# Tracing Source Language Interference in Translation with Graph-Isomorphism Measures

**Koel Dutta Chowdhury**
Saarland Informatics Campus
Saarland University
koeldc@lst.uni-saarland.de

**Cristina España-Bonet**
DFKI GmbH, Saarbrücken
cristinae@dfki.de

**Josef van Genabith**
Saarland University
DFKI GmbH, Saarbrücken
josef.van_genabith@dfki.de

## Abstract

Previous research has used linguistic features to show that translations exhibit traces of source language interference and that phylogenetic trees between languages can be reconstructed from the results of translations into the same language. Recent research has shown that instances of translationese (source language interference) can even be detected in embedding spaces, comparing embeddings spaces of original language data with embedding spaces resulting from translations into the same language, using a simple Eigenvector-based divergence from isomorphism measure. To date, it remains an open question whether alternative graph-isomorphism measures can produce better results. In this paper, we $(i)$ explore Gromov-Hausdorff distance, $(ii)$ present a novel spectral version of the Eigenvector-based method, and $(iii)$ evaluate all approaches against a broad linguistic typological database (URIEL). We show that language distances resulting from our spectral isomorphism approaches can reproduce genetic trees on a par with previous work without requiring any explicit linguistic information and that the results can be extended to non-Indo-European languages. Finally, we show that the methods are robust under a variety of modeling conditions.

## 1 Introduction

The study of cross-linguistic variation has been a key focus of linguistics for genetic or typological classification of languages. Historical comparative linguistic methods determine genetic relationships between languages using concept lists of words with a common origin in multiple languages that share similar meaning and pronunciation (Swadesh, 1952; Dyen et al., 1992). Linguistic typology studies how distinct languages are, and what generalizations can be made regarding cross-linguistic variation on different levels of linguistic analysis and representations (Trask, 2000). Comrie (1989), for example, studies language variance in terms of their functional processes, whereas Cysouw (2013) measures language distance using structural features. More recent research indicates that semantic similarity between languages can serve as a quantitative means to determine cross-linguistic variation across languages. Seminal work of Eger et al. (2016) provides evidence that semantic alignment between languages can be explained by geographical factors. Likewise, Thompson et al. (2018) find that differences correlate with cultural distances among societies speaking the languages.

Conversely, it has also been shown that language differences are so profound that the structure of a language is approximately preserved even when translated into another language. This is often referred to as source language interference (Toury, 2012). Rabinovich et al. (2017) show that source languages of translations into the same target language can be clustered solely based on interference phenomena in the translations in the target language using simple linguistic features and that these clusters correspond with genetic distance. In a similar vein, Bjerva et al. (2019) find that comparable results can be established by clustering neural language model (NLM) based vectors using raw words, part-of-speech (POS) tags, phrase-structure or dependency-based input sequence representations of the data, showing that the distances between these learned language representations are more reflective of syntactic (structural) similarity rather than genetic relationship. Chowdhury et al. (2020) show that source language interference is even evident in simple word, POS, synset and semantic tag based embedding spaces computed from originally authored and data translated into the same target language. They use a graph-based Eigenvector (EV) divergence from iso-



*Proceedings of Recent Advances in Natural Language Processing*, pages 375–385
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_043

morphism distance measure (Søgaard et al., 2018), originally used for bilingual dictionary induction, to capture divergence from isomorphism between monolingual original and translation embedding spaces. With this, they quantify distances between the source languages of the translations and predict phylogenetic trees, and analyse the correlation between isomorphism measure based distances and genetic relations in language families.

However, to date, $(i)$ alternative graph-based distance metrics have not yet been explored for embedding-based approaches to detect translationese; $(ii)$ it is not clear how word-embedding based approaches fare under different data settings including $a)$ varying the number of most frequent words considered in the graphs, $b)$ different corpus sizes and $c)$ different word embedding architectures; $(iii)$ it is not clear how the previous approaches (using either linguistic feature vectors, NLM based feature vectors, or divergence from isomorphism graph-based distance between embedding spaces) compare on the same data against the commonly used gold standard phylogenetic tree of Serva and Petroni (2008) (SP08); $(iv)$ it is not clear how function words would affect graph-based distances; $(v)$ evaluation of the graph- and embedding-space approach against the broader URIEL typological data base (Littell et al., 2017) has not been carried out; and $(vi)$ it is not clear if the scope of this research can be expanded to include non-Indo-European languages.

In this paper, we show $(i)$ that Gromov-Hausdorff (GH) distance can be used as a distance metric to quantify divergence from isomorphism between simple embedding spaces in monolingual settings and develop a novel Spectral Graph-based (SGM) distance measure, extending the original EV-based approach; $(ii)$ that graph- and embedding-based distances are fairly robust under different data settings and that they are not sensitive to skip-gram or CBOW-based embeddings; $(iii)$ divergence from isomorphism graph-based measures using embeddings can reproduce genetic trees on a par with linguistic feature vector and NLM based approaches (Rabinovich et al., 2017; Bjerva et al., 2019); $(iv)$ that function words and concept lists are still relevant within this general approach; $(v)$ that graph- and embedding space-based distance metrics correlate not only with genetic features but also with geographical and syntactic ones (Littell et al., 2017); and $(vi)$ that this research can be extended to translations from non-Indo-European languages.

The rest of the paper is organised as follows. We review related work in Section 2. Section 3 introduces the concept of graph isomorphism and our SGM measure, together with EV and GH. We describe the experimental setting in Section 4. In Section 5, we report results on the isomorphism metrics, infer language family relationships and correlate them with linguistic benchmarks. We describe robustness experiments in Section 6 and compare to previous work in Section 7. Finally, we extend our analysis to non-Indo-European source languages in Section 8 and summarize and draw conclusions in Section 9.

## 2 Related Work

Representational distance between two languages refers to how different one language or language variety is from another. Several analyses (Malaviya et al., 2017; Oncevay et al., 2020) have attempted to disentangle the typological factors that influence language representational distance. Rabinovich et al. (2017) clustered languages based on linguistically inspired features of their translations into the same target language and show that syntactic footprints of the source language in the translations can be used to estimate phylogenetic similarities between their source languages. They use agglomerative clustering with variance minimization (Ward Jr, 1963) as linkage procedure and compare their generated trees ($P$) to the pruned gold-tree ($g$) (SP08) of Serva and Petroni (2008). Their comparison metric is the sum of squared deviations between each language pair's gold-tree distance $D_g$ and corresponding distance in their computed tree $D_P$ :

$$Dist(P, g) = \sum_{i,j} (D_P(l_i, l_j) - D_g(l_i, l_j))^2 \quad (1)$$

It is worth noting that the use of SP08 as a gold standard has also been questioned in the literature. Fortson IV (2011) observes that the SP08 approximation is only suited for a small subset of languages and that it fails to explain finer-grained inconsistencies in the Indo-European language family.

Bjerva et al. (2019) expand the work of Rabinovich et al. (2017) in their NLM- and sequence-based approach and argue that representational distance between languages can be better explained by structural relatedness than by language genetics. Chowdhury et al. (2020) use departures from

isomorphism based on the EV measure (Søgaard et al., 2018) on simple embeddings to infer genealogical distances. They compare different embedding spaces (word, POS, synset or semantic tags) constructed from translations into a single target language and the target language in terms of how similar their corresponding nearest neighborhood graphs are by analyzing their eigenvalues.

The similarity between languages can also be measured using the (dis)similarities between their discrete linguistic properties. Such properties are typically handcrafted and collected in typological databases such as URIEL (Littell et al., 2017) which lists a large inventory of properties for 8000 languages of various typological characteristics, such as overlap in syntactic features, or proximity along phoneme features (Cysouw, 2013). URIEL is a compilation of a variety of linguistic resources including the World Atlas of Language Structure, WALS (Dryer, 2009), PHOIBLE (Moran et al., 2014), Ethnologue (Lewis et al., 2015), and Glottolog (Nordhoff and Hammarström, 2011). Based on linguistic feature vectors, URIEL provides precomputed distance statistics between any language pairs stored in the database in terms of various metrics including genetic, geographical, syntactic, phonological, and phonetic inventory distances.

In this work, we follow the approach of Rabinovich et al. (2017) and evaluate our geometrical measures against the phylogenetic benchmark SP08. We compute the branching length directly from SP08, assuming it reflects the actual proportions. Additionally, we follow He et al. (2019) to compare our generated trees against the average of three precomputed measures of language distance, namely genetic, geographic, and syntactic distances based on the URIEL database.

## 3 Graph Isomorphism

We define the *distance between languages* based on word usage and the notion of isomorphism. An isomorphism $f$ between two metric spaces ($\mathcal{X}$, $d_{\mathcal{X}}$) and ($\mathcal{Y}$, $d_{\mathcal{Y}}$), where $\mathcal{X}$ and $\mathcal{Y}$ are two sets of words in two languages and $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are the metric distances, is a function $f : \mathcal{X} \mapsto \mathcal{Y}$ that is a distance preserving transformation i.e.: for all pairs of points $x_1$ and $x_2$ in $\mathcal{X}$ such that, $d_{\mathcal{Y}}(f(x_1), f(x_2)) = d_{\mathcal{X}}(x_1, x_2)$.

For a vocabulary $V = v_0, v_1, ...., v_n$ in language $\ell$, we define its graph as $G(V, E, w)$, where $V$ denotes the set of vertices corresponding to the vocabulary words; $E = e_0, e_1, ..., e_m$ is a set of edges; and every pair $\{v_i, v_j\}$ has a non-negative edge weight $w_{ij}$ associated with it. Our approach starts with mapping words $v_i^{\ell}$ in language $\ell$ onto points $\mathbf{v_i}^{\ell}$ using distributional semantics methods. Each language is then represented with its own graph $G^{\ell}$. After mapping words onto points $\mathbf{v_i}^{\ell}$ as vectors, the distance between words is defined as the distance between their vectors. We quantify the similarity between languages $\ell_1$ and $\ell_2$ through a distance function between their graphs $d(G^{\ell_1}, G^{\ell_2})$. In what follows we make the concept mentioned above more concrete.

### 3.1 Gromov-Hausdorff (GH) Distance

The first measure we use to quantify the similarity between languages is the Gromov-Hausdorff distance (GH) proposed by Patra et al. (2019).

Given two metric spaces ($\mathcal{X}$, $d_{\mathcal{X}}$) and ($\mathcal{Y}$, $d_{\mathcal{Y}}$), we start with the Hausdorff distance, defined as:

$$d_{\mathrm{H}}(\mathcal{X}, \mathcal{Y}) = \max \left\{ \sup_{x \in \mathcal{X}} d(x, \mathcal{Y}), \sup_{y \in \mathcal{Y}} d(y, \mathcal{X}) \right\} \tag{2}$$

where $d(a, \mathcal{B}) = \inf_{b \in \mathcal{B}} \| a - b \|_2$ is the distance of point $a$ in $\mathcal{A}$ from set $\mathcal{B}$. Informally, it is the largest distance needed to travel from a point in $\mathcal{A}$ to a point in $\mathcal{B}$.

However, the Hausdorff distance is easily affected by isometric transformations. The GH distance which is the infimum of the Hausdorff distances under all possible isometric transformations is a more robust measure. By contrast, the GH distance reduces the distance over the isometric transforms $f$ and $g$ between $\mathcal{X}$ and $\mathcal{Y}$ as follows:

$$d_{\mathrm{GH}}(\mathcal{X}, \mathcal{Y}) = \inf_{f,g} d_{\mathrm{H}}(f(\mathcal{X}), g(\mathcal{Y})) \tag{3}$$

The computation of Hausdorff distance is NP-hard, and hence we follow Patra et al. (2019) and compute the Bottleneck distances (Chazal et al., 2009) which are considered to be reasonable lower-bounds.

### 3.2 Spectral Graph-based Matching (SGM)

Our second measure is based on the graph-based Eigenvector similarity method. Søgaard et al. (2018) used this similarity to measure the distance between two embedding matrices corresponding to two languages $\ell_1$ and $\ell_2$, via their Laplacian matrices, $\mathcal{L}$. They argue that the Laplacian eigenvalues are good compact representations for the

graph Laplacian, and that their comparison can consequently capture the degree of isomorphism.

Although similar in spirit to their approach, our method to build the underlying graphs ($G^{\ell_1}$ and $G^{\ell_2}$) differs. We use the same idea to model differences between two embedding spaces $\mathcal{X}$ and $\mathcal{Y}$ for the single target language translations from different source languages, as proposed by (Søgaard et al., 2018) but our method to build the underlying graphs ($G^{\ell_1}$ and $G^{\ell_2}$) differs from their approach. While they extract the nearest neighbors by computing the cosine similarity of the cross-lingual word pairs, we take inspiration from the `Isomap` algorithm of Tenenbaum et al. (2000) and build a weighted connected graph over the data points to capture better neighborhood relations. Weights $w_{ij}$ correspond to the distance between points $i$ and $j$ in the input space ($\mathcal{X}$, $d_{\mathcal{X}}(i, j)$). We connect each point only to its $K$ nearest neighbors to consider more geometrical information on the interaction between all vectors within the initial space to improve the graph characterization of the spaces. The value $K$(=6) is chosen to have similar edge density for all graphs. We estimate the geodesic distances between vertices (points) in the input space using shortest-path distances obtained with Dijkstra's algorithm (Dijkstra et al., 1959) on the constructed graph to minimize the sum of the weights of their constituent edges. The subsequent distance matrix represents the basis for our graphs. From this point onwards, the computation of the Laplacian matrices and the final measure $\Delta$ is as in Søgaard et al. (2018), where

$$\Delta = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2 . \quad (4)$$

First for $\mathcal{L}_1$, we find the smallest $k$ in Equation 4 such that the sum of its $k$ largest eigenvalues $\sum_{i=1}^{k} \lambda_{1i}$ is at least 90% of the sum of all its eigenvalues. Similarly, we find another $k$ for $\mathcal{L}_2$, and take the smallest $k$ of these two, such that, $k = min(k_1, k_2)$. The graph similarity metric returns a value in the half-open interval $[0, \infty)$, where values closer to zero indicate better isometry. We compare our SGM metric with the metric of Søgaard et al. (2018) (referred to as EV) in Section 5.

## 4 Experimental Setting

In this section, we provide information on the data, and the vector spaces used for computing deviation of isomorphism. Since we quantify language similarity based on the degree of isomorphism in monolingual spaces, we independently train monolingual word embeddings for the target language and translations into that target language.

### 4.1 Data

We use the same setup as Rabinovich et al. (2017), Bjerva et al. (2019) and Chowdhury et al. (2020), and use the comparable portion of Europarl (Koehn, 2005) with translations from 21 European Union languages into English to minimize the impact of domain difference. The tokens per language vary, ranging from 67 k tokens for Maltese to 7.2 M for German. We refer to the multiple translations into English as $L_j$'s, where $j$=1,2,...,$n$; and to originally written text in English as $L_o$.

We select the subset of translations from 16 languages covering four families: *Romance* (French ($fr$), Italian ($it$), Spanish ($es$), Romanian ($ro$), Portuguese ($pt$)), *Germanic* (Dutch ($nl$), German ($de$), Swedish ($sv$), Danish ($da$)), *Slavic* (Czech ($cs$), Slovak ($sk$), Slovenian ($sl$), Polish ($pl$), Bulgarian ($bg$)) and *Baltic* (Latvian ($lv$) and Lithuanian ($lt$)) into English and English original ($en$) text.

For these 17 datasets, we define two settings, the *full data* condition and the *small data* condition to investigate the effect of data settings on our methods. The former makes use of the complete Europarl edition available for a language (recall that data size differs widely); for the latter, we randomly extract $m$ sentences, where $m$ corresponds to the lower-bound data-size of our translationese data, i.e., the size of the Latvian corpus (118,525 words). We shuffle and randomly subsample $m$ sentences with the same seed for all target-side language data. We report results for the *full data* setting and use the *small data* for robustness checks and comparisons with existing literature.

### 4.2 Vector Spaces

Our data are original English ($L_o$) or translations from language $j$ into English ($L_j$'s). For each data set we induce separate monolingual embeddings in the *full* and *small data* conditions, from their respective tokenised (Koehn et al., 2007) and lowercased data using fastText (Bojanowski et al., 2017).

We train 300 dimensional embeddings with words with more than 5 occurrences in the data. We use skip-gram with negative sampling (Mikolov et al., 2013) with standard hyper-parameters (character $n$-grams of sizes 3 to 6, and a learning rate of

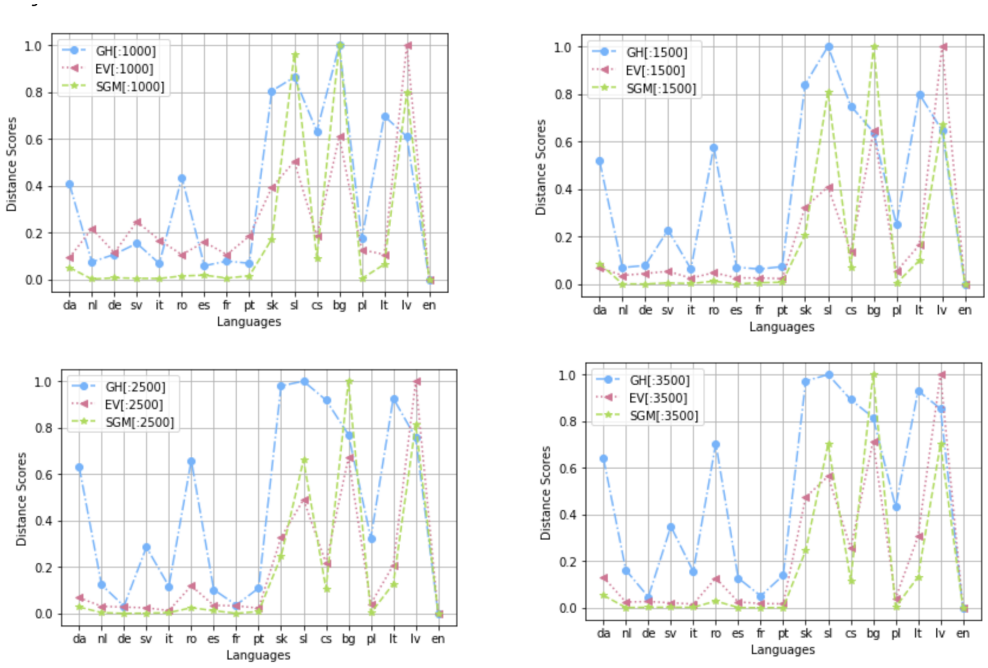



Figure 1: Normalised distances between embedding spaces for original $en$ and translations into $en$ from 16 languages given by our three distance measures using 4 different number of data points.

0.025). Afterwards, embeddings are mean centered and unit normalised. For comparison purposes, we also create vector spaces using the CBOW algorithm and standard hyper-parameters.

## 5 Results and Evaluation

We analyze the behaviour of the different distance measures: the Gromov-Hausdorff distance (GH) and the two Eigenvector similarity-based ones (SGM and EV). We apply them to $(i)$ infer language families, $(ii)$ reconstruct phylogenetic trees, and finally $(iii)$ perform correlation analysis against two benchmarks (SP08 and URIEL).

### 5.1 Language Distance Measures

First, we perform an experiment to determine how distant the vector spaces created from $L_o$ and $L_j$'s are. We compute each metric over the top-$n$ most frequent common words in our data, where $n \in \{1000,1500,2500,3500\}$ to explore the behaviour of the measures with different graph sizes. Notice that having the same number of components (vertices and edges) is a condition for isomorphism.

Results with the normalised distances are displayed in Figure 1. The behaviour for the metrics varies with respect to the number of the most frequent top-$n$ datapoints considered. SGM is the most stable measure across all configurations, showing most variance for 1000 points. EV shows larger variability in distinguishing language similarity, while GH results are relatively stable with respect to larger datapoints (2500 and 3500 points). These variations are due to the different nature of the metrics. GH calculates the distance between the spaces only on the subset of $n$ words, while SGM weights the nearest neighbors of a word which could lie outside the top-$n$ to build the initial graph. Thus SGM considers more *context* for each point and thus needs less datapoints to successfully describe the space. As expected, results with EV and SGM are closer to each other than to GH because they follow a similar methodology, except that SGM retains more context than EV.

We observe that, in all plots, the differences between original English and translations from Germanic, followed by Romance languages are the lowest, indicating that vector spaces of these languages are closer to each other in terms of semantic embedding space based isomorphism measures. However, isomorphism weakens consistently with increased linguistic distances of Baltic and Slavic families irrespective of the method used, providing evidence that language distance in semantic space is higher for etymologically distant language pairs. Additionally, we observe some outliers varying from measure to measure. GH puts $ro$ far from the other Romance languages and $da$ and $sv$ are placed relatively far from other Germanic relatives.

On the other hand, SGM (and to lesser extent EV) locates $pl$ close to $en$.

### 5.2 Reconstructing Language Phylogeny

Figure 2[1] shows evidence that deviations from isomorphism between semantic spaces computed from $L_j$'s into a common target language ($en$) and originally authored text ($L_o$) in $en$ reflect linguistic notions of distance between the source languages and the source language families of the translations. This is evidence for an important aspect of translationese, namely source language interference, in semantic space. Below we further investigate whether the distance in semantic space signal can be used to infer phylogenetic trees.

In our predicted trees, Figure 2, we observe trends that indicate groupings based on morphological or other typological properties. We identify some well known language–language relationships in all three trees showing high similarity between English and other Germanic languages, with some divergences —for example, $sv$ is located far away from its other Germanic counterparts under GH reconstructions and $pl$ is always misplaced into the Germanic-Romance language group, despite its Slavic origin. The influence of geographical factors such as language contact or structural interactions (*Balkan Sprachbund*) can also explain some of the interesting divergences. Overall, the trees exhibit coarse-grained language family contour traces, i.e., Baltic and Slavic languages are close together, while Germanic and Romance form another group in most of the cases.

Our simple embedding-based results provide evidence that translationese is reflected in semantic spaces and that without reliance on fine-grained linguistic knowledge, differences in semantic embeddings space are powerful enough to detect important language differences related to linguistic typology in semantic spaces, corroborating previous results of Chowdhury et al. (2020). This further reconfirms in word embedding based semantic space earlier findings of Rabinovich et al. (2017), Bjerva et al. (2019) which used manual feature engineering or NLMs with a focus on morphologic and syntactic structure.

[1]The clusters are computed over 3500 datapoints for each metric.

### 5.3 Correlation with Typology, Geography and Phylogeny Benchmarks

Above we observed how predicted trees not only show genetic effects, but also other characteristics that might be due to the geographic proximity and not to just to phylogenetic evolution. In this section, we compare our language classification predictions, Figure 2, against linguistic benchmarks. We estimate Kendall correlations between our generated trees and SP08 (representing genetic similarities), and our trees and the averaged URIEL features introduced in Section 2 (representing other rich typological similarities beside genetic ones). The Kendall correlation between the two benchmarks SP08 and the selection of URIEL features is 0.56 reflecting the different nature of the two benchmarks. Although the genealogical distance is common in both, the source for this kind of information is different.

Our results, summarised in the top rows of Table 1, show that correlations between predicted trees and URIEL are higher than with SP08, demonstrating that other factors besides genetics are reflected in the semantic spaces. SGM reproduces the genetic SP08 benchmark better than EV and GH, while GH clearly correlates better with structural URIEL features, followed by SGM and EV. This corroborates NLM-based findings of Bjerva et al. (2019) in our semantic word embedding based spaces: the differences and similarities between languages and language representations go beyond genetic (dis)similarities. Further, we find that while correlations are better under full data conditions, they exhibit a similar behaviour in simulated small-data scenarios, suggesting that our graph-based approaches are effective in a variety of data settings.

## 6 Robustness Analysis

After showing that exploiting departures from isomorphism between spaces can be used to predict relations between languages, we analyze the impact of various modeling assumptions and different training conditions that might have an effect in skewing the results.

### 6.1 Data Size Effects

Large differences in data sizes between high and low-resource languages have played a pivotal role in the performance of monolingual embeddings (Vulić et al., 2020; Sahlgren and Lenci, 2016). In our work, to some extent this is already minimized

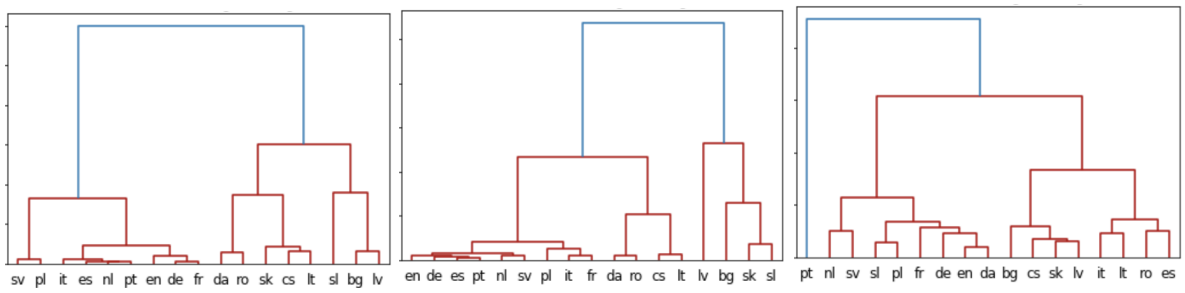



Figure 2: Clustering based on the distance matrix obtained for GH *(left)*, EV *(middle)* and SGM *(right)*.

| # Points | *SP08* | | | *URIEL* | | |
|---|---|---|---|---|---|---|
| | **GH** | **SGM** | **EV** | **GH** | **SGM** | **EV** |
| *Full data condition* | | | | | | |
| 1000 | 0.32 | 0.52 | 0.43 | 0.39 | 0.38 | 0.24 |
| 1500 | 0.40 | 0.30 | 0.26 | 0.51 | 0.31 | 0.36 |
| 2500 | 0.42 | 0.38 | 0.39 | 0.57 | 0.36 | 0.43 |
| 3500 | 0.40 | 0.39 | 0.31 | 0.58 | 0.45 | 0.36 |
| FW | 0.40 | 0.30 | 0.32 | 0.44 | 0.45 | 0.30 |
| Swadesh | 0.30 | 0.32 | 0.11 | 0.36 | 0.43 | 0.09 |
| *Small data condition* | | | | | | |
| 1000 | 0.11 | 0.45 | 0.21 | 0.21 | 0.39 | 0.20 |
| 1500 | 0.29 | 0.37 | 0.21 | 0.49 | 0.27 | 0.21 |
| 2500 | 0.39 | 0.45 | 0.23 | 0.40 | 0.39 | 0.30 |
| 3500 | 0.27 | 0.46 | 0.11 | 0.36 | 0.35 | 0.12 |

Table 1: Mean Kendall correlations of predicted trees with SP08 and average URIEL for various number of datapoints and the function words experiment (FW).

by taking only the most frequent $n$ words to estimate the distances between embedding spaces, but still the quality of even these embeddings might differ. To examine the impact of the data size for our experiments, we use the embeddings obtained under the *small data* condition (see Section 4) and compare the results in the bottom rows of Table 1.

The results show that SGM correlates best with SP08 under all training conditions (number of datapoints and corpus size), but the correlation decreases with respect to URIEL features. GH shows good correlation in some instances (1500 and 2500 datapoints) for both SP08 and URIEL, while EV, shows no consistent correlation. For EV, we consider frequent words and mutual nearest neighbors, thus in the *small data* condition, it has even less access to contexts. Our spectral graph-based measure SGM, which is inspired by the ideas of node representation in contemporary geometric and manifold learning (Cayton, 2005), provides more intuitive understanding of linguistic distances than what is offered in Chowdhury et al. (2020) under varied data settings.

### 6.2 Word Embedding Effects

Köhn (2015) showed that different methods to obtain word embeddings (CCA, skip-gram, CBOW, GloVe, etc.) behave similarly when capturing syntactic and morphological information. We check that this is also the case with our distance methods by comparing the performance obtained with skip-gram and CBOW architectures, and observe small variations with similar global trends. To give an example, correlation results for SGM under the full data condition with CBOW and skip-gram vary only in the $\pm 0.05$ range.

We also performed experiments with lower dimensions (50,100,200) which may lead to reduced expressivity, but, very interestingly, we obtained similar performance as we did with 300 dimensions. For example, on 100-dimensional monolingual word embeddings, the differences are: GH ($\pm 0.077$), SGM ($\pm 0.013$), and EV ($\pm 0.088$).

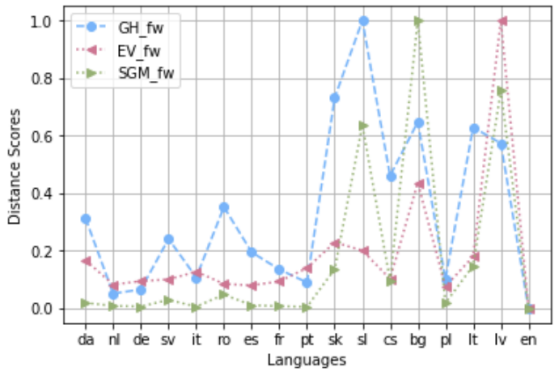
## 7 Comparison with Previous Approaches

### 7.1 Leaf-Node Tree Distances

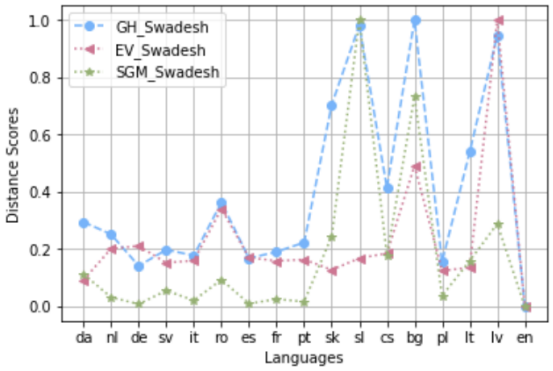
In order to compare our results with previous work, we calculate tree distances using the leaf-node distance in Equation 1 previously defined in Rabinovich et al. (2017), and compare with the best results on SP08 in Rabinovich et al. (2017) and Bjerva et al. (2019). We report our results in the small and the large data conditions for 1500 most frequent datapoints obtained with different metrics in Table 2[2]. All distances are normalized to a zero-one scale[3].

[2]Notice that the results of Table 1 and Table 2 cannot be directly compared as Table 2 is computed after summing over all possible pairs of the leaves (languages), while Table 1 shows the association with benchmarks (SP08 and URIEL) keeping only originally authored English as its source. Table 1 follows Chowdhury et al. (2020) to correlate the results with the benchmarks and Table 2 compares the findings to Rabinovich et al. (2017); Bjerva et al. (2019).

[3]Although an overall correlation similarity analysis based on confusion matrices would be more optimal, we perform




(a) Function Words




(b) Swadesh List

Figure 3: Normalised distances between original and translationese *en*. Embedding spaces are created with only functional words from the Europarl data (a) and Swadesh wordlist (b).

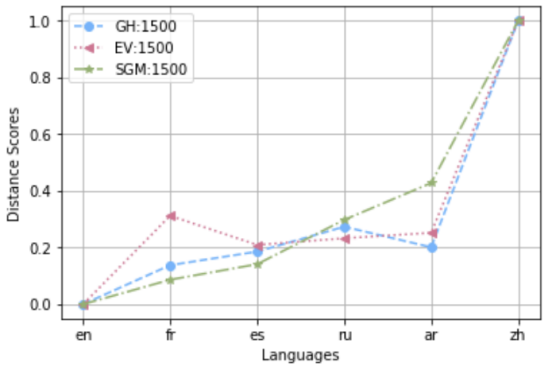



Figure 4: Normalised distances between original and translationese *en*. Embedding spaces are created with UN parallel corpus (Tolochinsky et al., 2018).

| | **Rabinovich** *et al.* 2017 | **Bjerva** *et al.* 2019 | **Our proposal** Full | Small |
|---|---|---|---|---|
| Words | – | 0.53 | | |
| FW | 0.43 | – | – | – |
| POS | 0.35 | 0.52 | – | – |
| FW+POS | 0.36 | 0.56 | – | – |
| PS | – | 0.36 | – | – |
| DepRel | – | 0.32 | – | – |
| GH | – | – | 0.37 | 0.38 |
| EV | – | – | 0.57 | 0.56 |
| SGM | – | – | 0.54 | 0.58 |

Table 2: Mean distance between SP08 and reconstructed phylogenetic trees as compared to previous literature using words, function words (FW), parts of speech (POS), phrase structures (PS) and dependency relations (DepRel) as features.

According to the mean distance, our simple embedding and graph-based approaches, especially GH can reproduce genetic trees on a par with previous work without requiring any explicit linguistic information.

Unlike previous methods which rely on surface-level features of the source language, our graph-based isomorphism analysis is unsupervised and still is able to detect important language differences related to linguistic distances. Of all the methods, GH is the closest to SP08, followed by SGM and EV in the *full data* settings while the trend for SGM-EV is reversed under the *small data* condition.

### 7.2 Function and Content Words

To control topical skew, we investigate whether our approaches of *departure from isomorphism* works on non-lexical representations. To this end, we first focus on function words which introduce and identify key discourse referents and represent relationships between entities but are considered to be not well-modeled by distributional semantics (Bernardi et al., 2015). We use the list of function words defined in Koppel and Ordan (2011) to construct the language distance measure of Section 5.1 in Figure 3$(a)$. In this case, the number of data points is 468, well below the minimum number of points used with content words (1000).

The performance of all three methods show similar trends as in Figure 1. The figure demonstrates that function words are able to capture departures of isomorphism in a similar way as the complete set of words, indicating that source languages carry over grammatical constructs into the translation product, corroborating in simple embedding space prior findings of Rabinovich et al. (2017) and Bjerva et al. (2019) with function words.

Additionally, we explore the much smaller cog-

tree distance analysis because not all underlying confusion matrices of the previous approaches are available.

nate collection of Swadesh word lists (Swadesh, 1952) to capture the relatedness between languages in Table 1 and the language distance computed from their embeddings is shown in Figure 3($b$). As this concept-aligned resource ensures a consistent set of word-lists across all our languages, thereby enhancing comparability, these findings are particularly important. The results in Table 1 show that the large context (6 neighbors) exploited by SGM estimations exceeds other isomorphism methods, while highlighting the limitations of EV in low-data regimes with limited access to contexts.

## 8 Analysis for non-Indo-European Source Languages

Previous research (Rabinovich et al., 2017; Bjerva et al., 2019; Chowdhury et al., 2020) focused on investigating translationese and source language interference for European language families. Here we extend this work, for the first time, to the best of our knowledge, to translations from non-Indo-European languages into English. We explore the language distance measures of Section 5.1 on the UN corpus (Tolochinsky et al., 2018) which consists of translations covering typologically different languages such as Arabic ($ar$) and Chinese ($zh$), as well as Indo-European languages (i.e., Russian ($ru$), Spanish ($es$) and French ($fr$)). The embedding spaces are created in the same manner as for Europarl dataset.

We show results with 1500 words[4] in Figure 4 and observe the following trends: compared with translations from $ar$ and $zh$, the difference between original English and translations from $fr$ and $es$ tend to be smaller, the distance to $zh$ is the largest, and that within the European language family distances is mostly $fr < es < ru$. This is in line with our previous results in Figure 1 on the same-domain monolingual Europarl data under different data settings. However, despite these general trends, GH and EV measured distance scores are similar for $es$, $ar$ and $ru$, while EV has $fr$ more distant to $en$ than $es$, $ru$ and $ar$. This is something that would need to be further explored in future work. Of all measures, SGM, which captures more context from the interaction between data-points and their neighborhoods, accords best with linguistic expectations about language (dis)similarities.

[4]We have comparable trends in all of our configurations with varying number of points.

## 9 Conclusion

In this paper we contribute to the ongoing line of research in computational typology, exploring the potential of translations into a single target language that retains the traces of the source languages to reflect the distances between them. Specifically, we propose an alternative graph-based distance measure to explore (dis)similarities between languages. Our results show that simple graph- and embedding-based distance based methods perform on a par with the best results achieved by previous approaches based on linguistic features in detecting source language interference in translations. We compare Gromov-Hausdorff and our novel Spectral Graph based approach with the original Eigenvector-based divergence from isomorphism measure (EV) against URIEL and SP08, show that our alternative graph isomorphism measures outperform EV and, for the first time, expand translationese research to non-Indo-European languages. We perform robustness tests to verify that our methods are stable under a variety of modeling conditions.

In future work, we aim to leverage our estimated similarities to better explain transfer behavior (local information spreading from one language to the other) in downstream applications such as machine translation.

## Acknowledgments

We would like to thank the reviewers for their insightful comments and feedback. This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding.

## References

Raffaella Bernardi, Gemma Boleda, Raquel Fernández, and Denis Paperno. 2015. Distributional semantics in use. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 95–101.

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.

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.

Lawrence Cayton. 2005. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1.

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. Wiley Online Library.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2020. Understanding translationese in multi-view embedding spaces. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6056–6062.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Michael Cysouw. 2013. Chapter predicting language-learning difficulty. In *Approaches to measuring linguistic differences*. De Gruyter.

Edsger W Dijkstra et al. 1959. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Matthew S Dryer. 2009. Problems testing typological correlations with the online WALS. *Linguistic Typology*, 13(1):121–135.

Isidore Dyen, Joseph B Kruskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical society*, 82(5):iii–132.

Steffen Eger, Armin Hoenen, and Alexander Mehler. 2016. Language classification from bilingual word embedding graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3507–3518, Osaka, Japan.

Benjamin W Fortson IV. 2011. *Indo-European language and culture: An introduction*, volume 30. John Wiley & Sons.

Junxian He, Zhisong Zhang, Taylor Berg-Kirkpatrick, and Graham Neubig. 2019. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3211–3223. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1318–1326. Association for Computational Linguistics.

M Paul Lewis, GF Simons, and CD Fennig. 2015. Ethnologue: Languages of the world. *Dallas, Texas: SIL International*.

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

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Steven Moran, Daniel McCloy, and Richard Wright. 2014. PHOIBLE online.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980, Austin, Texas. Association for Computational Linguistics.

Maurizio Serva and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Bill Thompson, Sean Roberts, and Gary Lupyan. 2018. Quantifying semantic similarity across languages. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*.

Elad Tolochinsky, Ohad Mosafi, Ella Rabinovich, and Shuly Wintner. 2018. The un parallel corpus annotated for translation direction. *arXiv preprint arXiv:1805.07697*.

Gideon Toury. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.

Robert Lawrence Trask. 2000. *The dictionary of historical and comparative linguistics*. Psychology Press.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.