The Secret is in the Spectra: Predicting Cross-lingual Task Performance with Spectral Similarity Measures

Haim Dubossarsky[1] **Ivan Vulić**[1] **Roi Reichart**[2] **Anna Korhonen**[1]
[1] Language Technology Lab, University of Cambridge
[2] Faculty of Industrial Engineering and Management, Technion, IIT
{hd423, iv250, alk23}@cam.ac.uk roiri@ie.technion.ac.il

Abstract

Performance in cross-lingual NLP tasks is impacted by the (dis)similarity of languages at hand: e.g., previous work has suggested there is a connection between the expected success of bilingual lexicon induction (BLI) and the assumption of (approximate) isomorphism between monolingual embedding spaces. In this work we present a large-scale study focused on the correlations between monolingual embedding space similarity and task performance, covering thousands of language pairs and four different tasks: BLI, parsing, POS tagging and MT. We hypothesize that statistics of the spectrum of each monolingual embedding space indicate how well they can be aligned. We then introduce several isomorphism measures between two embedding spaces, based on the relevant statistics of their individual spectra. We empirically show that **1)** language similarity scores derived from such spectral isomorphism measures are strongly associated with performance observed in different cross-lingual tasks, and **2)** our spectral-based measures consistently outperform previous standard isomorphism measures, while being computationally more tractable and easier to interpret. Finally, our measures capture complementary information to typologically driven language distance measures, and the combination of measures from the two families yields even higher task performance correlations.

1 Introduction

The effectiveness of joint multilingual modeling and cross-lingual transfer in cross-lingual NLP is critically impacted by the actual languages in consideration (Bender, 2011; Ponti et al., 2019). *Characterizing, measuring, and understanding this cross-language variation* is often the first step towards the development of more robust multilingually applicable NLP technology (O'Horan et al., 2016; Bjerva et al., 2019; Ponti et al., 2019). For instance, selecting suitable source languages is a prerequisite for successful cross-lingual transfer of dependency parsers or POS taggers (Naseem et al., 2012; Ponti et al., 2018; de Lhoneux et al., 2018). In another example, with all other factors kept similar (e.g., training data size, domain similarity), the quality of machine translation also depends heavily on the properties and language proximity of the actual language pair (Kudugunta et al., 2019).

In this work, we contribute to this research endeavor by proposing a suite of spectral-based measures that capture the degree of isomorphism (Søgaard et al., 2018) between the monolingual embedding spaces of two languages. Our main hypothesis is that the potential to align two embedding spaces and learn transfer functions can be estimated through the differences between the monolingual embeddings' spectra. We therefore discuss representative statistics of the spectrum of an embedding space (i.e., the set of the singular values of the embedding matrix), such as its condition number or its sorted list of singular values. We then derive measures for the isomorphism between two embedding spaces based on these statistics.

To validate our hypothesis, we perform an extensive empirical evaluation with a range of cross-lingual NLP tasks. This analysis reveals that our proposed spectrum-based isomorphism measures better correlate and explain greater variance than previous isomorphism measures (Søgaard et al., 2018; Patra et al., 2019). In addition, our measures also outperform standard approaches based on linguistic information (Littell et al., 2017),

The first part of our empirical analysis targets bilingual lexicon induction (BLI), a cross-lingual task that received plenty of attention, in particular as a case study to investigate the impact of cross-language variation on task performance (Søgaard et al., 2018; Artetxe et al., 2018). Its popularity stems from its simple task formulation and reduced

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2377–2390,
November 16–20, 2020.

resource requirements, which makes it widely applicable across a large number of language pairs (Ruder et al., 2019b).

Prior work has empirically verified that for some language pairs BLI performs remarkably well, and for others rather poorly (Søgaard et al., 2018; Vulić et al., 2019). It attempted to explain this variance in performance by grounding it in the differences between the monolingual embedding spaces themselves. These studies introduced the notion of *approximate isomorphism*, and argued that it is easier to learn a mapping function (Mikolov et al., 2013; Ruder et al., 2019b) between language pairs whose embeddings are approximately isomorphic, than between languages pairs without this property (Barone, 2016; Søgaard et al., 2018). Subsequently, novel methods to quantify the degree of isomorphism were proposed, and were shown to significantly correlate with BLI scores (Zhang et al., 2017; Søgaard et al., 2018; Patra et al., 2019).

In this work, we report much higher correlations with BLI scores than existing isomorphism measures, across a variety of state-of-the-art BLI approaches. While previous work was limited only to coarse-grained analysis with a small number of language pairs (i.e., < 10), our study is the first large-scale analysis that is focused on the relationship between quantifiable isomorphism and BLI performance. Our analysis covers hundreds of diverse language pairs, focusing on typologically, geographically and phylogenetically distant pairs as well as on similar languages.

We further show that our findings generalize beyond BLI, to cross-lingual transfer in dependency parsing and POS tagging, and we also demonstrate strong correlations with machine translation (MT) performance. Finally, our spectral-based measures can be combined with typologically driven language distance measures to achieve further correlation improvements. This indicates the complementary nature of the implicit knowledge coded in continuous semantic spaces (and captured by our spectral measures) and the discrete linguistic information from typological databases (captured by the typologically driven measures).

2 Quantifying Isomorphism with Spectral Statistics

Following the distributional hypothesis (Harris, 1954; Firth, 1957), word embedding models learn the meaning of words according to their co-occurrence patterns. Hence, the word embedding space of a language whose words are used in diverse contexts is intuitively expected to encode richer information and greater variance than the word embedding space of a language with more restricting word usage patterns. The difference between two monolingual embedding spaces may also result from other reasons, such as the difference between the training corpora on which the embedding induction algorithm is trained, and the degree to which this algorithm accounts for the linguistic properties of each of the languages.

While the exact combination of factors that govern the difference between the embedding spaces of different languages is hard to figure, this difference is likely to be indicative of the quality of cross-lingual transfer. This is particularly true when the embedding spaces are used by cross-lingual transfer algorithms. Our core hypothesis is that the difference between two monolingual spaces can be quantified by spectral statistics of the two spaces.

2.1 Spectrum Statistics

Given a d-dimensional embedding matrix $\mathbf{X}$, we perform Singular Value Decomposition (SVD) and obtain a diagonal matrix $\mathbf{\Sigma}$ whose main diagonal comprises of d singular values, $\sigma_1, \sigma_2, \ldots, \sigma_d$, sorted in a descending order.[1] Our aim is to quantify the difference between two embedding spaces by comparing statistics of their singular values. We next describe such statistics and in §2.2 use them to measure the isomorphism between the spaces.

Condition Number. In numerical analysis, a function's condition number measures the extent of change of the function's output value conditioned on a small change in the input (Blum, 2014). Consider the case of $\varphi : \mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are two embedding spaces mapped via φ. The condition number, $\kappa(\mathbf{X})$, represents the degree to which small perturbations in the input $\mathbf{X}$ are amplified in the output $\varphi(X)$. Following Higham et al. (2015), we compute the condition number of an input matrix $\mathbf{X}$ with d singular values as the ratio between its first (largest) and last (smallest) singular values:

$$\kappa(\mathbf{X}) = \frac{\sigma_1}{\sigma_d} \quad (1)$$

Why is it a relevant statistic? A smaller condition number denotes a more "stable" matrix that is less

[1] $\mathbf{X}$ is mean-centered, column means have been subtracted and are equal to zero.

sensitive to perturbations. Consequently, learning a transfer function φ from one embedding space to another is more robust to noise when dealing with spaces with smaller $\kappa(\mathbf{X})$. We thus expect that embedding matrices with high condition numbers might impede the learning of good transfer functions in cross-lingual NLP: A function learnt on an embedding space that is sensitive to small perturbations may not generalize well.

Are small singular values reliable? Small singular values are associated with noise, or with the least important information, and many noise reduction techniques remove them (Ford, 2015). If the smallest singular value indeed captures noise, this might affect the condition number (Eq. (1)). It is thus crucial to distinguish between "small but significant" and "small and insignificant" singular values. This is what we do below.

Effective Rank. Given sorted singular values, how can we determine the last effective singular value? For a matrix with d singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$, the ϵ-numerical rank can be defined as: $r_\epsilon = \min\{r : \sigma_r \geq \epsilon\}$, which means that singular values below a certain threshold are removed. However, this formulation introduces a dependency on the hyper-parameter ϵ. To avoid this, Roy and Vetterli (2007) proposed an alternative method that considers the full spectrum of singular values before computing the so-called *effective rank* of the input matrix $\mathbf{X}$:

$$erank(\mathbf{X}) = \left\lfloor e^{H(\Sigma)} \right\rfloor \tag{2}$$

where $H(\Sigma)$ is the entropy of the matrix $\mathbf{X}$'s normalized singular value distribution $\bar{\sigma}_i = \frac{\sigma_i}{\sum_{i=1}^{d} \sigma_i}$, computed as $H(\Sigma) = -\sum_{i=1}^{d} \bar{\sigma}_i \log \bar{\sigma}_i$. $erank(\mathbf{X})$, rounded to the smaller integer, yields the index of the last singular value that is considered significant, and is interpreted as the effective dimensionality, or rank, of the matrix $\mathbf{X}$. If d is the dimensionality of the embedding space $\mathbf{X}$, and we assume that the number of word vectors in $\mathbf{X}$ is typically much larger than d, it then holds that (see Roy and Vetterli (2007)):

$$1 \leq erank(\mathbf{X}) \leq rank(\mathbf{X}) \leq d \tag{3}$$

The dimensionality of an embedding space is intuitively assumed to be equal to the dimensionality of its constituent vectors: the matrix rank. Effective rank undermines this assumption: with effective rank matrices of the same 'initial dimensionality' can have very different 'true dimensionalities' (Yin and Shen, 2018). Effective rank is used for various problems outside NLP, such as source localization for acoustic (Tourbabin and Rafaely, 2015) and seismic (Leeuwenburgh and Arts, 2014) waves, video compression (Bhaskaranand and Gibson, 2010), and for the evaluation of implicit regularization in neural matrix factorization (Arora et al., 2019). We propose to use it to inform and improve the estimation of the condition number.

Effective Condition Number. We replace σ_d in Eq. (1) with the singular value at the position of $\mathbf{X}$'s effective rank (see Eq. (2)), and compute the *effective condition number* κ_{ecn} as follows:

$$\kappa_{ecn}(\mathbf{X}) = \frac{\sigma_1}{\sigma_{erank(\mathbf{X})}} \tag{4}$$

In §5 we empirically validate the quality of the effective condition number in comparison to the standard condition number.

Having defined spectral statistics of an embedding space, we move to define means of comparing two spaces using these statistics.

2.2 Spectral-Based Isomorphism Measures

The statistics described in §2.1 capture properties of a single embedding space, but it is not straightforward how to employ them in order to quantify the similarity between two distinct embedding spaces. In what follows, we introduce isomorphism measures based on the spectral statistics.

Let us assume two embedding matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, with their condition numbers, $\kappa(\mathbf{X_1})$ and $\kappa(\mathbf{X_2})$. We combine the two numbers using the harmonic mean function (HM) to derive an isomorphism measure between two embedding spaces, COND-HM$(\mathbf{X}_1, \mathbf{X}_2)$:

$$\text{COND-HM}(\mathbf{X}_1, \mathbf{X}_2) = \frac{2 \cdot \kappa(\mathbf{X_1})\kappa(\mathbf{X_2})}{\kappa(\mathbf{X_1}) + \kappa(\mathbf{X_2})} \tag{5}$$

We similarly define the ECOND-HM measure over $\kappa_{ecn}(\mathbf{X_1})$ and $\kappa_{ecn}(\mathbf{X_2})$.

Why harmonic mean? The higher the (effective) condition number of an embedding space, the higher its sensitivity to perturbations (i.e., the performance of transfer functions will be low). We view the condition number as a constraining factor on transferability, but what is the right way to evaluate the 'transferability potential' of two spaces via their condition numbers? There are multiple ways to combine two condition numbers, but we

have empirically validated (§5) that HM is a robust choice that outperforms some other possibilities (e.g., the arithmetic mean). We hypothesize this is because HM treats large discrepancies between two numbers in a manner that leans towards the smaller one (unlike e.g. arithmetic mean). Two noisy and two stable embedding spaces would have high and low HMs, respectively, but a noisy embedding space and a stable one would have an HM that leans towards the stable one.[2] Our results suggest that embedding spaces with small condition numbers can often tolerate noisy mappings from embedding spaces with high condition numbers, which might result from the improved stability of the former spaces.

Singular Value Gap. In addition to COND-HM and ECOND-HM, we introduce another measure that empirically quantifies the divergence between the full spectral information of two embedding spaces. This measure quantifies the gap between the singular values obtained from the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ sorted in descending order. We define the measure of Singular Value Gap (SVG) between two d-dimensional spaces $\mathbf{X}_1$ and $\mathbf{X}_2$, as the squared Euclidean distance between the corresponding sorted singular values after log transform:

$$SVG(\mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{d} (\log \sigma_i^1 - \log \sigma_i^2)^2 \quad (6)$$

where σ_i^1 and σ_i^2, $i = 1, \ldots, d$ are the sorted singular values characterizing the two embedding matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. The intuition here is that two embedding spaces with similar singular values at the same index will be more isomorphic and therefore easier to align into a shared space, and enable more effective cross-lingual transfer.

In summary, this section has presented methods that estimate the degree of isomorphism between any given pair of embedding spaces, which may differ in their language, training corpus, embedding induction algorithm or in other factors. While the focus of our empirical analysis (§4, §5) is cross-language learning and transfer, we note that the scope of our methods may be wider, and that they have not been developed only with cross-lingual learning in mind.

[2] Constraining factors are usually handled by taking the minimum (or maximum) values; however, this approach would impede the validity of our correlation analysis (see later §4) as it would artificially decrease the variance in only one variable (the language with the lowest measure will be systematically chosen). We refer the reader to Appendix C for an analysis.

3 Related Work and Baselines

We now provide an overview of prior research that focused on two relevant themes: **1)** measuring approximate isomorphism between two embedding spaces, and **2)** more generally, quantifying the (dis)similarity between languages, going beyond isomorphism measures. The discussed approaches will also be used as the main baselines later in §5.

Measuring Approximate Isomorphism. We focus on two standard isomorphism measures from prior work which are most similar to our work, and use them as our main baselines. The first measure, termed **Isospectrality (IS)** (Søgaard et al., 2018), is based on spectral analysis as well, but of the Laplacian eigenvalues of the nearest neighborhood graphs that originate from the initial embedding spaces $\mathbf{X}_1$ and $\mathbf{X}_2$ (for further technical details see Appendix A). Søgaard et al. (2018) argue that these eigenvalues are compact representations of the graph Laplacian, and that their comparison reveals the degree of (approximate) isomorphism. Although similar in spirit to our approach, constructing nearest neighborhood graphs (and then analyzing their eigenvalues) removes useful information on the interaction between all vectors from the initial space, which our spectral method retains.

The second measure is the **Gromov-Hausdorff distance (GH)** introduced by Patra et al. (2019). It measures the maximum distance of a set of points to the nearest point in another set, or in other words the worst case distance between two metric spaces $\mathcal{X}$ and $\mathcal{Y}$ (for further technical details see again Appendix A). Patra et al. (2019) propose this distance to test how well two language embedding spaces can be aligned under an isometric transformation.

While both IS and GH were reported to have strong correlations with BLI performance in prior work, they have not been evaluated in large-scale experiments before. In fact, the correlations were computed on a very small number of language pairs (IS: 8 pairs, GH: 10 pairs). Further, both measures do not scale well computationally. Therefore, for computational tractability, the scores are computed only on the sub-matrices spanning the sub-spaces of the most frequent subsets from the full embedding spaces (IS: 10k words, GH: 5k words). In this work, we provide full-fledged empirical analyses of the two measures on a much larger number of pairs from diverse languages, and compare them against the spectral-based measures introduced in §2. The fact that the proposed spectral-based meth-

ods are grounded in linear algebra theory (cf. §2.1) also arguably provides a more intuitive understanding of their theoretical underpinning than what is currently offered in the relevant prior work.

Measuring Language Similarity. At the same time, distances between language pairs can also be captured through (dis)similarities in their *discrete* linguistic properties, such as overlap in syntactic features, or proximity along the phylogenetic language tree. The properties are typically handcrafted, and are extracted from available typological databases such as the World Atlas of Languages (WALS) (Dryer and Haspelmath, 2013) or URIEL (Littell et al., 2017), among others (O'Horan et al., 2016; Ponti et al., 2019). Such distances were found useful in guiding and informing cross-lingual transfer tasks (Cotterell and Heigold, 2017; Agić, 2017; Lin et al., 2019; Ponti et al., 2019).

In particular, we compare against three precomputed measures of language distance based on the URIEL typological database (Littell et al., 2017). **Phylogenetic distance (PHY)** is derived from the hypothesized phylogenetic tree of language descent. **Typological distance (TYP)** is computed based on the overlap in syntactic features of languages from the WALS database (Dryer and Haspelmath, 2013). **Geographic distance (GEO)** is obtained from the locations where languages are spoken; see the work of Littell et al. (2017) for more details.

We use these isomorphism measures and linguistic measures as language distance measures. We simply compute language distance between two languages L_1 and L_2 as $LDist(L_1, L_2) = D(\mathbf{X}_1, \mathbf{X}_2)$, where $D = \{$SVG, COND-HM, ECOND-HM, GH, IS, PHY, TYP, GEO$\}$. Later in §5 we show that "proxy" language distances originating from the proposed spectral-based isomorphism measures (see §2.2) correlate better with cross-lingual transfer scores across several tasks, than these language distances which are based on discrete linguistic properties. We verify that implicit knowledge coded in continuous embedding spaces and linguistic knowledge explicitly coded in external databases often complement each other.

4 Experimental Setup

The conducted empirical analyses can be divided into two major parts. First, we run large-scale BLI analyses across several hundred language pairs from dozens of languages, comparing the correlation of spectral-based isomorphism measures (§2.2) and all baselines (§3) with performance of a wide spectrum of state-of-the-art BLI methods. Second, we run further correlation analyses with performances in cross-lingual downstream tasks: dependency parsing, POS tagging, and MT. We first provide the details of the experiments that are shared between the two parts, and then provide further specifics of each experimental part.

Monolingual Word Embeddings. For all isomorphism measures (SVG, COND-HM, ECOND-HM, GH and IS) and languages in our analyses we use publicly available 300-dim monolingual fastText word embeddings pretrained on Wikipedia with exactly the same default settings (see Bojanowski et al. (2017)), length-normalized and trimmed to the 200k most frequent words.[3]

Isomorphism Measures: Technical Details. For our spectral-based measures, we compute a full SVD decomposition (i.e., no dimensionality reduction) of the embedding space. We compute SVG scores for BLI based on the first 40 singular values, which we empirically found to produce slightly better results;[4] for the other tasks we use all singular values. For IS and GH, we replicate the experimental setup from prior work: we compute the IS score over the top 10k most frequent words in each monolingual space, while the GH score is computed over the top 5k words from each monolingual space.[5]

4.1 Bilingual Lexicon Induction

We conduct correlation analyses of the results from previous studies that report BLI scores for a large number of language pairs. On top of that, we complement the existing results from previous research with new results obtained with state-of-the-art BLI methods, applied to additional language pairs.

BLI Setups and Scores. Vulić et al. (2019) ran BLI experiments on 210 language pairs, spanning 15 diverse languages. Their training and test dictionaries (5k and 2k translation pairs) are derived from **PanLex** (Baldwin et al., 2010; Kamholz et al., 2014). We complement the original 210 pairs with additional 210 language pairs of 15 closely related (European) languages using dictionaries extracted from PanLex following the procedure of Vulić et al. (2019). With the additional language set, the aim is

[3] fasttext.cc/docs/en/pretrained-vectors.html

[4] The average gains were around 10%; we note that running SVG with the full set of singular values also outperforms the baseline measures.

[5] github.com/joelmoniz/BLISS

to probe if isomorphism measures can also capture more subtle and smaller language differences.[6]

We also analyze the BLI results of 108 language pairs from **MUSE** (Conneau et al., 2018). This dataset systematically covers English, with 88 translation pairs that involve English as either the source or target language. Finally, we analyze the available BLI results of Glavaš et al. (2019) (referred to as **GTrans**) that are based on dictionaries obtained from Google Translate and include 28 language pairs spanning 8 different languages. For the full list of language pairs involved in previous BLI studies, we refer the reader to prior work (Conneau et al., 2018; Glavaš et al., 2019; Vulić et al., 2019).

BLI Methods in Comparison. The scores in each BLI setup were computed by several state-of-the-art BLI methods based on cross-lingual word embeddings, briefly described here. **1)** SUP is the standard supervised method (Artetxe et al., 2016; Smith et al., 2017) that learns a mapping between two embedding spaces based on a training dictionary by solving the orthogonal Procrustes problem (Schönemann, 1966). **2)** SUP+ is another standard supervised method that additionally applies a variety of pre-processing and post-processing steps (e.g., whitening, dewhitening, symmetric reweighting) before and after learning the mapping matrix, see (Artetxe et al., 2018). **3)** UNSUP is a fully unsupervised method based on the "similarity of monolingual similarities" heuristic to extract the seed dictionary from monolingual data. It then uses an iterative self-learning procedure to improve on the initial noisy dictionary (Artetxe et al., 2018). For more technical details on the fully unsupervised model, we refer the reader to prior work (Ruder et al., 2019a; Vulić et al., 2019).[7]

In sum, our analyses are conducted in three BLI setups (PanLex, MUSE, GTrans) and examine three types of state-of-the-art mapping-based methods, both supervised and unsupervised (SUP, SUP+, UNSUP). Altogether, these span 556 language pairs, and cover both related and distant languages.[8] Following prior work (Glavaš et al., 2019), our BLI evaluation measure is Mean Reciprocal Rank (MRR). We note that identical findings emerge from running the correlation analyses based on Precision@1 scores in lieu of MRR.

4.2 Downstream Tasks

Following the large-scale nature of our BLI analyses, we run similar correlation analyses on several downstream tasks that comprise a large number of (both similar and distant) language pairs.[9] We rely on results from a recent study of Lin et al. (2019) that focused on cross-lingual transfer performance in MT, dependency parsing, and POS tagging.[10]

Machine Translation. Lin et al. (2019) report BLEU scores when translating 54 source L_1 languages into English as the target language. We report correlations between the different language distance measures and these 54 BLEU scores.

Dependency Parsing. We base our analysis on the cross-lingual zero-shot parser transfer results of Lin et al. (2019): The standard biaffine dependency parser (Dozat and Manning, 2017; Dozat et al., 2017) is trained on the training portions of Universal Dependencies (UD) treebanks from 31 languages (Nivre et al., 2018), and is then used to parse the test treebank of each language, now used as the target language. We report correlations between the language distance measures and the Labeled Attachment Scores (LAS) for all combinations of 31 languages, resulting in 930 pairs.

POS Tagging. We use POS tagging accuracy scores reported by Lin et al. (2019). These scores span 26 low-resource target languages and 60 source languages which measure the utility of each source language to each of the 26 target languages in POS tagging. We use a sample of 840 language pairs for the correlation analysis, as 16 low-resource target languages and 49 source languages have readily available pretrained fastText vectors.

[6] The initial set of Vulić et al. (2019) comprises Bulgarian, Catalan, Esperanto, Estonian, Basque, Finnish, Hebrew, Hungarian, Indonesian, Georgian, Korean, Lithuanian, Norwegian, Thai, Turkish. The additional 210 language pairs are only composed of Germanic, Romance and Slavic languages. For a full list of the languages see Table 4 in the appendix.

[7] The SUP+ and UNSUP methods are based on the VecMap framework (`github.com/artetxem/vecmap`) which showed very competitive and robust BLI performance across a wide range of language pairs in recent comparative analyses (Glavaš et al., 2019; Vulić et al., 2019; Doval et al., 2019).

[8] We report all results for each BLI method, dictionary and language pairs in the supplementary material (and also here `https://tinyurl.com/skn5cf7`). We also report scores with another method, RCSLS (Joulin et al., 2018), benchmarked in the GTrans BLI setup (see Table 1).

[9] For the full list of languages that were analyzed throughout all our experiments see Table 4 in the appendix.

[10] For full details regarding the models used to compute the scores for each downstream task, we refer the interested reader to the work of Lin et al. (2019) and the accompanying repository: `https://github.com/neulab/langrank`. We note that scores for each language pair in each task have been produced with the same task architectures.

4.3 Correlation Analyses and Statistical Tests

All scores from isomorphism measures and BLI scores were log-transformed[11] prior to any correlation computation. We report Pearson's correlation coefficients in all tasks. This allows us to investigate which of the different individual measures is most important to predict task performance.

Regression Analyses. The individual (i.e., single-variable) analyses are not sufficient to account for the complex interdependencies between the distance measures themselves, and how they interact with task performance when combined. Therefore, we also use standard linear stepwise regression model (Hocking, 1976; Draper and Smith, 1998):

$$\mathbf{Y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon. \quad (7)$$

Here, task performance $\mathbf{Y}$ is predicted using a set of regressors, $x_1, \ldots, x_n$ (i.e., SVG, COND-HM, ECOND-HM, GH, IS, PHY, TYP, GEO), that are added to the model incrementally only if their marginal addition to predicting $\mathbf{Y}$ is statistically significant ($p < .01$). This method is useful for finding variables (i.e., in our case distance measures) with *maximal* and *unique* contribution to the explanation of $\mathbf{Y}$, when the variables themselves are strongly cross-correlated, as in our case.

This model is able to: (a) discern which variables overlap in their information; (b) detect variables that complement each other; and (c) evaluate their joint contribution in predicting task performance. We compute the regression model's score for all statistically significant variables, and report its square-root, $\hat{r}$.

Importantly, $\hat{r}$ is *not* a one-number description of a language, but rather an illustrative quantification of the joint contribution of several *different* distance measures to the explanation of $\mathbf{Y}$. Its goal is to investigate potential gains achieved through the combination of several distance measures, as opposed to using a single-best measure. The distance measures that are found statistically significant in the regression analyses are marked by superscripts over $\hat{r}$ (see later in Tables 1 and 2).

5 Analyses and Results

The results are summarized in Tables 1 and 2. The first main finding is that our proposed spectral-based isomorphism measures are strongly correlated with performance across all tasks and settings.[12] In fact, they show the strongest individual correlations with task performance among all isomorphism measures and linguistic distances alike. The only exception is the MT task, where our measures fall short of TYP (see Table 2), although we mark that they still hold a strong advantage over the baseline GH and IS isomorphism measures that do not seem to capture any useful language similarity properties needed for the MT task.

ECOND-HM systematically outperforms COND-HM on 2 of 3 BLI datasets and 2 of 3 downstream tasks, validating our assumption that discarding the smallest singular values reduces noise. Additionally, SVG shows greater stability across tasks and datasets than ECOND-HM. A general finding across all tasks is that our spectral measures are the most robust isomorphism measures: they substantially outperform the widely used baselines GH and IS.

As stepwise regression discerns between overlapping and complementing variables (see §4.3), another finding indicates that our spectral measures are complemented by linguistically driven language distances. Indeed, their combination achieves very high correlation scores. The results demonstrate this across all tasks and settings (see bottom rows of the tables). For instance, when combining spectral measures with the linguistic distances, the regression model reaches outstanding correlation scores up to $r = .91$ on PanLex BLI (Table 1); with 420 language pairs, PanLex is the most comprehensive BLI dataset in our study. In addition, GH and IS are not chosen as significant regressors in the stepwise regression model, which indicates that they capture less information than our spectral methods.[13] Overall, the regression results support the notion that conceptually different distances (i.e., isomorphism-based versus measures based on linguistic properties) capture different properties of similarities between languages, which has a synergistic effect when they are combined.

Concerning individual tasks, we note that our spectral-based measures outperform the baselines

[11] This is an order-preserving transformation which is often used when the data distribution is skewed (as was the case with our BLI and isomorphism scores).

[12] The negative correlations between SVG and ECOND-HM scores and task performance have a clean interpretation: they are distance/dissimilarity measures, so high scores of those measures indicate less similar languages.

[13] This does not imply GH or IS do not capture information that is complementary to discrete linguistic information, only that the implicit knowledge captured by GH or IS cannot match that of our spectral-based methods, and their combination with linguistic distances obtained from external knowledge sources.

	PanLex			MUSE			GTrans		
	SUP*	SUP+	UNSUP	SUP	SUP+	UNSUP	SUP	RCSLS	UNSUP
1. SVG	-.74	-.76	**-.74**	**-.60**	-.60	**-.61**	-.59	-.59	-.53
2. COND-HM	-.66	-.64	-.56	-.58	-.47	-.39	-.46	-.47	-.42
3. ECOND-HM	**-.79**	**-.77**	-.70	-.46	-.38	-.39	**-.74**	**-.75**	**-.56**
4. GH	-.41	-.41	-.46	-.43	-.33	-.28	-.47	-.46	-.50
5. IS	-.51	-.53	-.49	-.42	-.42	-.34	-.33	-.33	-.36
6. PHY	-.55	-.55	-.41	-.53	-.62	-.34	-.51	-.50	-.52
7. TYP	-.57	-.52	-.38	-.47	**-.64**	-.44	-.59	-.59	-.43
8. GEO	-.62	-.66	-.62	-.57	-.54	-.35	-.22	-.22	-.35
$\hat{r}$	$.91^{1,3,6-8}$	$.91^{1,3,6-8}$	$.82^{1,3,8}$	$.69^{1,6}$	$.74^{7,8}$	$.61^{1}$	$.92^{3,6,8}$	$.93^{3,6,8}$	$.86^{3,6,8}$

Table 1: Correlations with BLI performance in three BLI setups, see §4.1. The best distance measure for each setup and BLI method is **bolded**. $\hat{r}$ is the score from the stepwise regression model, see §4.3. Superscripts indicate the distance measures that are statistically significant and included in the stepwise regression model (e.g., $.91^{1,3,6-8}$ means: SVG, ECOND-HM and all the linguistic distances have a combined contribution equivalent to 0.91 Pearson). *See the scatter plot in Appendix C.

	MT	DEP	POS
1. SVG	-.56	**-.79**	**-.52**
2. COND-HM	-.52	-.60	-.41
3. ECOND-HM	-.51	-.66	-.44
4. GH	-.16	-.70	**-.52**
5. IS	-.13	-.67	-.43
6. PHY	-.45	-.56	-.16
7. TYP	**-.66**	-.40	-.16
8. GEO	-.26	-.58	-.15
$\hat{r}$	$.74^{1,7}$	$.87^{1,6-8}$	$.59^{1,4,5,7}$

Table 2: Correlations with performance in three other cross-lingual tasks: Machine Translation (MT), dependency parsing (DEP), and POS tagging. Results for the best distance measure are highlighted in **bold**. $\hat{r}$ is computed using the stepwise regression model (see §4.3).

regardless of the underlying BLI method. Further, SVG is the most informative distance measure in parsing experiments, and all linguistic distances fall behind isomorphism measures. Combining linguistic distances with SVG increases the already high correlation from 0.79 to 0.87 (Table 2, second column). For the POS tagging task, SVG and GH are on par, and the combination with IS and TYP increase their correlation from 0.52 to 0.59 (Table 2, third column). This is the only analysis where baseline methods are found significant.

Additional results and analyses are provided in Appendix B. They further demonstrate that our measures also indicate transfer quality of different target languages for a given source language, and transfer quality of source languages for a given target language, for the tasks discussed in this paper.

6 Further Discussion and Conclusion

This work introduces two spectral-based measures, SVG and ECOND-HM, that excel in predicting performance on a variety of cross-lingual tasks. Both measures leverage information from singular values in different ways: ECOND-HM uses the ratio between two singular values, and is grounded in linear algebra and numerical analysis (Blum, 2014; Roy and Vetterli, 2007), while SVG directly utilizes the full range of singular values. We suspect that the use of the full range of singular values is what makes SVG more robust across different tasks and datasets, compared to ECOND-HM that shows greater variance, as observed in our results above.

While the spectral methods are computed solely on word vectors from Wikipedia, the results in the downstream tasks are computed with different sets of embeddings (e.g., multilingual embeddings for dependency parsing), or the embeddings are learnt during training (for POS tagging and MT). We believe that this discrepancy does not constitute a shortcoming of our analyses, but rather the opposite: spectral-based methods maintain their high correlations in the downstream tasks as well, and this supports the notion that these measures might indeed capture deeper linguistic information than mere similarities between embedding spaces.

Our use of effective rank in improving the condition number (via effective condition number) is also inspired by recent work that aimed to automatically detect true dimensionality of embedding spaces. However, previous work has taken an empirical approach by simply tuning embedding dimensionality to evaluation tasks at hand (Wang, 2019; Raunak

et al., 2019; Carrington et al., 2019). Our intention, on the other hand, is to extract the true embedding dimensionality directly from the embedding space. Another recent study (Yin and Shen, 2018) employed perturbation analysis to study the robustness of embedding spaces to noise in monolingual settings, and established that it is also related to effective dimensionality of the embedding space. All these inspired us to replace the standard matrix rank with effective rank when computing the condition number, and to introduce the statistic of effective condition number in §2.1.

Our study is also the first to compare language distance measures that are based on discrete linguistic information (Littell et al., 2017) with measures of isomorphism (i.e., our spectral-based measures, IS, GH), which can also be used as proxy language distance measures. Our findings, suggesting that it is possible to effectively combine these two types of language distance measures, call for further research that will advance our understanding of: 1) what knowledge is captured in monolingual and cross-lingual embedding spaces (Gerz et al., 2018; Pires et al., 2019; Artetxe et al., 2020); 2) how that knowledge complements or overlaps with linguistic knowledge compiled into lexical-semantic and typological databases (Dryer and Haspelmath, 2013; Wichmann et al., 2018; Ponti et al., 2019); and 3) how to use the combined knowledge for more effective transfer in cross-lingual NLP applications (Ponti et al., 2018; Eisenschlos et al., 2019).

The differences in embedding spaces of different languages do not only depend on linguistic properties of the languages in consideration, but also on other factors such as the chosen training algorithm, underlying training domain, or training data size and quality (Søgaard et al., 2018; Arora et al., 2019; Vulić et al., 2020). In future research we also plan an in-depth study of these factors and their relation to our spectral analysis.

We believe that the main insights from this study will inform and guide different cross-lingual transfer learning methods and scenarios in future work. These might range from choosing source languages for transfer in low-data regimes, over monolingual word vector induction guided by the spectral statistics, even to more effective hyperparameter search.

Acknowledgments

The work of IV and AK is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909) awarded to AK. HD is supported by the Blavatnik Postdoctoral Fellowship Programme. HD would like to thank Dr. Avital Hahamy for her contribution to the initiation of this work.

References

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

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. 2019. Implicit regularization in deep matrix factorization. In *Proceedings of NeurIPS*, pages 7411–7422.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*, pages 2289–2294.

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

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*.

Timothy Baldwin, Jonathan Pool, and Susan Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of COLING (Demo Papers)*, pages 37–40.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Malavika Bhaskaranand and Jerry D Gibson. 2010. Spectral entropy-based quantization matrices for H264/AVC video coding. In *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, pages 421–425.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Lenore Blum. 2014. *Alan Turing and the other theory of computation (expanded)*, volume 42 of *Lecture Notes in Logic*, pages 48–69. Cambridge University 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.

Rachel Carrington, Karthik Bharath, and Simon Preston. 2019. Invariance and identifiability issues for word embeddings. In *Proceedings of NeurIPS*, pages 15114–15123.

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

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of EMNLP*, pages 748–759.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2019. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. *CoRR*, abs/1908.07742.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR*.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Norman R. Draper and Harry Smith. 1998. *Applied Regression Analysis, 3rd Edition*. John Wiley & Sons.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of EMNLP*, pages 5702–5707.

John Rupert Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

William Ford. 2015. Chapter 15 - the singular value decomposition. In William Ford, editor, *Numerical Linear Algebra with Applications*, pages 299 – 320. Academic Press, Boston.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

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*, pages 710–721.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

N.J. Higham, M.R. Dennis, P. Glendinning, P.A. Martin, F. Santosa, and J. Tanner. 2015. *The Princeton Companion to Applied Mathematics*. Princeton University Press.

Ronald R. Hocking. 1976. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP*, pages 2979–2984.

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

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of EMNLP-IJCNLP*, pages 1565–1575.

Olwijn Leeuwenburgh and Rob Arts. 2014. Distance parameterization for efficient seismic history matching with the ensemble kalman filter. *Computational Geosciences*, 18(3-4):535–548.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of EMNLP*, pages 4992–4997.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*, pages 3125–3135.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of EACL*, pages 8–14.

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

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of ACL*, pages 629–637.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, et al. 2018. Universal Dependencies 2.3.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING*, pages 1297–1308.

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*, pages 184–193.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Edoardo Maria Ponti, Helen OHoran, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of ACL*, pages 1531–1542.

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference*, pages 606–610.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019a. Unsupervised cross-lingual representation learning. In *Proceedings of ACL: Tutorial Abstracts*, pages 31–38.

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

Peter H. Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

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

Vladimir Tourbabin and Boaz Rafaely. 2015. Direction of arrival estimation using microphone array processing for moving humanoid robots. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):2046–2058.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP*, pages 4398–4409.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of EMNLP*.

Yu Wang. 2019. Single training dimension selection for word embedding with PCA. In *Proceedings of EMNLP-IJCNLP*, pages 3588–3593.

Søren Wichmann, André Müller, Viveka Velupillai, Cecil H Brown, Eric W Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, et al. 2018. The ASJP database (version 18).

Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. In *Proceedings of NeurIPS*, pages 887–898.

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

A IS and GH

Isospectrality (IS) After length-normalizing the vectors, Søgaard et al. (2018) compute the nearest neighbor graphs using a subset of the top N most frequent words in each space, and then calculate the Laplacian matrices LP_1 and LP_2 of each graph. For LP_1, the smallest k_1 is then sought 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. The same procedure is used to find k_2. They then define $k = \min(k_1, k_2)$. The final IS measure Δ is then 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, consequently, the more isomorphic are the two embedding spaces.

Gromov-Hausdorff Distance (GH) It measures the worst case distance between two metric spaces $\mathcal{X}$ and $\mathcal{Y}$ with a distance function m as:

$$\mathcal{H}(\mathcal{X}, \mathcal{Y}) = \max\{\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} m(x, y), \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} m(x, y)\} \tag{8}$$

It computes the distance between the nearest neighbors that are farthest apart. The Gromov-Hausdorff distance then minimizes this distance over all isometric transforms $\mathcal{X}$ and $\mathcal{Y}$:

$$\mathcal{GH}(\mathcal{X}, \mathcal{Y}) = \inf_{f,g} \mathcal{H}(f(\mathcal{X}), g(\mathcal{Y})) \tag{9}$$

Computing $\mathcal{GH}$ directly is computationally intractable in practice, but it can be tractably approximated by computing the Bottleneck distance between the metric spaces (Chazal et al., 2009).

B Source and Target Selection Analysis

In addition to the correlation analyses in the main text that aggregate all language pairs, some tasks and datasets also support analyses where one language is fixed as a target language (i.e., *source-language selection analysis*), or as a source language (i.e., *target-language selection analysis*). Such analyses could inform us on how to choose the right transfer language. That is, given a target language one would like to transfer to, which is the best source language to transfer from, and vice versa. These analyses are conducted for the following tasks with sufficient language pairs: BLI with PanLex, parsing, and POS tagging.

For these analyses we report average correlation: across target languages in the source-language selection analysis, or across the source languages in the target-language selection analysis. We provide the percentage of times each compared measure scored the highest for a particular task and setting.

Stepwise regression analysis is not suitable for the selection analysis due to the limited number of language pairs in each language selection setup (e.g., PanLex offers 14 language pairs for each source- or target- language selection analysis). These conditions impede the statistical significance power of the tests which stepwise regression requires. We therefore opt for a standard multiple linear regression model instead; the regressors include the isomorphism measure with the highest individual correlation combined with the linguistic measures. Similarly to the stepwise analysis, we report the unified correlation coefficient, $\hat{r}$.

We observe that the same findings reported for the aggregated correlation analyses (Tables 1 and 2 in the main text) also hold for the language selection analyses (Table 3 below). This indicates that our spectral measures have an applicative value as they can facilitate better cross-language transfer.

We also observe interesting patterns in the selection analyses for the POS tagging task in Table 3: While the results in the target-language selection analysis largely follow the main-text results, the same does not hold for source-language selection (Table 3, POS Target and Source columns). We speculate that this is in fact an artefact of the experimental design of Lin et al. (2019). Their set of target languages deliberately comprises only truly low-resource languages, and such languages are expected to have lower-quality embedding spaces. Transferring to such languages is bound to fail with most source languages regardless of the actual source-target language similarity. The difficulty of this setting is reflected in the actual scores: average accuracy scores for the best source-target combination is 0.55 in the source-language selection analysis, and 0.92 for target-language selection.

C Single and Combined Analysis

We show (Figure 1) a single experimental condition (SUP method in the PanLex BLI dataset, leftmost column in Table 1 in the main text) to illustrate the data distribution and the correlation for spectral-based measures (e.g., ECOND-HM), and their improvement once this measure is combined with linguistic measures through regression analysis.

Task			**BLI**				**DEP**		**POS**	
Selection		Target			Source		Target	Source	Target	Source
	SUP	SUP+	UNSUP	SUP	SUP+	UNSUP				
SVG	-.57$^{31\%}$	-.56$^{31\%}$	**-.54**$^{35\%}$	-.53$^{28\%}$	**-.54**$^{29\%}$	**-.56**$^{38\%}$	**-.62**$^{56\%}$	**-.73**$^{68\%}$	-.51$^{4\%}$	-.10
COND-M	-.59	-.55	-.42	-.51	-.49	-.38	-.44	-.64	-.58	-.08
ECOND-M	**-.64**$^{31\%}$	**-.57**$^{25\%}$	-.44$^{17\%}$	**-.54**$^{25\%}$	-.53$^{24\%}$	-.43$^{17\%}$	-.51$^{3\%}$	-.70$^{6\%}$	**-.61**$^{10\%}$	.05
GH	-.32	-.33$^{3\%}$	-.27$^{7\%}$	-.31	-.28	-.26	-.51$^{16\%}$	-.63$^{13\%}$	-.58$^{34\%}$	-.11
IS	-.30	-.30	-.30$^{3\%}$	-.33$^{3\%}$	-.34	-.32$^{7\%}$	-.52$^{16\%}$	-.58$^{10\%}$	-.43$^{50\%}$	-.23
PHY	-.25$^{7\%}$	-.27$^{3\%}$	-.23$^{3\%}$	-.41$^{24\%}$	-.42$^{34\%}$	-.32$^{10\%}$	-.44	-.42	-.14	-.14
TYP	-.50$^{24\%}$	-.49$^{28\%}$	-.42$^{28\%}$	-.45$^{17\%}$	-.41$^{10\%}$	-.36$^{21\%}$	-.40$^{3\%}$	-.42	-.18$^{2\%}$	-.05
GEO	-.29$^{7\%}$	-.34$^{10\%}$	-.32$^{7\%}$	-.31$^{3\%}$	-.32$^{3\%}$	-.32$^{7\%}$	-.48$^{6\%}$	-.56$^{3\%}$	-.12	**-.33**
$\hat{r}$	.86	.83	.76	.83	.83	.77	.79	.85	.68	–

Table 3: Correlation scores in *source-language* (Source) and *target-language* (Target) selection analyses. The best distance measure per column is provided in **bold**. The percentage of cases a measure topped the others is shown in superscript (see details in Appendix B). $\hat{r}$ refers to the unified correlation coefficient from the multiple regression model (see details in Appendix B).

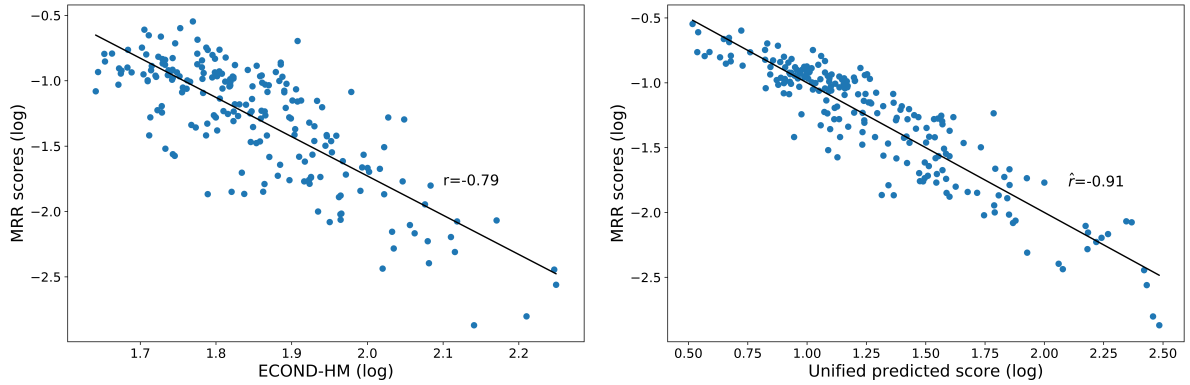

Figure 1: Scatter plots with least square regression lines for the SUP method in the PanLex BLI model (leftmost column in Table 1 of the main paper). The left panel presents results for the best single isomorphism measure, ECOND-HM. The right panel presents results for the combined unified model based on the regression analysis $\hat{r}$ that includes linguistic measures. $\hat{r}$'s sign was flipped (right panel) to make the graphs directly comparable.

Language family	Language	PanLex	MUSE	GTrans	MT	DEP	POS
Germanic (IE)	English	28	88	7	54	62	17
Germanic (IE)	German	28	10	7	1	62	17
Germanic (IE)	Dutch	28	2		1	62	17
Germanic (IE)	Swedish	28	2		1	62	17
Germanic (IE)	Danish	28	2		1	62	17
Germanic (IE)	Norwegian	28	2		1	62	17
Germanic (IE)	Afrikaans		2				65
Germanic (IE)	Faroese						50
Romance (IE)	Italian	28	10	7	1	62	17
Romance (IE)	Portuguese	28	10		1	62	17
Romance (IE)	Spanish	28	10		1	62	17
Romance (IE)	French	28	10	7	1	62	17
Romance (IE)	Romanian	28	2		1	62	17
Romance (IE)	Catalan	28	2			62	17
Romance (IE)	Galician				1		17
Romance (IE)	Latin					62	17
Slavic (IE)	Croatian	28	2	7	1	62	17
Slavic (IE)	Polish	28	2		1	62	17
Slavic (IE)	Russian	28	2	7	1	62	17
Slavic (IE)	Czech	28	2		1	62	17
Slavic (IE)	Bulgarian	56	2		1	62	17
Slavic (IE)	Bosnian		2		1		
Slavic (IE)	Macedonian		2		1		
Slavic (IE)	Slovak		2		1	62	17
Slavic (IE)	Slovenian		2		1	62	17
Slavic (IE)	Ukrainian		2		1	62	17
Slavic (IE)	Belarusian				1		65
Slavic (IE)	Serbian				1		17
Hellenic (IE)	Modern Greek		2		1		17
Balto-Slavic (IE)	Lithuanian	28	2		1		65
Balto-Slavic (IE)	Latvian		2			62	17
Indo-Iranian (IE)	Bengali		2		1		
Indo-Iranian (IE)	Persian		2		1		17
Indo-Iranian (IE)	Hindi		2		1	62	17
Indo-Iranian (IE)	Kurdish				1		
Indo-Iranian (IE)	Marathi				1		65
Indo-Iranian (IE)	Urdu				1		17
Indo-Iranian (IE)	Sanskrit						50
Celtic (IE)	Irish						65
Celtic (IE)	Breton						50
Albanian (IE)	Albanian		2		1		
Armenic (IE)	Armenian				1		65
Turkic	Turkish	28	2	7	1		17
Turkic	Azerbaijani				1		
Turkic	Kazakh				1		65
Turkic	Uighur						17
Semitic	Hebrew	28	2		1	62	17
Semitic	Arabic		2		1	62	17
Semitic	Amharic						50
Uralic	Hungarian	28	2		1		65
Uralic	Finnish	28	2	7	1	62	17
Uralic	Estonian	28	2		1	62	17
Austronesian	Indonesian	28	2		1	62	17
Austronesian	Malay		2		1		
Austronesian	Tagalog		2				50
Dravidian	Tamil		2		1		65
Dravidian	Telugu						65
Sino-Tibetan	Chinese		2		1	62	17
Sino-Tibetan	Burmese				1		
Japonic	Japanese		2		1	62	17
Kartvelian	Georgian	28			1		
Koreanic	Korean	28	2		1	62	17
Mongolic	Mongolian				1		
Niger-Congo	Yoruba						50
Austroasiatic	Vietnamese		2		1		17
Tai-Kadai	Thai	28	2		1		50
Isolate	Basque	28			1		17
Not defined	Esperanto	28			1		

Table 4: Summary of all the languages included in our analyses. The numbers in each cell indicate the number of different language pairs where each language was included, per each task and dataset. IE refers to the Indo-European language group.