Are All Good Word Vector Spaces Isomorphic?

Ivan Vulić[1]*, **Sebastian Ruder**[2]*, **Anders Søgaard**[3,4]*
[1] Language Technology Lab, University of Cambridge
[2] DeepMind
[3] Department of Computer Science, University of Copenhagen
[4] Google Research, Berlin
iv250@cam.ac.uk ruder@google.com soegaard@di.ku.dk

Abstract

Existing algorithms for aligning cross-lingual word vector spaces assume that vector spaces are approximately isomorphic. As a result, they perform poorly or fail completely on non-isomorphic spaces. Such non-isomorphism has been hypothesised to result almost exclusively from typological differences between languages. In this work, we ask whether non-isomorphism is also crucially *a sign of degenerate word vector spaces*. We present a series of experiments across diverse languages which show that, besides inherent typological differences, variance in performance across language pairs can largely be attributed to the size of the monolingual resources available, and to the properties and duration of monolingual training (e.g. "under-training").

1 Introduction

Word embeddings have been argued to reflect how language users organise concepts. The extent to which they do so has been evaluated, e.g., using semantic word similarity and association norms (Hill et al., 2015; Gerz et al., 2016), and word analogy benchmarks (Mikolov et al., 2013c). If word embeddings reflect more or less language-independent conceptual organisations, word embeddings in different languages can be expected to be near-isomorphic. Researchers have exploited this to learn linear transformations between such spaces (Mikolov et al., 2013a; Glavaš et al., 2019), which have been used to induce bilingual dictionaries, as well as to facilitate multilingual modeling and cross-lingual transfer (Ruder et al., 2019).

In this paper, we show that near-isomorphism arises only with sufficient amounts of training. This is of practical interest for applications of linear alignment methods for cross-lingual word embeddings. It furthermore provides us with an explanation for reported failures to align word vector spaces in different languages (Søgaard et al., 2018; Artetxe et al., 2018a), which has so far been largely attributed *only* to inherent typological differences.

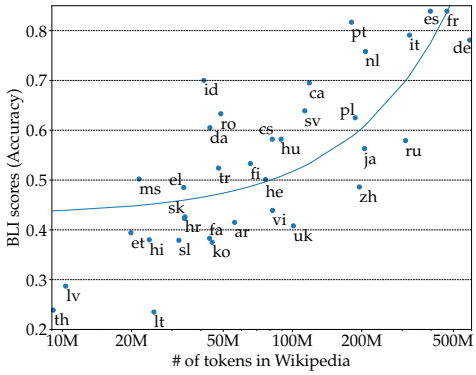


Figure 1: Performance of a state-of-the-art BLI model mapping from English to a target language and the size of the target language Wikipedia are correlated. Linear fit shown as a blue line (log scale).

In fact, the amount of data used to induce the monolingual embeddings is predictive of the quality of the aligned cross-lingual word embeddings, as evaluated on bilingual lexicon induction (BLI). Consider, for motivation, Figure 1; it shows the performance of a state-of-the-art alignment method—RCSLS with iterative normalisation (Zhang et al., 2019)—on mapping English embeddings onto embeddings in other languages, and its correlation ($\rho = 0.72$) with the size of the tokenised target language Polyglot Wikipedia (Al-Rfou et al., 2013).

We investigate to what extent the amount of data available for some languages and corresponding training conditions provide a *sufficient* explanation for the variance in reported results; that is, whether it is the full story or not: The answer is 'almost', that is, its interplay with inherent typological differences does have a crucial impact on the 'alignability' of monolingual vector spaces.

We first discuss current standard methods of

*All authors contributed equally to this work.

quantifying the degree of near-isomorphism between word vector spaces (§2.1). We then outline training settings that may influence isomorphism (§2.2) and present a novel experimental protocol for learning cross-lingual word embeddings that simulates a low-resource environment, and also controls for topical skew and differences in morphological complexity (§3). We focus on two groups of languages: **1)** Spanish, Basque, Galician, and Quechua, and **2)** Bengali, Tamil, and Urdu, as these are arguably spoken in culturally related regions, but have very different morphology. Our experiments, among other findings, indicate that a low-resource version of Spanish is as difficult to align to English as Quechua, challenging the assumption from prior work that the primary issue to resolve in cross-lingual word embedding learning is language dissimilarity (instead of, e.g., procuring additional raw data for embedding training). We also show that by controlling for different factors, we reduce the gap between aligning Spanish and Basque to English from 0.291 to 0.129, and similarly do not observe any substantial performance difference between Spanish and Galician, or Bengali and Tamil.

We also investigate the learning dynamics of monolingual word embeddings and their impact on BLI performance and isomorphism of the resulting word vector spaces (§4), finding training duration, amount of monolingual resources, preprocessing, and self-learning all to have a large impact. The findings are verified across a set of typologically diverse languages, where we pair English with Spanish, Arabic, and Japanese.

We will release our new evaluation dictionaries and subsampled Wikipedias controlling for topical skew and morphological differences to facilitate future research at: `https://github.com/cambridgeltl/iso-study`.

2 Isomorphism of Vector Spaces

Studies analyzing the qualities of monolingual word vector spaces have focused on intrinsic tasks (Baroni et al., 2014), correlations (Tsvetkov et al., 2015), and subspaces (Yaghoobzadeh and Schütze, 2016). In the cross-lingual setting, the most important indicator for performance has been the *degree of isomorphism*, that is, how (topologically) similar the structures of the two vector spaces are.

Mapping-based approaches The prevalent way to learn a cross-lingual embedding space, especially in low-data regimes, is to learn a mapping between a source and a target embedding space (Mikolov et al., 2013a). Such mapping-based approaches assume that the monolingual embedding spaces are isomorphic, i.e., that one can be transformed into the other via a linear transformation (Xing et al., 2015; Artetxe et al., 2018a). Recent unsupervised approaches rely even more strongly on this assumption: They assume that the structures of the embedding spaces are so similar that they can be aligned by minimising the distance between the transformed source language and the target language embedding space (Zhang et al., 2017; Conneau et al., 2018; Xu et al., 2018; Alvarez-Melis and Jaakkola, 2018; Hartmann et al., 2019).

2.1 Quantifying Isomorphism

We employ measures that quantify isomorphism in three distinct ways—based on graphs, metric spaces, and vector similarity.

Eigenvector similarity (Søgaard et al., 2018) Eigenvector similarity (EVS) estimates the degree of isomorphism based on properties of the nearest neighbour graphs of the two embedding spaces. We first length-normalise embeddings in both embedding spaces and compute the nearest neighbour graphs on a subset of the top most frequent N words. We then calculate the Laplacian matrices L_1 and L_2 of each graph. For L_1, we find the smallest k_1 such that the sum of its k_1 largest eigenvalues $\sum_{i=1}^{k_1} \lambda_{1i}$ is at least 90% of the sum of all its eigenvalues. We proceed analogously for k_2 and set $k = \min(k_1, k_2)$. The eigenvector similarity metric Δ is now the sum of the squared differences of the k largest Laplacian eigenvalues: $\Delta = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2$. The lower Δ, the *more* similar are the graphs and the more isomorphic are the embedding spaces.

Gromov-Hausdorff distance (Patra et al., 2019) The Hausdorff distance is a measure of the worst case distance between two metric spaces $\mathcal{X}$ and $\mathcal{Y}$ with a distance function d:

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = \max\{\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} d(x, y), \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} d(x, y)\}$$

Intuitively, it measures the distance between the nearest neighbours that are farthest apart. The Gromov-Hausdorff distance (GH) in turn minimizes this distance over all isometric transforms (orthogonal transforms in our case as we apply mean centering) $\mathcal{X}$ and $\mathcal{Y}$ as follows:

$$\mathcal{GH}(\mathcal{X}, \mathcal{Y}) = \inf_{f,g} \mathcal{H}(f(\mathcal{X}), g(\mathcal{Y}))$$

In practice, $\mathcal{GH}$ is calculated by computing the Bottleneck distance between the metric spaces (Chazal et al., 2009; Patra et al., 2019).

Relational similarity As an alternative, we consider a simpler measure inspired by Zhang et al. (2019). This measure, dubbed RSIM, is based on the intuition that the similarity distributions of translations within each language should be similar. We first take M translation pairs (m_s, m_t) from our bilingual dictionary. We then calculate cosine similarities for each pair of words (m_s, n_s) on the source side where $m_s \neq n_s$ and do the same on the target side. Finally, we compute the Pearson correlation coefficient ρ of the sorted lists of similarity scores. Fully isomorphic embeddings would have a correlation of $\rho = 1.0$, and the correlation decreases with lower degrees of isomorphism.[1]

2.2 Isomorphism and Learning

Non-isomorphic embedding spaces have been attributed largely to typological differences between languages (Søgaard et al., 2018; Patra et al., 2019; Ormazabal et al., 2019). We hypothesise that non-isomorphism is not solely an intrinsic property of dissimilar languages, but also a result of a poorly conditioned training setup. In particular, languages that are regarded as being dissimilar to English, i.e. non-Indo-European languages, are often also low-resource languages where comparatively few samples for learning word embeddings are available.[2] As a result, embeddings trained for low-resource languages may often not match the quality of their high-resource counterparts, and may thus constitute the main challenge when mapping embedding spaces. To investigate this hypothesis, we consider different aspects of poor conditioning as follows.

Corpus size It has become standard to align monolingual word embeddings trained on Wikipedia (Glavaš et al., 2019; Zhang et al., 2019). As can be seen in Figure 1, and also in Table 1, Wikipedias of low-resource languages are more than a magnitude smaller than Wikipedias of high-resource languages.[3] Corpus size has been shown to play a role in the performance of monolingual embeddings (Sahlgren and Lenci, 2016), but it is unclear how it influences their structure and isomorphism.

Training duration As it is generally too expensive to tune hyper-parameters separately for each language, monolingual embeddings are typically trained for the same number of epochs in large-scale studies. As a result, word embeddings of low-resource languages may be "under-trained".

Preprocessing Different forms of preprocessing have been shown to aid in learning a mapping (Artetxe et al., 2018b; Vulić et al., 2019; Zhang et al., 2019). Consequently, they may also influence the isomorphism of the vector spaces.

Topical skew The Wikipedias of low-resource languages may be dominated by few contributors, skewed towards particular topics, or generated automatically.[4] Embeddings trained on different domains are known to be non-isomorphic (Søgaard et al., 2018; Vulić et al., 2019). A topical skew may thus also make embedding spaces harder to align.

3 Simulating Low-resource Settings

As low-resource languages—by definition—have only a limited amount of data available, we cannot easily control for all aspects using only a low-resource language. Instead, we modify the training setup of a high-resource language to simulate a low-resource scenario. For most of our experiments, we use English (EN) as the source language and modify the training setup of Spanish (ES). Additional results where we modify the training setup of English instead are available as the supplemental material; they further corroborate our key findings. We choose this language pair as both are similar, i.e. Indo-European, high-resource, and BLI performance is typically very high. Despite this high performance, unlike English, Spanish is a highly inflected language. In order to inspect if similar patterns also hold across typologically more dissimilar languages, we also conduct simulation experiments with two other target languages with large Wikipedias in lieu of Spanish: Japanese (JA, an agglutinative language) and Arabic (AR, introflexive).

When controlling for *corpus size*, we subsample the target language (i.e.,, Spanish, Japanese, or Arabic) Wikipedia to obtain numbers of tokens comparable to low-resource languages as illustrated in

[1] There are other measures that quantify similarity between word vectors spaces based on network modularity (Fujinuma et al., 2019) and external resources such as sense-aligned corpora (Ammar et al., 2016), but we do not include them for brevity and because they show similar relative trends.

[2] There are obvious exceptions to this, such as Mandarin.

[3] Recent initiatives replace training on Wikipedia with training on larger CommonCrawl data (Grave et al., 2018; Conneau et al., 2020), but the large differences in corpora sizes between high-resource and low-resource languages are not removed.

[4] As one prominent example, a bot has generated most articles in the Swedish, Cebuano, and Waray Wikipedias.

ES Wikipedia sample		
# Sentences	# Tokens	Comparable Wikis
50k	1.3M	Amharic, Yoruba, Khmer
100k	2.7M	Ilocano, Punjabi
200k	5.4M	Burmese, Nepali, Irish
500k	13.4M	Telugu, Tatar, Afrikaans
1M	26.8M	Armenian, Uzbek, Latvian
2M	53.7M	Croatian, Slovak, Malay
5M	134.1M	Finnish, Indonesian
10M	268.3M	Catalan, Ukrainian

Table 1: Spanish Wikipedia samples of different sizes and comparable Wikipedias in other languages.

Table 1. When controlling for *training duration*, we take snapshots of the "under-trained" vector spaces after seeing an exact number of M word tokens (i.e., after performing M *updates*).

To control for *topical skew*, we need to sample similar documents as in low-resource languages. To maximise topical overlap, we choose low-resource languages that are spoken in similar regions as Spanish and whose Wikipedias might thus also focus on similar topics—specifically Basque (EU), Galician (GL), and Quechua (QU). These four languages have very different morphology. Quechua is an agglutinative language, while Spanish, Galician, and Basque are highly inflected. Basque additionally employs case marking and derivation. If non-isomorphism was entirely explained by language dissimilarity, we would expect even low-resource versions of Spanish to have high BLI performance with English. We repeat the same experiment with another set of languages with distinct properties but spoken in similar regions: Bengali, Urdu, and Tamil.

Typological differences however, may still explain part of the difference in performance. For instance, as we cannot simulate Basque by changing the typological features of Spanish[5], we instead make Spanish, Basque, Galician, and Quechuan "morphologically similar": we remove inflections and case marking through lemmatisation. We follow the same process for Bengali, Urdu, and Tamil.

4 Experiments and Analyses

4.1 Experimental Setup

Embedding algorithm Previous work has shown that learning embedding spaces with different hyper-parameters leads to non-isomorphic spaces (Søgaard et al., 2018; Hartmann et al., 2018). To control for this aspect, we train monolingual embeddings with fastText in the standard setup (skip-gram, character n-grams of sizes 3 to 6, a learning rate of 0.025, 15 negative samples, a window size of 5) (Bojanowski et al., 2017).[6] Unless specified otherwise, we train for 15 epochs.

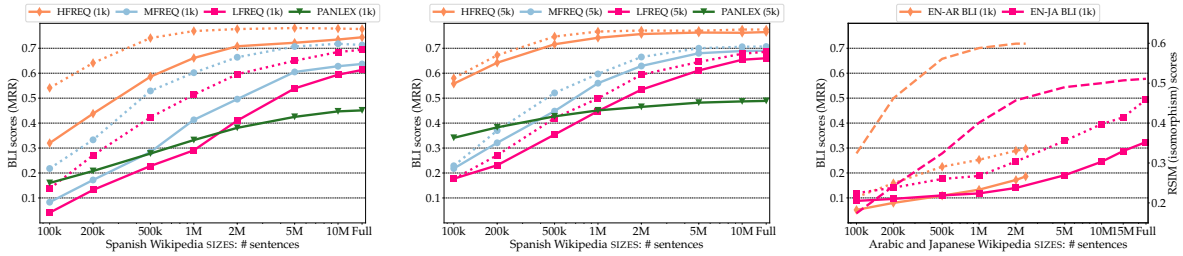
Mapping algorithm We use the supervised variant of VecMap (Artetxe et al., 2018a) for our experiments, which is a robust and competitive choice according to the recent empirical comparative studies (Glavaš et al., 2019; Vulić et al., 2019; Hartmann et al., 2019). VecMap learns an orthogonal transformation based on a seed translation dictionary with additional preprocessing and postprocessing steps, and it can additionally enable *self-learning* in multiple iterations. For further details we refer the reader to the original work (Artetxe et al., 2018a).

Evaluation We measure isomorphism between monolingual spaces using the previously described intrinsic measures: eigenvector similarity (EVS), Gromov-Hausdorff distance (GH), and relational similarity (RSIM). In addition, we evaluate on bilingual lexicon induction (BLI), a standard task for evaluating cross-lingual word representations. Given a list of N_s source words, the task is to find the corresponding translation in the target language as a nearest neighbour in the cross-lingual embedding space. The list of retrieved translations is then compared against a gold standard dictionary. Following prior work (Glavaš et al., 2019), we employ mean reciprocal rank (MRR) as evaluation measure, and use cosine similarity as similarity measure.

Training and test dictionaries Standard BLI test dictionaries over-emphasise frequent words (Czarnowska et al., 2019; Kementchedjhieva et al., 2019) whose neighbourhoods may be more isomorphic (Nakashole, 2018). To account for this, we create new evaluation dictionaries for English–Spanish that consist of words in different frequency bins: we sample EN words for 300 translation pairs respectively from (i) the top 5k words of the full English Wikipedia (HFREQ); (ii) the interval [10k, 20k] (MFREQ); (iii) the interval [20k, 50k] (LFREQ). The entire dataset (ALL-FREQ; 900 pairs) consists of (i) + (ii) + (iii). We exclude named entities as they are over-represented in many test sets (Kementchedjhieva et al., 2019) and in-

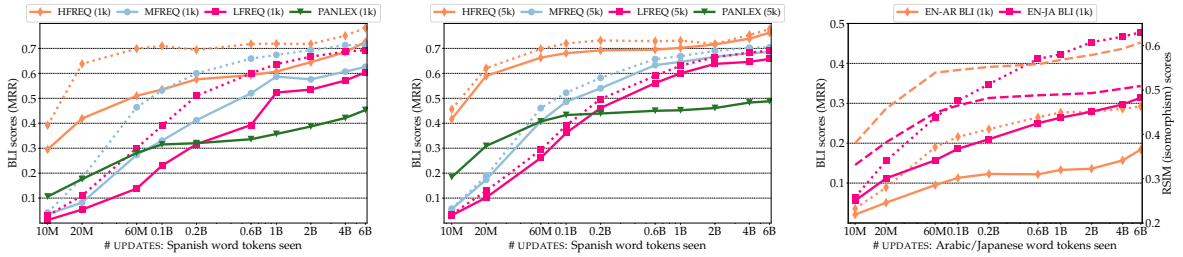
[5] Ravfogel et al. (2019) generate synthetic versions of English that differ from English in a single typological parameter. This process requires a treebank and is infeasible for all typological parameters of a language.

[6] We ran experiments and observed similar results with `word2vec` algorithms (Mikolov et al., 2013b), GloVe (Pennington et al., 2014) and fastText CBOW (Grave et al., 2018).



(a) EN → ES (1k) (b) EN → ES (5k) (c) EN → AR/JA (1k)

Figure 2: Impact of **dataset size** on BLI when aligning ES, AR, and JA vector spaces fully trained on corpora of different sizes (obtained through sampling from the full corpus) to an EN space fully trained on complete data. We report scores without self-learning (solid lines) and with self-learning (dotted lines; same colour) with seed dictionary sizes of **(a)** 1k and **(b)** 5k on our EN–ES BLI evaluation sets, while the corresponding isomorphism scores are provided in Figure 6 for clarity. **(c)** We again report scores without and with self-learning on EN–AR/JA BLI evaluation sets from the MUSE benchmark with 1k seed translation pairs. The results with 5k seed pairs for EN–AR/JA are available in the supplemental material. Dashed lines without any marks show isomorphism scores (computed by RSIM; higher is better) computed across different AR and JA snapshots.



(a) EN → ES (1k) (b) EN → ES (5k) (c) EN → AR/JA (1k)

Figure 3: Impact of **training duration** on BLI when aligning a partially trained Spanish (ES), Arabic (AR), and Japanese (JA) vector space, where snapshots are taken after seeing M word tokens in training, to the fully trained EN space. We report scores without self-learning (solid lines) and with self-learning (dotted lines; same colour) with seed dictionary sizes of **(a)** 1k and **(b)** 5k on our EN–ES BLI evaluation sets. For clarity, the corresponding isomorphism scores (and impact of training duration on isomorphism of vector spaces) over the same training snapshots for Spanish are shown in Figure 5. **(c)** We again report scores without and with self-learning on EN–AR/JA BLI evaluation sets from the MUSE benchmark with 1k seed translation pairs. The results with 5k seed pairs for EN–AR/JA are available in the supplemental material. Dashed lines without any marks show isomorphism scores (computed by RSIM; higher is better) computed across different AR and JA snapshots.

clude nouns, verbs, adjectives, and adverbs in all three sets. All 900 words have been carefully manually translated, and translations double-checked by a native Spanish speaker. There are no duplicates. We also report BLI results on the PanLex test lexicons (Vulić et al., 2019).

For English–Spanish, we create training dictionaries of sizes 1k and 5k based on PanLex (Kamholz et al., 2014) following the same procedure as Vulić et al. (2019). We exclude all words from ALL-FREQ from the training set. For EN–JA/AR BLI experiments, we rely on the standard training and test dictionaries from the MUSE benchmark (Conneau et al., 2018). Isomorphism scores with RSIM for EN–JA/AR are computed on a fixed random sample of 1k one-to-one translations from the respective MUSE training dictionary.[7] For learning monolingual embeddings, we use tokenised and sentence-split Polyglot Wikipedias (Al-Rfou et al., 2013). In §4.6, we process Wiki dumps, using Moses for tokenisation and sentence splitting. For lemmatisation of Spanish, Basque, Galician, Tamil, and Urdu we employ the UDPipe models (Straka and Straková, 2017). For Quechua and Bengali, we utilise the unsupervised Morfessor model provided by Polyglot NLP.

[7] Note that the *absolute* isomorphism scores between different pairs of languages are not directly comparable. The focus of the experiments is to follow the patterns of isomorphism change for each language pair separately.

4.2 Impact of Corpus Size

To evaluate the impact of the corpus size on vector space isomorphism and BLI performance, we shuffle the target language (i.e., Spanish, Arabic, Japanese) Wikipedias and take N sentences where $N \in \{$10k, 20k, 500k, 100k, 500k, 1M, 2M, 10M, 15M$\}$ corresponding to a range of low-resource languages (see Table 1). Each smaller dataset is a subset of the larger one. We learn target language embeddings for each sample, map them to the English embeddings using dictionaries of sizes 1k and 5k and supervised VecMap with and without self-learning, and report their BLI performance and isomorphism scores in Figure 2. Both isomorphism and BLI scores improve with larger training resources.[8] Performance is higher with a larger training dictionary and self-learning but shows a similar convergence behaviour irrespective of these choices. What is more, despite different absolute scores, we observe a similar behaviour for all three language pairs, demonstrating that our intuition holds across typologically diverse languages.

In English–Spanish experiments performance on frequent words converges relatively early, between 1-2M sentences, while performance on medium and low-frequency words continues to increase with more training data and only plateaus around 10M sentences. Self-learning improves BLI scores, especially in low-data regimes. Note that isomorphism scores increase even as BLI scores saturate, which we discuss in more detail in §5.

4.3 Impact of Training Duration

To analyse the effect of under-training, we align English embeddings with target language embeddings that were trained for a certain number of iterations/updates and compute their BLI scores. The results for the three language pairs are provided in Figure 3. As monolingual vectors are trained for longer periods, BLI and isomorphism scores improve monotonously, and this holds for all three language pairs. Even after training for a large number of updates, BLI and isomorphism scores do not show clear signs of convergence. Self-learning again seems beneficial for BLI, especially at earlier, "under-training" stages.

[8] We observe similar trends when estimating isomorphism on the 5k and 10k most frequent words in both languages.

4.4 Impact on Monolingual Mapping

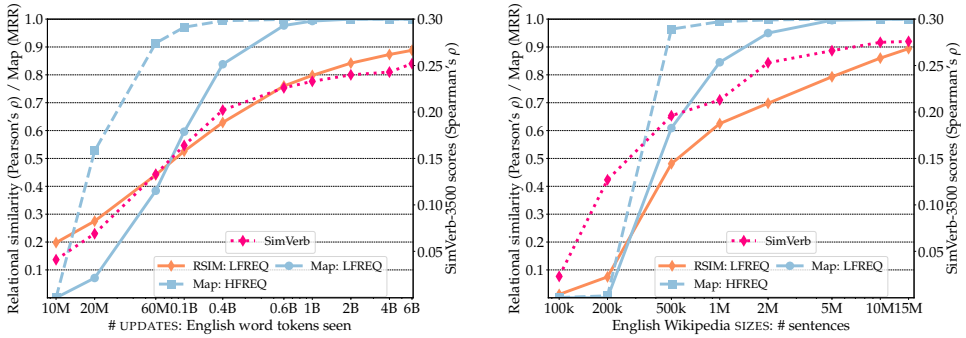
As a control experiment, we repeat the two previous experiments controlling for corpus size and training duration when mapping an English embedding space to another EN embedding space. Previous work (Hartmann et al., 2018) has shown that EN embeddings learned with the same algorithm achieve a perfect monolingual "BLI" score of 1.0 (mapping EN words to the same EN word). If typological differences were the only factor affecting the structure of embedding spaces, we would thus expect to achieve a perfect score also for shorter training and smaller corpus sizes. For comparison, we also provide scores on a standard monolingual word similarity benchmark, SimVerb-3500 (Gerz et al., 2016). We show results in Figure 4. We observe that BLI scores only reach 1.0 after 0.4B and 0.6B updates for frequent and infrequent words or with corpus sizes of 1M and 5M sentences respectively, which is more than the size of most low-resource language Wikipedias (Table 1). This clearly shows that even aligning EN to EN is challenging in a low-resource setting due to different vector space structures, and we cannot attribute performance differences to typological differences in this case.

4.5 Impact of Preprocessing

We next evaluate the impact of different forms of preprocessing. Specifically, we consider: **1)** No preprocessing (unnormalised vectors); **2)** Length normalisation (L2) only (required for orthogonal Procrustes); **3)** L2, mean centering (MC), followed by L2; used by VecMap (Artetxe et al., 2018a); and **4)** Iterative normalisation (Zhang et al., 2019).

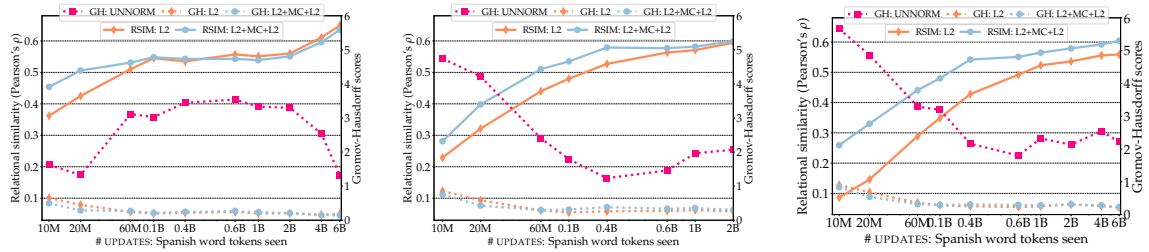
Iterative normalisation consists of multiple steps of L2 +MC +L2.[9] We have found it to achieve performance nearly identical to L2 +MC +L2, so we do not report it separately. We show results for the remaining methods in Figure 5 and Figure 6. For GH, using no preprocessing leads to much less isomorphic spaces, particularly for infrequent words during very early training. For RSIM with cosine similarity, L2 is equivalent to no normalisation as cosine applies length normalisation. L2 +MC +L2 leads to slightly better isomorphism scores overall compared to L2 alone, though it has a slightly negative impact on Gromov-Hausdorff scores over longer training duration. Most importantly, the results demonstrate that such preprocessing steps do have a profound impact on isomorphism between

[9] github.com/zhangmozhi/iternorm



(a) Updates: EN (b) Different Wiki sizes: EN

Figure 4: Monolingual "control" experiments when aligning **(a)** a partially trained EN vector space (after M updates, that is, seen word tokens) to a fully trained vector space, and **(b)** an EN vector space fully trained on Wikipedia of different sizes (number of sentences). We show RSIM scores, mapping performance (i.e., monolingual "BLI") on HFREQ and LFREQ EN words, and monolingual word similarity scores on SimVerb-3500.



(a) HFREQ subset (ES) (b) MFREQ subset (ES) (c) LFREQ subset (ES)

Figure 5: Impact of different monolingual vector space preprocessing strategies on isomorphism scores when aligning a partially trained ES vector space, where snapshots are taken after seeing M word tokens in training, to a fully trained EN vector space. We report RSIM (solid; higher is better, i.e., more isomorphic) and GH distance (dashed; lower is better) on **(a)** HFREQ, **(b)** MFREQ, and **(c)** LFREQ test sets.

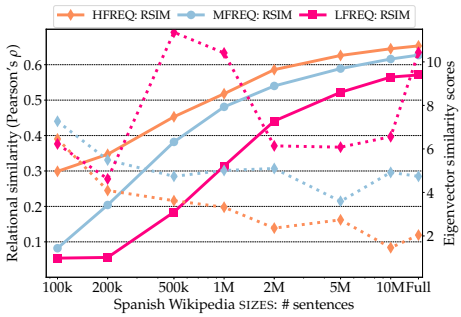


Figure 6: Impact of **dataset size** on vector space isomorphism when aligning an ES vector space fully trained on corpora of different sizes to an EN space fully trained on complete data. RSIM (solid lines; higher is better, i.e. more isomorphic) and Eigenvector similarity (dotted lines; lower is better) scores are reported. See Figures 2a-2b for the corresponding EN–ES BLI scores.

monolingual vector spaces.

4.6 Impact of Topical Skew and Morphology

To control for topical skew, we sample the Spanish Wikipedia so that its topical distribution is as close as possible to that of low-resource languages spoken in similar regions—Basque, Galician, and Quechua. To this end, for each language pair, we first obtain document-level alignments using the Wiki API[10]. We only consider documents that occur in both languages. We then sample sentences from the ES Wikipedia so that the number of tokens per document and the number of tokens overall is similar to the document-aligned sample of the low-resource Wikipedia. This results in topic-adjusted Wikipedias consisting of 14.3M tokens for ES and EU, 26.1M tokens for ES and GL, and 409k tokens for ES and QU. We additionally control for morphology by lemmatising the Wikipedia samples. For Spanish paired with each other language, we use training dictionaries that are similar in size and distribution. We learn monolingual embeddings on each subsampled Wikipedia corpus and align the resulting embeddings with English. We follow the same principle and sample the Bengali Wikipedia in the same way to make its topical distribution aligned with the samples of the Urdu and Tamil Wikipedias: this results in topic-adjusted

[10] https://www.wikidata.org/w/api.php

Wikipedias consisting of 3.8M for Bengali–Urdu, and 8.1M for Bengali–Tamil.

The results are provided in Table 2. We observe that inequality in training resources accounts for a large part of the performance gap. Controlling for topical skew and morphology reduces the gap further and results in nearly identical performance for Spanish compared to Quechua and Galician, respectively. For Galician, lemmatisation slightly widens the gap, likely due to a weaker lemmatiser. For Basque, the remaining gap may be explained by the remaining typological differences between the two languages.[11] We also observe similar patterns in experiments with BN, UR, and TA in Table 2: training with comparable samples with additional morphological processing reduces the observed gap in performance between EN–BN and EN–TA, as well as between EN–BN and EN–UR. This again hints that other factors besides inherent language dissimilarity are at play and contribute to reduced isomorphism between embedding spaces.

5 Further Discussion

Does isomorphism increase beyond convergence? In our experiments, we have measured how training monolingual word embeddings improves their isomorphism with word embeddings in other languages. In doing so, we made an interesting observation: *Isomorphism increases even beyond the point in the training process where validation (BLI) scores and training losses plateau.* To see this, consider Figure 7. One possible explanation for this may be that isomorphism increases as learning dynamics drive us into high-entropy solutions. Recent studies of learning dynamics in deep neural nets observe that flatter optima generalise better than sharp optima (Zhang et al., 2018): intuitively, it is because sharp minima correspond to more complex, likely over-fitted, models. Zhang et al. (2018) show that analogous to the energy-entropy competition in statistical physics, wide but shallow minima can be optimal if the system is undersampled. SGD is assumed to generalise well because its inherent anisotropic noise biases it towards higher entropy minima. We hypothesise a similar explanation of our observations in terms of energy-entropy competition. Once loss is minimised, the random oscillations due to SGD noise lead the weights toward a high-entropy solution. We hypothesise monolingual high-entropy minima are more likely to be isomorphic. A related possible explanation is that the increased isomorphism results from model compression. This is analogous to the idea of two-phase learning (Shwartz-Ziv and Tishby, 2017), whereby the initial fast convergence of SGD is related to sufficiency of the representation, while the later asymptotic phase is related to compression of the activations.

Do vocabularies align? If languages reflect the world, they should convey semantic knowledge in a similar way, and it is therefore reasonable to assume that with enough data, induced word embeddings should be isomorphic. On the other hand, if languages impose structure on our conceptualszation of the world, non-isomorphic word embeddings could easily arise. Studies that engage with speakers of different languages in the real world (Majid, 2010) are naturally limited in scope. Large-scale studies, on the other hand, have generally relied on distributional methods (Thompson et al., 2018), leading to a chicken-and-egg scenario. Vossen (2002) discuss mismatches between WordNets across languages, including examples of hypernyms without translation equivalents, e.g., *dedo* in Spanish (*fingers* and *toes* in English). Such examples break isomorphism between languages, but are relatively rare. Another approach to the question of vocabulary alignment is to study lexical organisation in bilinguals and how it differs from that of monolingual speakers (Pavlenko, 2009). While this paper obviously does not provide hard evidence for or against Sapir-Whorf-like hypotheses, our results suggest that the variation observed in BLI performance cannot trivially be attributed only to linguistic differences.

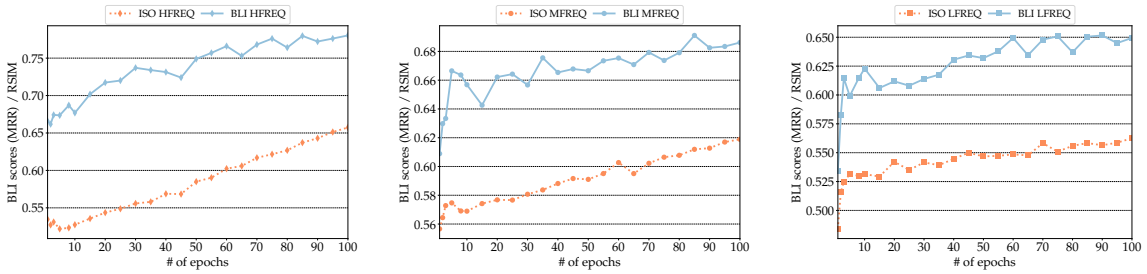
Impact of training conditions We find that degenerate conditions of monolingual training account for a significant part of the performance gap in bilingual lexicon induction with low-resource languages. This is in contrast to previous studies (Søgaard et al., 2018; Patra et al., 2019; Ormazabal et al., 2019) that generally attributed poor performance across languages predominantly to typological differences. While labelled data is generally assumed to be scarce, current methods tacitly assume that sufficient unlabelled data is available to learn good representations. Our results highlight that this is generally not the case for low-resource languages. We thus suggest to focus on methods

[11] The drop in performance for EN–ES in the setup with full Wikipedias when we analyze Spanish and Quechua is attributed to a smaller and lower-quality training lexicon (to make it comparable to the corresponding EN–QU lexicon).

	Basque (EU)			Quechua (QU)			Galician (GL)		
	EN-ES	EN-EU	ES/EU gap	EN-ES	EN-QU	ES/QU gap	EN-ES	EN-GL	ES/GL gap
Full wiki	0.757	0.466	0.291	0.572	0.066	0.506	0.757	0.689	0.068
Random sample	0.662	0.411	0.251	0.037	0.054	-0.017	0.680	0.663	0.017
Comparable sample	0.663	0.420	0.243	0.081	0.060	0.021	0.669	0.671	-0.002
Comp. sample + lemma	0.533	0.404	0.129	0.052	0.041	0.011	0.619	0.596	0.023

	Tamil (TA)			Urdu (UR)		
	EN-BN	EN-TA	BN/TA gap	EN-BN	EN-UR	BN/UR gap
Full wiki	0.253	0.152	0.101	0.118	0.132	-0.014
Random sample	0.193	0.124	0.069	0.076	0.093	-0.017
Comparable sample	0.196	0.131	0.065	0.108	0.112	-0.004
Comp. sample + lemma	0.152	0.121	0.031	0.072	0.070	0.002

Table 2: BLI scores (MRR) when mapping from a fully trained EN embedding space to one trained on full Wikipedia corpora, random samples, and topic-adjusted comparable samples of the same size with and without lemmatisation for Spanish (ES) and Basque (EU), Quechua (QU), and Galician (GL), respectively (**Top** table); Bengali (BN) and Tamil (TA), and BN and Urdu (UR) (**Bottom** table).



(a) HFREQ subset (ES) (b) MFREQ subset (ES) (c) LFREQ subset (ES)

Figure 7: Monolingual learning dynamics and isomorphism (RSIM): We align a partially trained ES vector space, after seeing M word tokens, with a fully trained EN vector space, and evaluate on (a) HFREQ, (b) MFREQ, and (c) LFREQ test sets. While BLI performance plateaus, the isomorphism score (computed with RSIM) does not.

that can transfer even with few unlabelled samples or to procur data from other sources, e.g. typologically similar languages with more resources.

Importance of word frequency We also find that word frequency has a strong effect on isomorphism and BLI performance. Word frequency is an understudied property as standard BLI datasets focus only on frequent words. Similar to previous work (Czarnowska et al., 2019), we find that BLI scores are generally lower on less frequent words. In addition, we demonstrate that graphs corresponding to less frequent words are less isomorphic and that they take longer to converge. This demonstrates the importance of studying representations across the entire frequency spectrum.

Self-learning and normalisation Finally, we observe self-learning and normalisation to have a significant impact on isomorphism. Self-learning has been found useful with small training dictionaries (Artetxe et al., 2017; Søgaard et al., 2018; Hartmann et al., 2019). Our experiments demonstrate that it is beneficial even with larger dictionaries in low-resource setups and that it leads to gains in all settings. Normalisation is useful particularly in low-resource setups and for infrequent words.

6 Conclusion

We have provided a series of analyses that demonstrate together that non-isomorphism is not—as previously assumed—primarily and solely a result of typological differences between languages, but due in large part to degenerate vector spaces and discrepancies between monolingual training regimes and data availability. Through controlled experiments in simulated low-resource scenarios, also involving languages with different morphology that are spoken in culturally related regions, we found that such vector spaces mainly arise from poor conditioning during training. The study suggests that besides improving our alignment algorithms for distant languages (Vulić et al., 2019), we should also focus on improving monolingual word vector spaces, and monolingual training conditions.

Acknowledgments

The work of IV is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). AS is supported by a Google Focused Research Award.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of CoNLL 2013*, pages 183–192.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of EMNLP 2018*, pages 1881–1890.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL 2017*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL 2018*, pages 789–798.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of AAAI 2018*, pages 5012–5019.

Marco Baroni, Georgiana Dinu, and German Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014*, pages 238–247.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. 2009. Gromov-Hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*, volume 28, pages 1393–1403.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR 2018*.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! A comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of EMNLP 2019*, pages 974–983.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul. 2019. A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. In *Proceedings of ACL 2019*, pages 4952–4962.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP 2016*, pages 2173–2182.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL 2019*, pages 710–721.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC 2018*, pages 3483–3487.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2018. Why is unsupervised alignment of English embeddings from different algorithms so hard? In *Proceedings of EMNLP 2018*, pages 582–586.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2019. Comparing unsupervised word translation methods step by step. In *Proceedings of NeurIPS 2019*, pages 6031–6041.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of LREC 2014*, pages 3145–3150.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of EMNLP 2019*, pages 3336–3341.

A Majid. 2010. Comparing lexicons cross-linguistically. In *Oxford Handbook of the Word*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS 2013*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT 2013*, pages 746–751.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of EMNLP 2018*, pages 512–522.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of ACL 2019*, pages 4990–4995.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of ACL 2019*, pages 184–193.

Aneta Pavlenko, editor. 2009. *The Bilingual Mental Lexicon*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of NAACL-HLT 2019*, pages 3532–3542.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of EMNLP 2016*, pages 975–980.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL 2018*, pages 778–788.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Bill Thompson, Sean Roberts, and Gary Lupyan. 2018. Quantifying semantic alignment across languages. *MPI Tech Report*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP 2015*, pages 2049–2054.

Piet Vossen. 2002. WordNet, EuroWordNet and Global WordNet . *Revue française de linguistique appliquée*, VII.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP 2019*, pages 4407–4418.

Chao Xing, Chao Liu, Dong Wang, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL-HLT 2015*, pages 1005–1010.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of EMNLP 2018*, pages 2465–2474.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Intrinsic subspace evaluation of word embedding representations. In *Proceedings of ACL 2016*, pages 236–246.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of EMNLP 2017*, pages 1934–1945.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shojo? Cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of ACL 2019*, pages 3180–3189.

Yao Zhang, Andrew Saxe, Madhu Advani, and Alpha Lee. 2018. Energy–entropy competition and the effectiveness of stochastic gradient descent in machine learning. *Molecular Physics*, 116.

A Supplemental Material

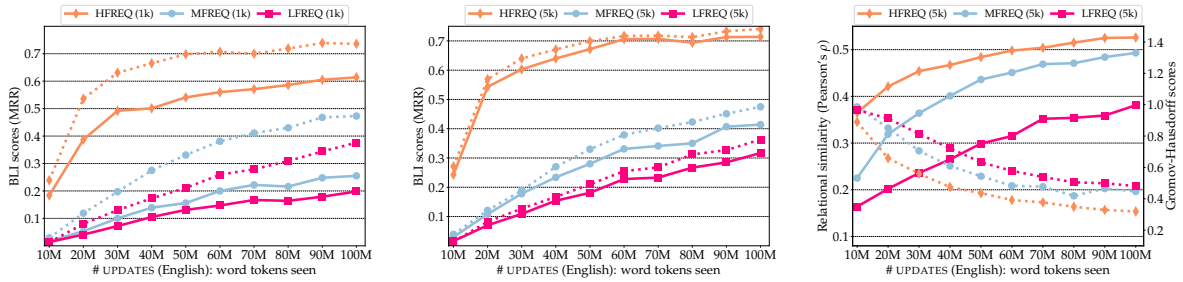
Additional experiments that further support the main claims of the paper have been relegated to the supplemental material for clarity and compactness of presentation. We provide the following additional information:

- **Table 3**. It provides "reference" BLI scores and scores stemming from isomorphism measures when we align fully trained EN and ES spaces, that is, when we rely on standard 15 epochs of fastText training on respective Wikipedias.
- **Figure 8** and **Figure 9** show BLI and isomorphism scores at very early stages of training, both for EN and ES. In other words, one vector space is fully trained, while we take early-training snapshots (after seeing only 10M, 20M, ..., 100M word tokens in training) of the other vector space. The results again stress the importance of training corpus size as well as training duration—early training stages clearly lead to suboptimal performance and non-isomorphic spaces. However, such shorter training durations (in terms of the number of tokens) are often encountered "in the wild" with low-resource languages.
- **Figure 10b** and **Figure 10a** show the results with 5k seed translation pairs in different training regimes for EN-AR and EN-JA experiments. The results with 1k seed translation pairs are provided in the main paper.
- **Figure 11a** and **Figure 11b** demonstrate the impact of vector space preprocessing (only L2-normalization versus L2 + mean centering + L2) on the RSIM isomorphism scores in different training regimes for EN-AR and EN-JA experiments.
- **Table 4.** It provides additional isomorphism scores, not reported in the paper, again stressing the importance of monolingual vector space preprocessing before learning any cross-lingual mapping.

(The actual tables and figures start on the next page.)

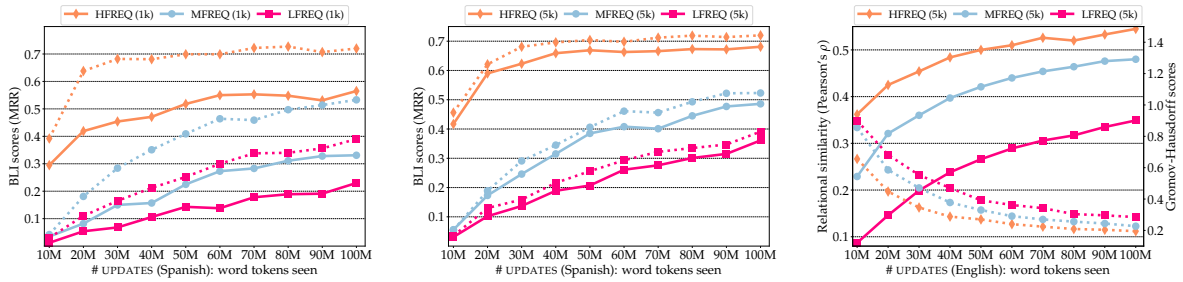
Full EN Vector Space – Full ES Vector Space					
	HFREQ	MFREQ	LFREQ	PANLEX	MUSE
BLI: Supervised (1k)	0.733	0.631	0.621	0.448	0.489
BLI: Supervised+SL (1k)	0.774	0.711	0.695	0.492	0.536
BLI: Supervised (5k)	0.759	0.685	0.658	0.490	0.533
BLI: Supervised+SL (5k)	0.778	0.704	0.691	0.491	0.538
RSIM	0.652	0.633	0.559	–	–
Gromov-Hausdorff	0.274	0.208	0.205	–	–
Eigenvector Similarity	4.86	6.32	10.95	–	–

Table 3: Reference BLI (MRR reported) and isomorphism scores (all three measures discussed in the main paper are reported, computed on L2-normalised vectors) in a setting where we *fully* train both English and Spanish monolingual vector spaces (i.e., training lasts for 15 epochs for both languages) on the *full* data, without taking snapshots at earlier stages, and without data reduction simulation experiments.



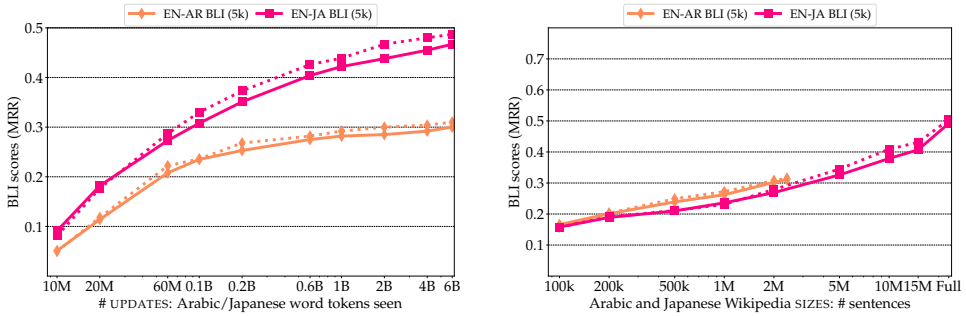
(a) Early updates: EN (1k) (b) Early updates: EN (5k) (c) Early updates: EN (isomorphism)

Figure 8: Impact of training duration on BLI and isomorphism, with a focus on the *early training stages*. BLI scores (a+b) and isomorphism (c) measures of aligning a partially trained EN vector space, where snapshots are taken after seeing N word tokens in training, to a fully trained ES vector space with a seed dictionary of 1k words (a) and 5k words (b) on the three evaluation sets representing different frequency bins. (c) shows how embedding spaces become more isomorphic over the course of training as measured by second-order similarity (on different frequency bins; solid lines, higher is better) and by Gromov-Hausdorff distance (dotted lines of the same colour and with the same symbols; lower is better).



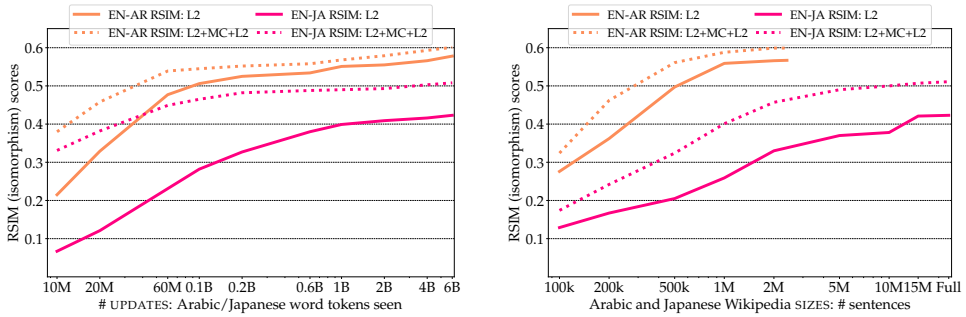
(a) Early updates: EN (1k) (b) Early updates: EN (5k) (c) Early updates: EN (isomorphism)

Figure 9: Impact of training duration on BLI and isomorphism, with a focus on the *early training stages*. BLI scores (a+b) and isomorphism (c) measures of aligning a partially trained ES vector space, where snapshots are taken after seeing N word tokens in training, to a fully trained EN vector space with a seed dictionary of 1k words (a) and 5k words (b) on the three evaluation sets representing different frequency bins. (c) shows how embedding spaces become more isomorphic over the course of training as measured by second-order similarity (on different frequency bins; solid lines, higher is better) and by Gromov-Hausdorff distance (dotted lines of the same colour and with the same symbols; lower is better).



(a) Updates: AR and JA (b) Different Wiki sizes: AR and JA

Figure 10: EN–AR/JA BLI scores on the MUSE BLI benchmark relying on 5k seed pairs for learning the alignment. **(a)** Results with partially trained AR and JA vector spaces where snapshots are taken after M updates (i.e., impact of training duration); **b)** Results with AR and JA vector spaces induced from data samples of different sizes (i.e., impact of dataset size). See the main paper (Figure 2c and Figure 3c) for BLI scores with 5k seed translation pairs.



(a) Updates: AR and JA (b) Different Wiki sizes: AR and JA

Figure 11: The impact of **(a)** training duration and **(b)** dataset size on EN–AR/JA isomorphism scores, also showing the impact of vector space preprocessing steps. We report the RSIM measure (higher is better, i.e., more isomorphic).

	EN (full) - ES (snapshot)				EN (snapshot) - ES (full)			
	EVS		GH		EVS		GH	
# Updates	UNNORM	L2+MC+L2	UNNORM	L2+MC+L2	UNNORM	L2+MC+L2	UNNORM	L2+MC+L2
100M	89.1 [42.7]	6.46 [5.88]	3.03 [3.19]	0.22 [0.30]	43.4 [20.9]	6.74 [26.3]	1.91 [3.66]	0.23 [0.33]
200M	162 [24.5]	3.67 [2.86]	3.45 [2.16]	0.25 [0.31]	38.3 [50.0]	9.15 [18.0]	2.51 [3.39]	0.27 [0.38]
600M	235 [26.2]	7.47 [3.13]	3.54 [1.80]	0.27 [0.29]	26.1 [20.0]	4.51 [15.0]	3.50 [2.21]	0.34 [0.37]
1B	125 [21.3]	4.69 [5.90]	3.31 [2.32]	0.24 [0.28]	45.2 [32.4]	12.5 [13.7]	3.35 [1.49]	0.27 [0.30]
2B	252 [23.4]	5.88 [6.44]	3.03 [2.13]	0.21 [0.31]	25.4 [9.43]	16.5 [10.1]	3.71 [2.03]	0.25 [0.29]
4B	360 [21.0]	8.78 [7.28]	2.54 [2.55]	0.15 [0.29]	141 [22.5]	5.33 [7.99]	3.04 [1.37]	0.18 [0.31]
6B	411 [16.1]	6.96 [8.22]	1.32 [2.22]	0.15 [0.23]	191 [16.7]	8.98 [11.4]	3.22 [0.99]	0.15 [0.37]

Table 4: Eigen Vector Similarity (EVS) and Gromov-Hausdorff distance (GH) distance scores with two different monolingual vector space preprocessing strategies: (a) no normalisation at all (UNNORM); (c) L2-normalisation followed by mean centering (MC) and another L2-normalisation step, done as standard preprocessing in the VecMap framework (Artetxe et al., 2018a) (L2+MC+L2). We show the scores in relation to training duration (provided in the number of updates, i.e., seen word tokens), taking snapshots of the English or the Spanish vector space, and aligning it to a fully trained space on the other side. We show scores on HFREQ and [LFREQ] sets; lower is better (i.e., "more isomorphic").