distance function A , we may relax the constraint so only the structure (ordering) is preserved. This is realized by triplet margin loss:

$$\mathcal{L}_{\text{triplet}} = \max \begin{cases} 0 \\ \alpha + |f_\theta(x_a) - f_\theta(x_p)| \\ -|f_\theta(x_a) - f_\theta(x_n)| \end{cases} \quad (16)$$

We consider all possible ordered triplets of distinct words (x_a, x_p, x_n) such that $A(x_a, x_p) < A(x_a, x_n)$. We refer to x_a as the anchor, x_p as the positive example, and x_n as the negative example. We then minimize $\mathcal{L}_{\text{triplet}}$ over all valid triplets. This allows us to learn θ for an embedding function f_θ that preserves the local neighbourhoods of words defined by $A(x, x')$. In addition, we modify the function f_θ by applying attention to all hidden states extracted from the last layer of the LSTM encoder. This allows our model to focus on phonemes that are potentially more useful when trying to summarize the phonetic information in a word. A related approach was used by Yang and Hirschberg (2019) to learn acoustic word embeddings. Although contrastive learning is a more intuitive approach, it yielded only negative results: $(\exp(\|f_\theta(x_a) - f_\theta(x_p)\|^2)) / (\sum \exp(\|f_\theta(x_a) - f_\theta(x_n)\|^2))$.

Though metric learning and triplet margin loss have been applied previously to similar applications, we are the first to apply them using articulatory features and articulatory distance.

3.4. Phonetic Language Modeling

To shed more light into the true landscape of phonetic word embedding models, we describe here a model which did not perform well on our suite of tasks (in contrast to other models). A common

way of learning word embeddings now is to train on the masked language model objective, popularized by BERT (Devlin et al., 2019). We input articulatory features from PanPhon into several successive Transformer (Vaswani et al., 2017) encoder layers and a final linear layer that predicts the masked phone. Positional encoding is added to each input. We prepend and append [CLS] and [SEP] tokens, respectively, to the phonetic transcriptions of each word, before we look up each phone’s PanPhon features. We use [CLS] pooling—taking the output of the Transformer corresponding to the first token—to extract a word-level representation. Unlike BERT, we do not train on the next sentence prediction objective. In addition, we do not add an embedding layer because we are not interested in learning individual phone embeddings but rather wish to learn a word-level embedding.

4. Evaluation Suite (key contribution)

We now introduce the embedding evaluation metrics of our suite, the primary contribution of this paper. We draw inspiration from evaluating semantic word embeddings (Bakarov, 2018) and work on phonetic word embeddings (Parrish, 2017). In some cases, the distinction between intrinsic and extrinsic evaluations is tenuous (e.g., retrieval and analogies). The main characteristic of intrinsic evaluation is that they are efficiently computed and are not part of any specific application. In contrast, extrinsic evaluation metrics directly measure the usefulness of the embeddings for a particular task.

We evaluate with 9 phonologically diverse languages: Amharic*, Bengali*, English, French, German, Polish, Spanish, Swahili, and Uzbek. Languages marked with * use non-Latin script. The non-English data (200k tokens each) is from CC-100 (Wenzek et al., 2020; Conneau et al., 2020), while the English data (125k tokens) is from the CMU Pronouncing Dictionary (Group, 2014).

4.1. Intrinsic Evaluation

4.1.1. Articulatory Distance

The unifying desideratum for phonetic embeddings is that they should capture the concept of sound similarity. Recall from Section 2 that phonetic word embeddings are a function $f : \Sigma^* \rightarrow \mathbb{R}^d$. In the vector space of \mathbb{R}^d , there are two widely used notions of similarity S . The first is the *negative L_2 distance* and the other is the *cosine similarity*. Consider three words x , x' and x'' . Using either metric, $S(f(x), f(x'))$ yields the embedding similarity between x and x' . On the other hand, since we have prior notions of similarity S_P between the words, e.g., based on a rule-based function, we can use this to represent the similarity between the

words: $S_P(x, x')$. We want to have embeddings f such that $S \circ f$ produces results close to S_P . There are at least two ways to verify that the similarity results are close. First is exact equality. For example, if $S_P(x, x') = 0.5$, $S_P(x, x'') = 0.1$, we want $S(f(x), f(x')) = 0.5$, $S(f(x), f(x'')) = 0.1$. We can measure this using Pearson's correlation coefficient between $S \circ f$ and S_P . On the other hand, we may consider only the relative similarity values. Following the previous example, we would only care that $S(f(x), f(x')) > S(f(x), f(x''))$. In this case we use Spearman's correlation coefficient between $S \circ f$ and S_P . For the rule-based similarity metric S_P , we use *articulatory distance* (Mortensen et al., 2016), as described in Section 3.3.1.

4.1.2. Human Judgement

Vitz and Winkler (1973) asked people to judge the sound similarity of English words. For selected word pairs, we denote the collected judgements (scaled from 0—least similar to 1—identical) with the function S_H . For example, $S_H(\text{slant}, \text{plant}) = 0.9$ and $S_H(\text{plots}, \text{plant}) = 0.4$. Like the previous task, we find correlations between $S \circ f$ and S_H . We note S_H judgments were produced from a small English-only corpus. These limitations highlight the importance of including analyses with A , rather than S_H alone. In fact, A and S_H do not correlate positively, with Pearson coefficient -0.74 .

4.1.3. Retrieval

An important usage of word embeddings is the retrieval of associated words, which is also utilized in the analogies extrinsic evaluation and other applications. Success in this task means that the new embedding space has the same local neighbourhood as the original space induced by some non-vector-based metric. Given a word dataset \mathcal{W} and one word $w \in \mathcal{W}$, we sort $\mathcal{W} \setminus \{w\}$ based on both $S \circ f$ and S_P distance from w . Based on this ordering, we define the immediate neighbour of w based on S_P , denoted w_N and ask the question *What is the average rank of w_N in the ordering by $S \circ f$?* If the similarity given by $S \circ f$ is copying S_P perfectly, then the rank will be 0 because w_N will be the closest to w in $S \circ f$.

Again, for S_P we use the articulatory distance A (Section 3.3.1). Even though there are a variety of possible metrics to evaluate retrieval, we focus on the average rank. We further cap the retrieval neighborhood at $n = 1000$ samples and compute percentile rank as $\frac{n-r}{n}$. This choice is done so that the metric will be bounded between 0 (worst) and 1 (best), which will become important for overall evaluation later (Section 4.3).

Error analysis. We identify two types of errors in the retrieval task for the *Metric Learner* model with articulatory features. The first one are simply incorrect neighbours with low sound similarity, such as the word *carcass*, whose correct neighbour is *cardiss* but for which *krutick* is chosen. The next group are plausible ones, such as for the word *counterrevolutionary*, its neighbour in articulatory distance space *counterinsurgency* and the retrieved word *cardiopulmonary*. In this case we might even say that the retrieved word is closer.

4.2. Extrinsic Evaluation

4.2.1. Rhyme Detection

There are multiple types of word rhymes, most of which are based around two words sounding similarly. We focus on perfect rhymes: when the sounds from the last stressed syllables are identical. An example is *grown* and *loan*, even though the surface character form does not suggest it. Clearly, this task can be deterministically solved if one has access to the articulatory and stress information of the concerned words. Nevertheless, we wish to evaluate whether this information can be encoded in a fixed-length vector produced by f . We create a balanced binary prediction task for rhyme detection in English and train a small multi-layer perceptron classifier on top of pairs of word embeddings. The linking hypothesis is that the higher the accuracy, the more useful information for the task there is in the embeddings.

4.2.2. Cognate Detection

Cognates are words in different languages that share a common origin. We include *loanwords* alongside genetic cognates. Similarly to rhyme detection, we frame cognate detection as a binary classification task where the input is a potential cognate pair. CogNet (Batsuren et al., 2019) is a large cognate dataset of many languages, making it ideal to evaluate the usefulness of phonetic embeddings. We add non-cognate, distractor pairs in the dataset by finding the orthographically closest word that is not a known cognate. For example, *plant*_{EN} and *plante*_{FR} are cognates, while *plant*_{EN} and *plane*_{EN} are not. Although cognates also preserve some of the similarities in the meaning, we detect them using phonetic characteristics only.

4.2.3. Sound Analogies

Just as distributional semantic vectors can complete word-level analogies such as *man* : *woman* \leftrightarrow *king* : *queen* (Mikolov et al., 2013b), so too should well-trained phonetic word embeddings capture sound analogies. For example of a sound analogy, consider */din/* : */tin/* \leftrightarrow */zin/* : */sin/*. The

		INTRINSIC			EXTRINSIC		OVERALL	
Model		Human Sim.	Art. Dist.	Retrieval	Analogies	Rhyme	Cognate	
		(Pearson)	(Pearson)	(rank perc.)	(Acc@1)	(accuracy)	(accuracy)	
Ours	Metric Learner	0.46	0.94	0.98	84%	83%	64%	0.78
	Triplet Margin	0.65	0.96	1.00	100%	77%	66%	0.84 *
	Count-based	0.82	0.10	0.84	13%	79%	68%	0.56
	Autoencoder	0.49	0.16	0.73	50%	61%	50%	0.50
Others'	Poetic Sound Sim.	0.74	0.12	0.78	35%	60%	57%	0.53
	phoneme2vec	0.77	0.09	0.80	17%	88%	64%	0.56
	Phon. Sim. Embd.	0.16	0.05	0.50	0%	51%	52%	0.29
Semantic	BPEmb	0.23	0.08	0.60	5%	54%	66%	0.36
	fastText	0.25	0.12	0.64	2%	58%	68%	0.38
	BERT	0.10	0.34	0.69	4%	58%	63%	0.40
	INSTRUCTOR	0.60	0.12	0.73	7%	54%	66%	0.45

Table 1: Embedding method performance in our evaluation suite. Higher number is always better.

difference within the pairs is $[\pm\text{voice}]$ in the first phoneme segment of each word.

With this intuition in mind, we define a *perturbation* as a pair of phonemes (p, q) differing in one articulatory feature. We then create a sound analogy corpus of 200 quadruplets $w_1 : w_2 \leftrightarrow w_3 : w_4$ for each language, with the following procedure:

1. Choose a random word $w_1 \in \mathcal{W}$ and one of its phonemes on random position i : $p_1 = w_{1,i}$.
2. Randomly select two perturbations of the same phonetic feature so that $p_1 : p_2 \leftrightarrow p_3 : p_4$, for example $/t/ : /d/ \leftrightarrow /s/ : /z/$.
3. Create w_2 , w_3 , and w_4 by duplicating w_1 and replacing $w_{1,i}$ with p_2 , p_3 , and p_4 . The new words w_2 , w_3 , and w_4 do not have to be a real word in the language but we are still interested in analogies in the space of all possible words and their detection. This is possible only for *open* embeddings.

We apply the above procedure 1 or 2 times to create 200 analogous quadruplets with 1 or 2 perturbations (evenly split). We then measure the Acc@1 to retrieve w_4 from $\mathcal{W} \cup \{w_4\}$. We simply measure how many often the closest neighbour of $w_2 - w_1 + w_3$ is w_4 . Our analogy task is different from that of Parrish (2017) who focused on morphological derivation² and that of Silfverberg et al. (2018), which show that phoneme embeddings learned via the word2vec objective demonstrate sound analogies at the *phoneme* level. We consider sound analogies at the *word* level.

4.3. Overall Score

Since all the measured metrics are bounded between 0 and 1, we can define the *overall* score

²Example *decide* : *decision* \leftrightarrow *explode* : *explosion*.

for our evaluation suite as the arithmetic average of results from each task. We mainly consider the results of all available languages averaged but later in Section 5.3 discuss results per language as well. To allow for future extensions in terms of languages and tasks, this evaluation suite is versioned, with the version described in this paper being v1.0.

5. Evaluation

We now compare all the aforementioned embedding models using our evaluation suite. We show the results in Table 1 with three categories of models. Our models trained using some articulatory features or distance supervision (Section 3) are given first, followed by other phonetic word embedding models (Section 2). We also include non-phonetic word embeddings, not as a fair baseline for comparison but to show that these embeddings are different from phonetic word embeddings and are not suited for our tasks: fastText (Grave et al., 2018), BPEmb (Heinzerling and Strube, 2018), BERT (Devlin et al., 2019) and INSTRUCTOR (Su et al., 2022). We chose these embeddings because they are open (i.e., they provide embeddings even to words unseen in the training data). All of these embeddings except for BERT and INSTRUCTOR are 300-dimensional (see Section 5.5).

5.1. Model Comparison

In Table 1 we show the performance of all previously described models. The *Triplet Margin* model is best overall, outperforming *Metric Learner*, despite its less direct supervision in training. However, it also requires the longest time to train.³

³The overall GPU budget for all included experiments is 100 hours on GTX 1080 Ti. We include reproducibility details in the code repository.

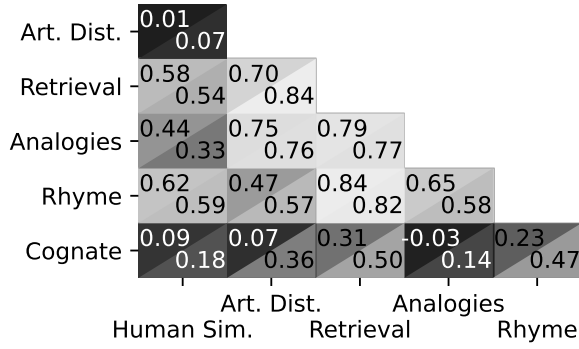


Figure 2: Spearman (upper left) and Pearson (lower right) correlations between performance on suite tasks. All models from Table 1 are used.

Surprisingly, the best model for human similarity is a simple *count-based* model. Semantic word embeddings perform worse than explicit phonetic embeddings, most notably on human similarity and analogies. However, they do perform reasonably on cognate detection.

We now examine how much the performance on one task (particularly an intrinsic one) is predictive of performance on another task. We measure this across all systems in Table 1 and revisit this topic later for creating variations of the same model. For lexical/semantic word embeddings, [Bakarov \(2018\)](#) notes that the individual tasks *do not correlate* among each other. In Figure 2, we find the contrary for some of the tasks (e.g., retrieval-rhyme or retrieval-analogies). Importantly, there is no strong negative correlation between any tasks, suggesting that performance on one task is not a tradeoff with another.

Model	Art.	IPA	Text
Metric Learner	0.78	0.64	0.62
Triplet Margin	0.84	0.84	0.79
Autoencoder	0.50	0.41	0.41
Count-based	-	0.56	0.51

Table 2: Overall performance of models with various input features. Art. = articulatory features.

5.2. Input Features

For all of our models, it is possible to choose the input feature type, which has an impact on the performance, as shown in Table 2. Unsurprisingly, the more phonetic the features are, the better the resulting model is. In the *Metric Learner* and *Triplet Margin* models we are still using supervision from the articulatory distance, and despite that, the input features play a major role.

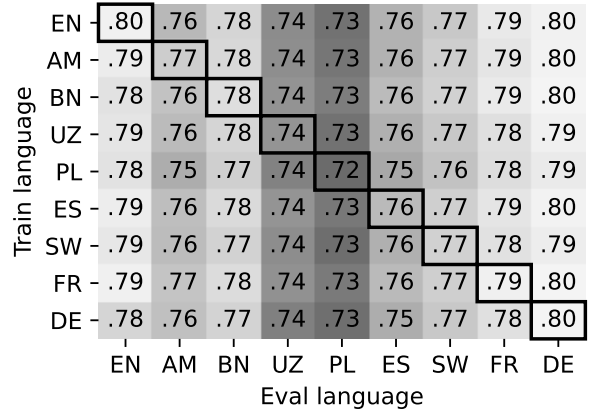


Figure 3: Suite score of Metric Learner with articulatory features trained on one language and evaluated on another one. Diagonal shows models trained and evaluated on the same language.

5.3. Transfer Between Languages

Recall from Section 3.3 that there are multiple feature types that can be used for our phonetic word embedding model: orthographic characters, IPA characters and articulatory feature vectors. It is not surprising that the characters as features provide little transferability when the model is trained on a different language than it is evaluated on. The transfer between languages for a different model type, shown in Figure 3, demonstrates that not all languages are equally challenging (e.g. Polish is more challenging than German). Furthermore, the articulatory features appear to be very useful for generalizing across languages. This echoes the findings of [Li et al. \(2021\)](#), who also break down phones into articulatory features to share information across, possibly unseen, phones.

5.4. Embedding Topology Visualization

The differences between feature types in Table 2 may not appear very large. Closer inspection of the clusters in the embedding space in Figure 4 reveals, that using the articulatory feature vectors or IPA features yields a vector space which resembles one induced by the articulatory distance the most. This is in line with *A* (articulatory distance, Section 3.3.1) being calculated using articulatory features and is used for the model supervision.

5.5. Dimensionality and Train Data Size

So far we used 300-dimensional embeddings. This choice was motivated solely by the comparison to other word embeddings. Now we examine how the choice of dimensionality, keeping all other things equal, affects individual task performance. The results in Figure 5 (top) show that neither too small

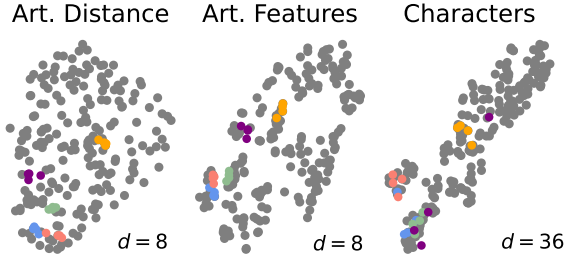


Figure 4: T-SNE projection of articulatory distance and embedding spaces from the metric learning models with articulatory or character features. Each point corresponds to one English word. Differently coloured clusters were selected in the articulatory distance space (left) and highlighted in other spaces. d is the average distance within the clusters normalized with average distance between points (unitless). Articulatory Features (center) result in tighter clusters than Characters (right).

nor too large a dimensionality is useful for the proposed tasks. Furthermore, there is little interaction between the task type and dimensionality. As a result, model ranking based on each task is very similar across dimensions, with Spearman and Pearson correlations of 0.61 and 0.79, respectively.

A natural question is how data-intensive the proposed metric learning method is. For this, we constrained the training data size and show the results in Figure 5 (bottom). Similarly to changing the dimensionality, the individual tasks react to changing the training data size without an effect of the task variable. The Spearman and Pearson correlations are 0.64 and 0.65, respectively.

6. Discussion

6.1. The Field of Phonology

Phonological features, especially articulatory features, play a strong role in phonology since Bloomfield (1993) and the work of Prague School linguists (Trubetskoy, 1939; Jakobson et al., 1951). The widely used articulatory feature set employed by PanPhon originates in the monumental *Sound Pattern of English* (Chomsky and Halle, 1968), which assumes a universal set of discrete phonological features and that all speech sounds in all languages consist of vectors of these features. The similarity between these feature vectors should capture the similarity between sounds. This position is born out in our results. These features encode a wealth of knowledge gained through decades of linguistic research on how the sound systems of languages behave, both synchronically and diachronically. While there is evidence that

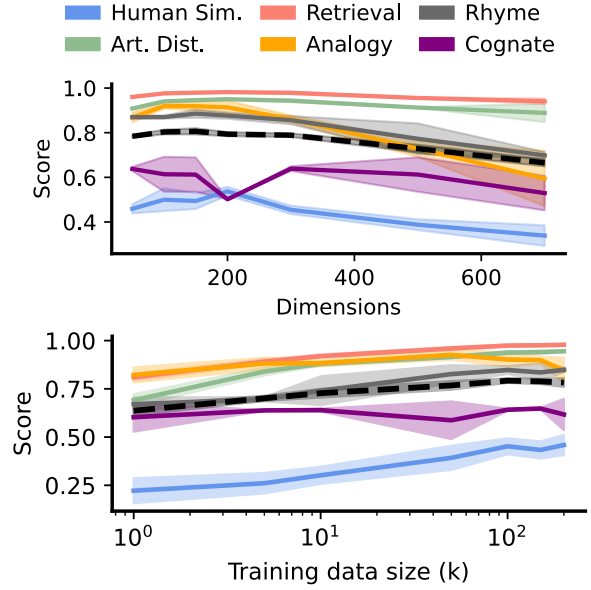


Figure 5: Metric Learner performance with **varying dimensionality** (top) and **varying training data size** (bottom) with articulatory features. Bands show 95% confidence intervals from t-distribution.

phonological features are emergent rather than universal (Mielke, 2008), these results suggest they can nevertheless contribute robustly to computational tasks. Phonetic word embeddings also represent more closely how humans and, in particular, children, interact with language (through sound rather than abstract meaning). Their study may have further applications in the fields of phonetics and phonology.

6.2. Applications

Phonetic word embeddings are more niche than their semantic counterparts but there are many applications shown to benefit from them.

- **Cognate/loanword detection** (Rama, 2016; Nath et al., 2022b,a). Along with semantic similarity, phonetic similarity measured in some latent transformation of articulatory features suggests cognacy or lexical borrowing.
- **Multilingual named entity recognition** (Bharadwaj et al., 2016; Chaudhary et al., 2018). Learning word embeddings from PanPhon features enables cross-lingual transfer for named entity recognition since named entities will likely bear pronunciation similarities across languages.
- **Keyphrase extraction** (Ray Chowdhury et al., 2019; Fahd Saleh Alotaibi and Gupta, 2022). Keyphrase extraction from Tweets for disaster relief can leverage PanPhon features to take advantage of the tendency for orthographic vari-

ants of the same word across different Tweets to share similar pronunciations.

- **Spelling correction** (Tan et al., 2020; Zhang et al., 2021). Imbuing word embeddings with pronunciation similarity helps in correcting typing mistakes by substituting words with their phonetic transcription and similar-sounding words. Another approach is to pretrain a spelling-correction model on phonetic units.
- **Phonotactic learning** (Mirea and Bicknell, 2019; Romero and Salamea, 2021). Phonetic information is a necessary part in deriving phonotactic patterns and vector representations.
- **Multimodal word embeddings** (Zhu et al., 2020, 2021). Phonetic and syntactic information can be incorporated into semantic word embeddings.
- **Spoken language understanding** (Chen et al., 2018, 2021; Fang et al., 2020). Training with phoneme embeddings can reduce errors from confusing phonetically similar words in automatic speech recognition so that such errors do not propagate to downstream natural language understanding tasks.
- **Language identification** (Zhan et al., 2021; Salesky et al., 2021). Phonological features help in distinguishing between languages and their identification.
- **Poetry generation** (Talafha and Rekabdar, 2021; Yi et al., 2018). Word sounds and their pronunciations are critical for poetry and incorporation of this information helps in automatic poetry generation.
- **Linguistic analysis** (Hamilton et al., 2016; Ryskina, Maria and Rabinovich, Ella and Berg-Kirkpatrick, Taylor and Mortensen, David R. and Tsvetkov, Yulia, 2020; Francis et al., 2021). Apart from direct applications, there exist many investigations and analyses on what phonological and phonetic features are encoded by speakers. Phonological word embeddings are one tool by which this can be studied.

6.3. Limitations and Ethics

As hinted in Section 5.1, we evaluate models that use supervision from some of the tasks during training. Specifically, the metric learning models have an advantage on the articulatory distance task. Nevertheless, the models perform well also on other, more unrelated tasks and we also provide models without this supervision. We also do not make any distinction between training and development data. This is for a practical reason because some of the methods we use for comparison are

not open embeddings and need to see all concerned words during training.

Another limitation of our work is that we train on phonemic transcriptions, which cannot capture finer grained phonetic distinctions. Phonemic distinctions may be sufficient for applications such as rhyme detection, but not for tasks such as phone recognition or dialectometry.

We attempted to be inclusive with the language selection and do not foresee any ethical issues.

7. Future Work

After having established the standardized evaluation suite, we wish to pursue the following:

- enlarging the pool of languages,
- including more tasks in the evaluation suite,
- contextual phonetic word embeddings,
- new models for phonetic word embeddings.

8. Bibliographical References

- Felipe Almeida and Geraldo Xexéo. 2019. *Word embeddings: A survey*. *arXiv:1901.09069*.
- Amir Bakarov. 2018. *A survey of word embeddings evaluation methods*. *arXiv:1801.09536*.
- Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. *CogNet: A large-scale cog-nate database*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145.
- Aurélien Bellet, Amaury Habrard, and Marc Sebban. 2015. *Metric learning*. Morgan & Claypool.
- Akash Bharadwaj, David R Mortensen, Chris Dyer, and Jaime G Carbonell. 2016. *Phonologically aware neural model for named entity recognition in low resource transfer settings*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472.
- Leonard Bloomfield. 1993. *Language*. University of Chicago Press.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. *Enriching word vectors with subword information*. *Transactions of the association for computational linguistics*, 5:135–146.
- Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. *From word to sense embeddings: A survey on vector representations of meaning*. *Journal of Artificial Intelligence Research*, 63:743–788.

- Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R Mortensen, and Jaime G Carbonell. 2018. [Adapting word embeddings to new languages with morphological and phonological subword representations](#). *arXiv:1808.09500*.
- Qian Chen, Wen Wang, and Qinglin Zhang. 2021. [Pre-training for spoken language understanding with joint textual and phonetic representation learning](#). In *Interspeech 2021*. ISCA.
- Yi-Chen Chen, Sung-Feng Huang, Chia-Hao Shen, Hung-yi Lee, and Lin-shan Lee. 2018. [Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval](#). In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 941–948.
- Noam Chomsky and Morris Halle. 1968. [The Sound Pattern of English](#). Harper & Row.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.
- Vishal Gupta, Fahd Saleh, Alotaibi, Saurabh Sharma and Savita Gupta. 2022. [Keyphrase extraction using enhanced word and document embedding](#). *IETE Journal of Research*, 0(0):1–13.
- Anjie Fang, Simone Filice, Nut Limsopatham, and Oleg Rokhlenko. 2020. [Using phoneme representations to build predictive models robust to ASR errors](#). In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 699–708. Association for Computing Machinery.
- David Francis, Ella Rabinovich, Farhan Samir, David Mortensen, and Suzanne Stevenson. 2021. [Quantifying cognitive factors in lexical decline](#). *Transactions of the Association for Computational Linguistics*, 9:1529–1545.
- Sahar Ghannay, Yannick Esteve, Nathalie Camelin, and Paul Deléglise. 2016. [Evaluation of acoustic word embeddings](#). In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 62–66.
- William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [Diachronic word embeddings reveal statistical laws of semantic change](#). *arXiv preprint arXiv:1605.09096*.
- Yushi Hu, Shane Settle, and Karen Livescu. 2020. [Multilingual jointly trained acoustic and written word embeddings](#). *arXiv:2006.14007*.
- Roman Jakobson, Gunnar Fant, and Morris Halle. 1951. [Preliminaries to Speech Analysis: The Distinctive Features and their Correlates](#). Language.
- Mahmut Kaya and Hasan Şakir Bilge. 2019. [Deep metric learning: A survey](#). *Symmetry*, 11:1066.
- Quoc Le and Tomas Mikolov. 2014. [Distributed representations of sentences and documents](#). In *International conference on machine learning*, pages 1188–1196. PMLR.
- Xinjian Li, Juncheng Li, Florian Metze, and Alan W Black. 2021. [Hierarchical phone recognition with compositional phonetics](#). In *Interspeech*, pages 2461–2465.
- Jeff Mielke. 2008. [The emergence of distinctive features](#). Oxford University Press.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. [Efficient estimation of word representations in vector space](#). *arXiv:1301.3781*.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. [Linguistic regularities in continuous space word representations](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.
- 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.
- David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. [PanPhon: A resource for mapping IPA segments to articulatory feature vectors](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.
- Abhijnan Nath, Rahul Ghosh, and Nikhil Krishnaswamy. 2022a. [Phonetic, semantic, and articulatory features in Assamese-Bengali cognate detection](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 41–53. Association for Computational Linguistics.

- Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022b. [A generalized method for automated multilingual loanword detection](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013.
- Allison Parrish. 2017. [Poetic sound similarity vectors using phonetic features](#). In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.
- Taraka Rama. 2016. [Siamese convolutional networks for cognate identification](#). In *Proceedings of COLING, the 26th International Conference on Computational Linguistics*, pages 1018–1027.
- Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2019. [Keyphrase extraction from disaster-related tweets](#). In *The world wide web conference*, pages 1555–1566.
- David Romero and Christian Salamea. 2021. [On the use of phonotactic vector representations with fasttext for language identification](#). *Conversational Dialogue Systems for the Next Decade*, pages 339–348.
- Ryskina, Maria and Rabinovich, Ella and Berg-Kirkpatrick, Taylor and Mortensen, David R. and Tsvetkov, Yulia. 2020. [Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods](#). In *Proceedings of the Society for Computation in Linguistics*, volume 3.
- Elizabeth Salesky, Badr M. Abdullah, Sabrina J. Mielke, Elena Klyachko, Oleg Serikov, Edoardo Ponti, Ritesh Kumar, Ryan Cotterell, and Ekaterina Vylomova. 2021. [SIGTYP 2021 shared task: Robust spoken language identification](#).
- Rahul Sharma, Kunal Dhawan, and Balakrishna Pailla. 2021. [Phonetic word embeddings](#). *arXiv:2109.14796*.
- Miikka P. Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. [Sound analogies with phoneme embeddings](#). In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 136–144.
- Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. [One embedder, any task: Instruction-finetuned text embeddings](#). *arXiv:2212.09741*.
- Sameerah Talafha and Banafsheh Rekabdar. 2021. [Poetry generation model via deep learning incorporating extended phonetic and semantic embeddings](#). In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 48–55.
- Min Tan, Dagang Chen, Zesong Li, and Peng Wang. 2020. [Spelling error correction with BERT based on character-phonetic](#). In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, pages 1146–1150.
- Nikolai Trubetskoy. 1939. [Grundzüge der Phonologie](#), volume VII. Travaux du Cercle Linguistique de Prague.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *Advances in neural information processing systems*, 30.
- Paul C Vitz and Brenda Spiegel Winkler. 1973. [Predicting the judged “similarity of sound” of English words](#). *Journal of Verbal Learning and Verbal Behavior*, 12(4):373–388.
- Liu Yang and Rong Jin. 2006. [Distance metric learning: A comprehensive survey](#). *Michigan State University*, 2(2):4.
- Zixiaofan Yang and Julia Hirschberg. 2019. [Linguistically-informed training of acoustic word embeddings for low-resource languages](#). In *Interspeech*, pages 2678–2682.
- Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. 2018. [Chinese poetry generation with a working memory model](#).
- Qingran Zhan, Xiang Xie, Chenguang Hu, and Haobo Cheng. 2021. [A self-supervised model for language identification integrating phonological knowledge](#). *Electronics*, 10(18).
- Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang. 2021. [Correcting chinese spelling errors with phonetic pre-training](#). In *Findings of the Association for Computational Linguistics 2021*, pages 2250–2261.
- Wenhao Zhu, Shuang Liu, and Chaoming Liu. 2021. [Incorporating syntactic and phonetic information into multimodal word embeddings using graph convolutional networks](#). In *ICASSP International Conference on Acoustics, Speech and Signal Processing*, pages 7588–7592. IEEE.

- Wenhao Zhu, Shuang Liu, Chaoming Liu, Xiaoya Yin, and Xiaping Xu. 2020. [Learning multimodal word representations by explicitly embedding syntactic and phonetic information](#). *IEEE Access*, 8:223306–223315.
- Conneau, Alexis and Khandelwal, Kartikay and Goyal, Naman and Chaudhary, Vishrav and Wenzek, Guillaume and Guzmán, Francisco and Grave, Edouard and Ott, Myle and Zettlemoyer, Luke and Stoyanov, Veselin. 2020. [Unsupervised Cross-lingual Representation Learning at Scale](#).
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. [Learning word vectors for 157 languages](#). In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Carnegie Mellon Speech Group. 2014. [The Carnegie Mellon Pronouncing Dictionary 0.7b](#). Carnegie Mellon University.
- Benjamin Heinzerling and Michael Strube. 2018. [BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.
- Wenzek, Guillaume and Lachaux, Marie-Anne and Conneau, Alexis and Chaudhary, Vishrav and Guzmán, Francisco and Joulin, Armand and Grave, Edouard. 2020. [CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data](#). European Language Resources Association.