When Elote, Choclo and Mazorca are *not* the Same. Isomorphism-Based Perspective to the Spanish Varieties Divergences

Cristina España-Bonet[1] and **Ankur Bhatt**[1,2,3] and
Koel Dutta Chowdhury[2] and **Alberto Barrón-Cedeño**[4]
[1]DFKI GmbH, [2]Saarland University, Saarland Informatics Campus, Germany
[3]Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Germany
[4]Universitá di Bologna, Forlì, Italy
{cristinae, ankur.bhatt}@dfki.de,
koel.duttachowdhury@uni-saarland.de, a.barron@unibo.it

Abstract

Spanish is an official language in 20 countries; in 19 of them, it arrived by means of overseas colonisation. Its close contact with several coexisting languages and the rich regional and cultural diversity has produced varieties that divert from each other. We study these divergences with a data-based approach and according to their qualitative and quantitative effects in word embeddings. We generate embeddings for Spanish in 24 countries and examine the topology of the spaces. Due to the similarities between varieties —in contrast to what happens to different languages in bilingual topological studies— we first scrutinise the behaviour of three isomorphism measures in (quasi-)isomorphic settings: relational similarity, Eigenvalue similarity, and Gromov-Hausdorff distance. We then use the most trustworthy measure to quantify the divergences among varieties. Finally, we use the departures from isomorphism to build relational trees for the Spanish varieties by hierarchical clustering, and observe that *voseo* is the phenomenon that leaves the strongest imprint in the embeddings.

1 Introduction

Language is a reflection of the needs and behaviours of the community that uses and continually transforms it. One language spoken by diverse communities and/or in various regions can exhibit different characteristics. Spanish is a prototypical example: it lies only behind Chinese in terms of number of native speakers (Eberhard et al., 2023) and, different from it, it is a global overseas language (Ammon, 2010) spoken across c. 11.7 M km^2 by people with diverse cultures and needs.

Originating in the Iberian peninsula as a dialect of Latin, Spanish spread throughout America as a consequence of colonisation. The contact with the indigenous languages present in America in the 16th century, subsequent immigration fluxes, diverse language policies, and societal differences have created a wide variety of *Spanishes*. Figure 1 shows the linguistic zones. Some of these factors operate at the country level (e.g., language policies), but most of them operate at the regional level, where a region may be part of a single country or span across several countries. Consequently, political borders do not uniquely define the varieties.

We study the Spanish varieties using data-based approaches. Since large amounts of textual data for Spanish are only available with, at best, country of origin identifiers, one of our goals is to investigate whether natural language processing (NLP) techniques allow to derive relations among countries and varieties from them. Although the varieties are intrinsically different, divergences among them are less prominent than divergences among languages (e.g., Spanish from Mexico and Spanish from Spain are more similar than Spanish and Portuguese). Because of this, methods in NLP that are adequate and work well in multilingual settings might not properly work for language varieties.

For the study, we create per-country word embeddings and examine the topology of the embedding spaces and their relations using isomorphism metrics, which measure distances between embedding spaces and, in our case, between language varieties. We question whether these measures, used mostly in bilingual scenarios, could be adequate in monolingual settings. We widely explore their performance in controlled quasi-isomorphic scenarios (being our conclusions also relevant for bilingual scenarios) and then use the most reliable configurations to measure distances among our embedding spaces and to derive relational trees. Finally, we interpret the Spanish data-based tree in terms of linguistic characteristics. The work aims at two interrelated goals: (*i*) stressing and evaluating isomorphism measures when applied to language variation and (*ii*) studying Spanish varieties in a new data-based approach to gain linguistic insights. Data and models are available on the project website.[1]

[1] https://cereal-es.github.io/CEREAL/

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 56–77

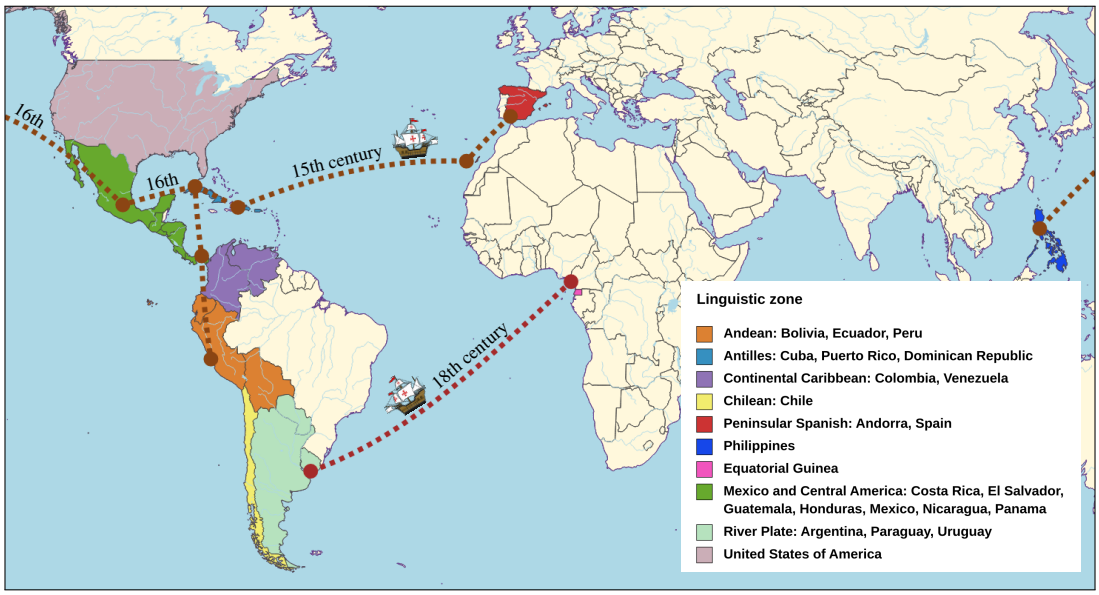


Figure 1: Common geographic Spanish linguistic zones as described by the *Real Academia Española*. The mapping between these linguistic zones and the Spanish varieties might not be one-to-one (Lipski, 2012).

2 The Origins of the Spanish Varieties

Spanish was first derived from Latin in contact with pre-Roman languages in the Iberian peninsula. Different aspects caused proto-Spanish to move away from other Romance languages to be (e.g., Catalan, Portuguese). Two come from long-term rulings during which no language imposition existed. During the Visigothic ruling, multiple words of Germanic origin were introduced, such as `guerra` (war), `riqueza` (wealth) and `yelmo` (helmet). The Arab–Berber control of (up to) two-thirds of the Iberian peninsula, from the 8th until the 15th century, imported novel knowledge and new artifacts, resulting in the introduction of more than $4k$ Arabisms to the lexicon (Alatorre, 1989, p. 74, 80).

The fall of the last remains of Arabic and Mozarabic language varieties came from the 13th until the end of the 15th century, a period during which the kingdom of Castille influenced the speech of neighbouring kingdoms, such as Leon and Aragon (Penny, 2002, p. 19). Internally, two Spanish norms contended: the one of Toledo (later on Madrid) and the one of Seville.

The path of overseas Spanish expansion started with the conquest of the Canary Islands, from which Columbus departed to arrive in the Caribbean in the late 15th century, unfolding in the conquest of Hispaniola (current Dominican R./Haiti) and Cuba. It is from Cuba that the conquest of both Mexico and Peru was launched, as depicted in Figure 1. The norm exported to the new territories was guided by the origin of the migrant population (e.g., priests, soldiers, settlers). Since almost 50% left from Andalusia and Extremadura (López Morales, 1998), these regions hold Spanish varieties derived from the Seville norm. As a result, they share phenomena such as *seseo*,[2] aspirate `/h/`, and an absence of formal/informal differentiation for the second person plural: only `ustedes` is used in America while both `ustedes` (formal) and `vosotros` (informal) are used in Spain (Penny, 2002, pp. 22–23).

It is assumed that the well-established connections across some regions kept the varieties of Spanish in Mexico and Peru closer to those from Spain. A weaker influence on more remote or difficult to reach regions (e.g., Argentina, Paraguay, Uruguay, Central America) allowed for the organic development of farther varieties. The most accepted theory (Penny, 2002, p. 25) is that the influence was driven by centres of power and strength of communication. Consequently, most of Mexico, Peru, Bolivia, and Andean Ecuador share the retention of syllable-final `/s/`, influenced by central Peninsular settlers, whereas other areas miss it, influenced by Southern Peninsular and Canarian Spanish. The pronoun `tú` (you), as used in Spain, is predominant in Mexico, the Caribbean, most of Bolivia and Peru, and part of Venezuela, whereas `vos` competes with `tú` in more remote areas (e.g., Chile, Ecuador,

[2] Phenomenon in which c and s share the phoneme `/s/`.

Colombia) and is predominant in others (e.g., Argentina, Uruguay, Paraguay). The phenomenon is called *voseo* and implies a change in the verbal forms used after the pronoun (Benavides, 2003).

Another relevant factor in the development of the varieties is the long-term influence of coexisting languages. During the whole period of Spanish rule, the native language of most of the population in America was prehispanic, be it of American or of African origin (López Morales, 1998, p. 3). As a result, Spanish in different regions incorporated large vocabularies from other languages. Also, phonetics was affected. One of the most distinctive features of (South) Argentinian and Uruguayan speakers is *zheísmo* (Staggs, 2019), the pronunciation of both `y` and `ll` as `[ʒ]` due to the influence of the local Amerindian languages. Nowadays, the sound has drifted in some regions to `[ʃ]` being called *sheísmo*.

Since independence, Spanish varieties in American countries have been significantly influenced by imported languages. During the 19th and 20th centuries, heavy immigration from Italy to Argentina (Cuadrado Rey, 2020) introduced lexical borrowings from Italian (e.g., `gamba` to refer to leg), Neapolitan and others (Bihan, 2011). Due to geographical vicinity and global influence, two foreign languages were the most influential in this same period: French in Spain and English elsewhere (e.g., `computadora` from English is used in most of Latin America vs `ordenador`, from the French `ordinateur`, in Spain).

Due to these intricacies, there is no straight line to draw on a map to separate the Spanish varieties. The varieties might form a continuum and scholars suggest different categorisations. The closest to our purposes is based on the lexicon; Henríquez Ureña (1921) distinguishes the varieties according to the indigenous language substrate: Nahuatl, Caribbean languages, Quechua, Mapudungun/Araucanian, and Guarani. Lipski (2012) defines 21 varieties (11 for Spain, 10 for America), and Soler Montes (2015) defines 8 (3 for Spain, 5 for America). In the former, the classification is based mainly upon phonetic, lexical, and morphosyntactic features; in the latter on geolinguistics. But these are only two examples. Sippola (2021) summarises 4 classifications taking into account geographical variations mostly including phonetic features. Still, the geographical variations are not aligned with geopolitical borders and this has an impact on data-based approaches. Even though few resources exist for the linguistically motivated varieties with city of origin indications (Robelo, 1904; Prieto and Roseano, 2013; Albelda Marco and Estellés, sd), large textual corpora in Spanish are, at best, tagged only with country of origin (Gonçalves and Sánchez, 2014; Tellez et al., 2023; RAE, 2024; España-Bonet and Barrón-Cedeño, 2024).

3 Isomorphism in NLP

Early empirical results using bilingual dictionaries (Youn et al., 2016) and vector embeddings calculated on textual corpora (Mikolov et al., 2013) show that concepts in natural language are structured in a similar way across languages. Vector embeddings in different languages appear to be isomorphic —or at least geometrically similar (Marchisio, 2023). However, other studies show that isomorphism does not always hold, and the more distant a pair of languages or the domain is, the weaker the isomorphism (Søgaard et al., 2018; Patra et al., 2019; Marchisio et al., 2020). But language and domain are not the only factors, differences in training corpus size, training time or the algorithm used to compute the embeddings have a significant effect too (Vulić et al., 2020; Marchisio et al., 2020).

Isomorphism metrics have been introduced in the context of bilingual lexical induction (BLI) where most of the previous conclusions have been drawn. In this context, metrics are used to quantify the similarity (or distance) between embedding spaces of different languages and to observe how they correlate with BLI accuracy.

Several metrics deal with word embeddings from different points of view. Isospectral metrics treat embedding spaces as graphs in the context of spectral graph theory: with respect to the spectral characteristics (e.g., eigenvalues and eigenvectors) of the matrix structures (e.g., adjacency and Laplacian matrices) that represent an embedding space. The Eigenvector similarity distance (Søgaard et al., 2018), the effective condition number (Dubossarsky et al., 2020) and the Spectral Graph-based Matching distance (Dutta Chowdhury et al., 2021) are examples. Isometric measures treat word embeddings as coordinates in a metric space. Earth Mover's distance is a measure of the closeness between the distribution of two sets of words (Zhang et al., 2017) and Relational Similarity is the Pearson's correlation between their cosine similarities (Vulić et al., 2020). The Gromov-Hausdorff distance scores the largest distance between a word from one space and the nearest neigh-

bours from the other space after an isometric transformation between the spaces (Patra et al., 2019).

The mathematical definition of isomorphism, in which two structures are either isomorphic or not, is an approximation in NLP. In NLP, one deals with *degrees of isomorphism* between representative substructures instead. Due to the large vocabularies, and the richness and nuances of natural language, embedding spaces are usually represented by a subgraph/subset formed by up to 5–10 k words. The number of words and which words are used is an ad-hoc decision.

Going beyond the correlation between the isomorphism scores and BLI, the previous metrics have been used to quantify the isomorphism between embedding spaces. When multiple metrics are used, it becomes evident that they do not correlate with each other (Dubossarsky et al., 2020; Dutta Chowdhury et al., 2021; España-Bonet and Barrón-Cedeño, 2022; Marchisio et al., 2022).

In this work, we analyse relations among 24 varieties of Spanish using isomorphism metrics. We expect differences across varieties of the same language to be much smaller than across different languages. Therefore, we first calibrate the isomorphism measures in isomorphic settings —same language, same training data, same embedding algorithm, and hyperparameters (Section 6). As these metrics do not correlate, this allows us to determine the best metric and configuration (number and selection of words) to perform the fine-grained analysis among varieties (Section 7).

4 Isomorphism Measures

We select three measures that capture the isomorphic/isometric degree between two embedding spaces E_1 and E_2 represented by nearest-neighbour graphs G_1 and G_2 and sets of points S_1 and S_2. We assume that the embeddings E_1 and E_2 are mean-centred and length-normalised.

Relational similarity (RS) (Vulić et al., 2020). One can presume that the similarity between words is distributed similarly in different spaces and, so, the cosine similarity of aligned words should be similar in both spaces. RS uses a list with k words from E_1 aligned to k words from E_2 (a dictionary) and calculates the cosine similarities between all the pairs of words in E_1 and E_2 independently:

$$\begin{aligned} &\mathrm{sim}_{E_1}(S^p, S^r) \;\; \forall S^p, S^r, p \neq r \in \mathrm{list}(E_1) \\ &\mathrm{sim}_{E_2}(S^p, S^r) \;\; \forall S^p, S^r, p \neq r \in \mathrm{list}(E_2) \end{aligned} \quad (1)$$

RS is the Pearson correlation ρ between the sorted lists of similarities resulting from the spaces:

$$\mathrm{RS} = \rho\left(\mathrm{sim}_{E_1}^{\mathrm{sorted}}, \mathrm{sim}_{E_2}^{\mathrm{sorted}}\right). \quad (2)$$

Eigenvector similarity (EV) distance (Søgaard et al., 2018). A total of k words in E_i are used to construct n-nearest neighbour unweighted graphs G_i. The nearest neighbours are extracted by computing the cosine similarity between the k words in E_i and all the words in E_j. Given G_i, EV estimates the degree of isomorphism from the eigenvalues of the Laplacian of G_1 and G_2. Let the Laplacian be

$$L_i = D_i - A_i(G_i), \quad (3)$$

where A_i is the adjacency matrix of G_i, and D_i is the diagonal matrix of degrees. After computing the Laplacian eigenvalues, following Søgaard et al. (2018), one finds the smallest m such that the sum of the m largest Laplacian eigenvalues is <90% of the total. Using the smallest m of E_1 and E_2, EV is defined as

$$\mathrm{EV} = \sum_{j=1}^{m} \left(\lambda_{1j} - \lambda_{2j}\right)^2, \quad (4)$$

where λ_{ij} are the top j eigenvalues of L_i.

Gromov-Hausdorff (GH) and Bottleneck distances (Patra et al., 2019). GH is an isometric measure that treats word embeddings as coordinates in a metric space. It gives the worst-case distance (E_1 vs E_2) of nearest neighbours in a shared embedding space after an optimal isometric transformation.

For each word x in S_i, one finds its nearest neighbour y in S_j (NN_j). The Hausdorff distance H is the largest of the two distances:

$$\mathrm{H} = \max\left(\mathrm{dist}(x_1, \mathrm{NN}_2), \mathrm{dist}(x_2, \mathrm{NN}_1)\right). \quad (5)$$

The Gromov-Hausdorff distance is the infimum of the Hausdorff distances under all possible orthogonal transformations. Since computing GH is an NP-hard problem, the Bottleneck distance B, bounded by GH, is used as an approximation (Chazal et al., 2009). B is the shortest distance for which there exists a perfect matching between the points p and r of the persistent diagrams[3] built from S_1 and S_2:

$$\mathrm{B} = \inf_{matches} \max_{(p,r)} ||p - r||_\infty. \quad (6)$$

[3]A persistent diagram is a set of points in $\mathbb{R}^2$ in the half-plane above the diagonal.

Country & Code		CEREAL Segments	CEREAL Words	CEREAL Vocab.	Twitter Vocab.
Andorra	ad	13,023	543,047	2,671	–
Argentina	ar	20,950,705	986,413,066	284,191	673,424
Bolivia	bo	975,429	49,518,821	53,799	47,012
Chile	cl	12,079,476	548,257,312	199,493	282,737
Colombia	co	8,323,794	375,326,751	163,212	324,635
Costa Rica	cr	825,513	37,760,657	45,893	103,086
Cuba	cu	1,919,998	93,368,177	82,275	18,682
Dominican R.	do	1,183,336	48,726,587	52,409	108,655
Ecuador	ec	1,624,269	66,662,454	64,312	147,560
Spain	es	20,950,705	880,495,659	596,842	571,196
Eq. Guinea	gq	4,050	329,469	1,698	1,167
Guatemala	gt	561,714	23,421,191	35,860	95,252
Honduras	hn	656,212	24,971,660	35,707	60,580
Mexico	mx	20,875,244	912,645,564	250,313	438,136
Nicaragua	ni	405,935	18,921,537	31,345	68,605
Panama	pa	448,974	18,431,387	31,268	111,635
Peru	pe	5,066,369	213,937,404	122,884	178,113
Philippines	ph	1,382	75,761	405	–
Puerto Rico	pr	128,103	5,619,179	15,062	23,062
Paraguay	py	775,101	33,771,401	46,513	124,162
El Salvador	sv	401,348	17,068,212	29,433	73,833
USA	us	376,839	21,335,770	34,368	292,465
Uruguay	uy	1,804,329	85,809,183	75,491	200,032
Venezuela	ve	1,201,624	55,514,289	59,334	271,924

Table 1: Number of segments and words used to compute the variety-specific word embeddings.

Data Points (Word) Selection In all the measures above, we characterise each embedding space by k words ($k \in \{100, 500, 1000, 2500, 5000\}$) following 5 criteria:

• **Most frequent words (Frequent, MFW).** We use the top-k words in an embedding space ranked by frequency. This is the standard choice for EV and GH in previous work.

• **Random words (Random).** We randomly select k words within the top half of the frequency-ranked embeddings.

• **Aligned random words (Random BiDict).** As in Random, but only words that appear simultaneously in the two spaces are considered. This is equivalent to using a bilingual lexicon in the general case, which is the standard choice for RS.

• **Numbers.** Random k numbers appearing simultaneously in the two spaces.

• **Named Entities (NEs).** Random k NEs appearing simultaneously in the two spaces. The list of NEs contains 3,416 single words extracted from the CoNLL-2002 shared task (Tjong Kim Sang, 2002).

We adapt the implementation for RS, EV, and GH in Vulić et al. (2020)[4] to consider our lists.

[4] https://github.com/cambridgeltl/iso-study

5 Variety-Specific Embedding Spaces

We use the CEREAL corpus (España-Bonet and Barrón-Cedeño, 2024) to obtain embedding spaces for 24 varieties of Spanish. CEREAL contains documents in Spanish extracted from OSCAR (Open Super-large Crawled Aggregated coRpus version 22.01 (Ortiz Suárez et al., 2019; Abadji et al., 2021) and annotated with the country of origin. We use the documents in CEREAL where the country of publication is codified in the URL and discard documents whose country was inferred automatically (CEREALex). To compute the embeddings, we eliminate sentences having only punctuation and numbers, as well as those with at least one Arabic, Chinese, Cyrillic or Greek character. We then normalise and tokenise the texts using Moses' scripts (Koehn et al., 2007) and lowercase. Table 1 shows the statistics of the final dataset together with the code we use to identify each variety. We estimate fastText (Bojanowski et al., 2017) embeddings using the default skip-gram configuration to train 300-dimensional embeddings for tokens appearing at least 20 times.

The amount of text in Spanish from Spain in CEREAL is significantly larger than for the other varieties (70.5 M segments for es vs 20.9 M for ar, the second largest). For comparability reasons, we use a subset of 20 M segments to train es embeddings. With this, ar, es, and mx have a similar amount of training data, whereas ad, gq, ph, pr, sv and us have less than 0.5 M segments and are discarded for our high-resourced experiments.

Our calibration experiments (Section 6) are done with Spanish from Spain embeddings. We create 10 models from CEREAL: 5 models using 5 seeds for fastText on a fixed subset of 20 M segments (model perturbation) and 5 models using 5 random extractions of 20 M segments over the whole 70.5 M segments (data perturbation).

Our exploration experiments (Section 7) consider embeddings for the 24 varieties. We generate 3 embedding models per variety with different seeds and show the mean in our results. In this case, we also use existing Twitter embeddings (currently X) for 22 varieties (Tellez et al., 2023).[5] Since the training corpus is not available, we use their pretrained embeddings with a single run. Otherwise, our setting with CEREAL is comparable to theirs except for the minimum frequency of in-vocabulary

[5] https://ingeotec.github.io/regional-spanish-models

tokens (the default being 5 in their case) and the fact that they remove diacritics from the data.

6 Isomorphism Measures Calibration

If isomorphism metrics are a good measure to account for distances among languages, they should drop to zero when computing the distance between embeddings of a single language —or approach 1 when they imply correlations. As described in Section 3, differences in size and domain of the training data and in the algorithm used to train the embeddings affect their performance. In this section, we isolate all these factors and evaluate the metrics in an isomorphic setting: same variety (es), same corpus (CEREAL) and same algorithm (skip-gram). We perturb the basic setting by (*i*) applying several initialisations for training the embeddings while keeping skip-gram and all its parameters constant (model perturbation) and (*ii*) subsampling the training data from a larger dataset (data perturbation). With these variations, we aim to study the robustness of the metrics to minor changes and at determining the best configuration for each of them. This study provides insights on the feasibility of using one or more isomorphism metrics to explore relations between language varieties in Section 7.

Model Perturbation We use 5 embedding models for es trained with different seeds on the same data (i.e. the vocabulary is the same for all 5). We calculate RS, EV and GH for 10 combinations of embeddings with the 25 possible configurations of Section 4. Detailed results for the pairwise combinations are in Appendix A and the mean over the 10 combinations is in Table 2.

For all three metrics, there are definite trends when the mean is considered, but the trends are less evident when looking at individual embedding pairs. The variations, due only to different runs, are significant. Frequent and Random BiDict perform the best; i.e. distances EV and GH are the smallest and correlation RS the highest. In this setting, the most frequent words in both es spaces are the same and therefore behave similarly to a dictionary —this does not need to happen in the data perturbation setting and even less in the general multilingual setting. As expected, random words unrelated across spaces perform the worst. Also, numbers and NEs do not perform well (except for RS with numbers). This might be related to the fact that they cluster in a specific region of the space and cannot represent the topology of the whole. In terms of the number of datapoints, EV performs best with few, GH with a large set and RS in this setting does not seem to be sensitive to the volume.

	Model Perturbation			Data Perturbation		
	RS↑	EV↓	GH↓	RS↑	EV↓	GH↓
Frequent						
100	0.989±0.000	2±1	0.02±0.00	0.860±0.056	2±1	0.03±0.01
500	0.982±0.001	2±1	0.02±0.00	0.314±0.032	3±1	0.02±0.00
1000	0.979±0.002	3±1	0.02±0.00	0.131±0.014	2±1	0.02±0.00
2500	0.976±0.002	3±1	0.01±0.00	0.038±0.000	4±1	0.01±0.00
5000	0.974±0.003	5±2	0.01±0.00	0.015±0.001	5±4	0.01±0.00
Random						
100	0.000±0.008	3±1	0.17±0.12	0.002±0.015	5±1	0.15±0.06
500	0.000±0.002	5±3	0.15±0.06	0.000±0.001	5±2	0.19±0.10
1000	0.000±0.000	6±2	0.16±0.07	0.000±0.000	5±1	0.13±0.05
2500	0.000±0.000	10±3	0.11±0.03	0.000±0.000	7±1	0.13±0.03
5000	0.000±0.000	14±7	0.07±0.01	0.000±0.000	14±5	0.12±0.04
Random BiDict						
100	0.959±0.002	1±1	0.03±0.01	0.884±0.008	3±2	0.05±0.02
500	0.959±0.002	3±1	0.02±0.00	0.882±0.004	4±1	0.03±0.01
1000	0.959±0.002	4±1	0.02±0.00	0.883±0.002	6±4	0.03±0.01
2500	0.960±0.002	7±3	0.02±0.00	0.883±0.001	7±2	0.03±0.01
5000	0.960±0.002	5±2	0.01±0.00	0.883±0.000	8±4	0.02±0.00
Numbers						
100	0.997±0.000	3±1	0.05±0.05	0.604±0.087	3±1	0.06±0.04
500	0.994±0.000	4±1	0.02±0.00	0.116±0.012	5±2	0.05±0.00
1000	0.993±0.001	5±2	0.02±0.00	0.061±0.008	9±6	0.02±0.00
2500	0.988±0.001	9±5	0.02±0.00	0.037±0.007	12±4	0.02±0.00
5000	0.985±0.001	7±1	0.02±0.00	0.022±0.003	11±3	0.05±0.01
NEs						
100	-0.003±0.013	3±1	0.10±0.03	-0.001±0.018	2±1	0.08±0.02
500	0.002±0.007	7±3	0.07±0.03	-0.003±0.007	3±1	0.07±0.03
1000	0.002±0.006	6±3	0.09±0.03	-0.002±0.004	7±2	0.07±0.02
2500	0.000±0.003	7±3	0.03±0.00	-0.002±0.002	7±2	0.05±0.02
5000	–	–	–	–	–	–

Table 2: Mean and standard deviation ($\mu \pm \sigma$) score for the three isomorphism metrics used in this study. Perfect isomorphism implies RS 1, and EV and GH 0.

Data Perturbation We use 5 embedding models for es trained with different random subsets of the same corpus (i.e. the vocabulary of the models is different). As before, we calculate RS, EV, and GH for a total of the 10 combinations of embeddings, and use 5 different types and a number of datapoints.

Contrary to what one could expect, the perturbation of the dataset —within the same corpus— does not bring more variability on the results of the metrics than the perturbation of the model as measured by the standard deviations (Table 2). The trends with respect to the number and types of points are also similar to the previous case; the global scores are slightly worse but compatible within the 1σ CIs for GH and EV; differences are larger for RS. Ideally, a good metric would score a distance of 0 (EV and GH) and correlation 1 (RS); EV achieves this at 2σ level, especially when using the most frequent words. We consider this configuration, EV (MFW

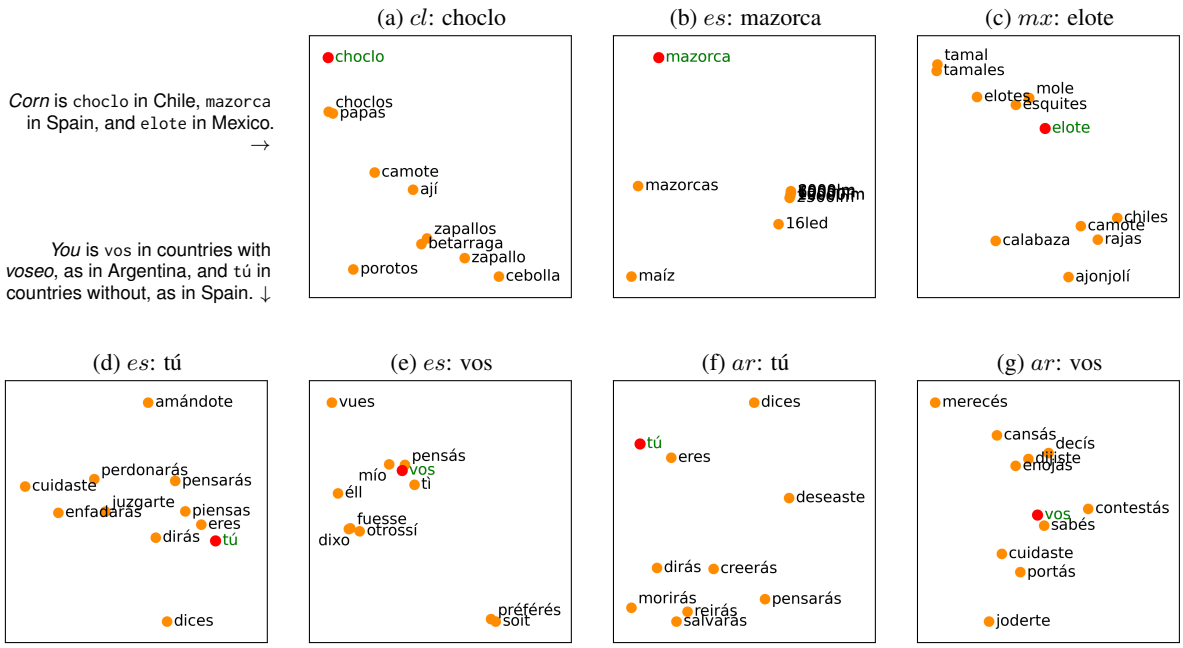


Figure 2: t-SNE projections (van der Maaten and Hinton, 2008) of the neighbourhood for the Spanish words equivalent to *corn* (top plots) and *you* (bottom plots).

100), the best metric to measure isomorphism.

This setting is close to our case of study: language varieties. We extract the data for training the embeddings from the same corpus and use skip-gram with the same configuration. Larger/smaller scores and standard deviations for the isomorphism metrics than the ones we see here should be attributed to language differences and to the quality of the embeddings given by the amount of training data per variety.

7 Spanish Varieties Relations

Qualitative Behaviour Different lexicons and cultural-dependent (near) synonyms change the topology of the embedding spaces. *Corn* in English translates into `elote` (from the Náhuatl `elotitutl`) in most of Mesoamerica, `choclo` (from the Quechua `chuqllu`) in South America, `mazorca` (from the Arabic `masúrqa`) in Colombia, Cuba and Spain and `jojoto` in Venezuela. The importance of this cereal in Spain is irrelevant in comparison to Mesoamerica, where it is so essential that it goes beyond staple food, and that changes the usage of the word.

This is reflected in Figure 2 (top plots), which shows the 10-nearest neighbours for `choclo` in cl, `mazorca` in es, and `elote` in mx. For cl and mx, `choclo` and `elote` are surrounded by other food-related words, but the intersection is almost null. For es, `mazorca` is surrounded mostly by words related to another sense of the word (a kind of light bulb). Appendix B, shows the same three words as located in all three embedding spaces —the three synonyms never appear in the same region of the space and `elote` does not even appear in cl. The surrounding words also vary from being local food when the word is shown in its native embedding space to foreign food when the word is in the embedding space corresponding to another country. Similarly, there are differences across varieties in the verbal forms usage and other grammatical issues, such as *voseo*, which also distort the embedding spaces. As Figure 2 (bottom plots) shows, for countries without *voseo*, such as Spain, the word `vos` is surrounded mostly by non-Spanish words (since it is a nearly-deprecated pronoun for this variety). In Argentina, we observe *verbal voseo*, that is, the usage of the modified 5th inflexion and the 7th verbal inflection instead of the 2nd inflection (e.g., `sabés` or `decís` rather than `sabes` or `dices`).

Isomorphism Following the results of Section 6, we select EV (MFW 100) for the main analysis and include the top-2 performing configurations per metric in the Appendix C as they give more insights on the behaviour of the metrics.

Figure 3 shows the results for EV. The heatmap combines the results with our `CEREAL` embeddings (top–right triangle) and the publicly available Twit-

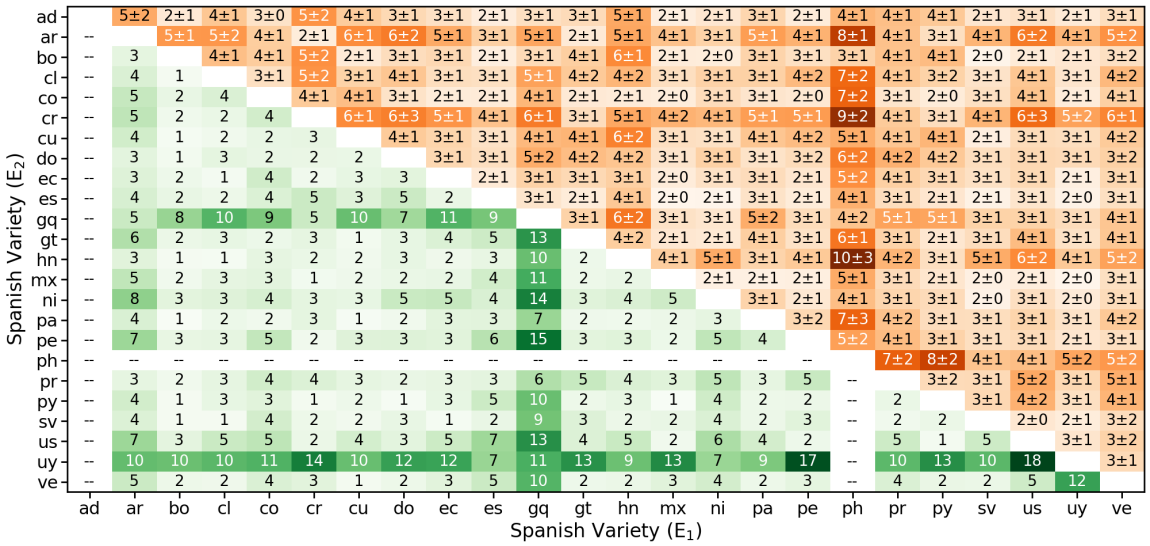


Figure 3: EV with 100 most frequent words for the 24 Spanish varieties. Top–right triangles (orange) correspond to the mean results with the CEREAL corpus and bottom–left triangle (green) to the Twitter corpus.

ter embeddings (bottom–left triangle). With our in-house embeddings, we calculate mean and standard deviation over 9 combinations per language pair, as we have three runs per variety. This is not possible with Twitter, so we expect more robust results with CEREAL.

Comparing the heatmap of EV with those of RS and GH, one sees that both but especially GH are sensitive to the size of the training data. The fact that GH needs to use more datapoints (5,000 MFW in contrast to the 100 for EV/RS) might make the effect of the data size stronger on GH. Varieties *ad*, *gq* and *ph* for CEREAL and *gq* for Twitter include less than 15 k training sentences. Pairs involving these varieties have statistically significant higher GH and lower RS values (and to a lesser extent also higher EV values) systematically for all the pairs. Vulić et al. (2020) showed in their correlation analysis between the isomorphism metrics and BLI accuracy that one needs at least 500 k training sentences to convergence in BLI accuracy. Therefore isomorphism scores with embeddings trained with less data might be suboptimal.

The results with CEREAL and Twitter are significantly different both in the absolute magnitude of the scores for pairs of varieties and in the relations between the varieties. This could be a consequence of the different volumes of training data but also of the differences in the register used in both genres. As Lipski (2012) notes, social factors are also relevant in the variation and both genres might be representative of different population profiles. The standard deviations in Figure 3 are of the same order of magnitude as in our calibration experiments (Section 6) and do not depend on the quality of the embeddings as measured by the data size. Therefore, differences in the scores across language pairs, that is, different departures from isomorphism, are representative of the distances (relations) among varieties. Next, we use hierarchical clustering to have a clearer overview of these relations.

Phylogenetic (Relational) Trees Clustering these results over varieties allows for building a Spanish phylogenetic tree. Notice that, strictly speaking, we do not construct a phylogenetic tree but a relational tree. Spanish was acquired in most American countries almost simultaneously, and varieties have been evolving in parallel since then. Also, word embeddings are static (in time) and they do not provide evolutionary relationships but an average snapshot of the language relationships. Following Dutta Chowdhury et al. (2020), we use agglomerative clustering with variance minimisation (Ward Jr, 1963) for this purpose.

We show the results for all the varieties and the best configuration option for RS and GH in Figure 4 as the comparison among metrics is especially relevant. Appendix D includes the remaining configurations. This representation makes more evident the fact that GH groups the varieties according to the amount of training data —and therefore the quality of the embeddings. On the left-hand side of the GH dendrogram are the least resourced varieties: *ph*, *ad* and *gq*. On the right-hand side are

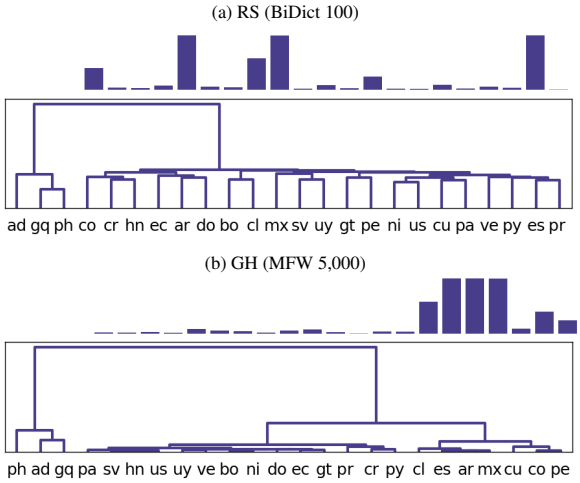


Figure 4: Relational trees derived from the CEREAL embeddings. The distribution of data is shown on top of the dendrograms as an illustration.

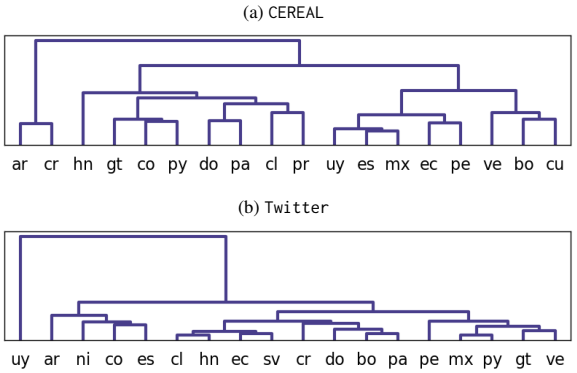


Figure 5: Relational trees for the subset of the highest resourced varieties with EV (MFW 100).

the highest resourced ones: *cl*, *es*, *ar*, *mx*, *cu*, *co* and *pe*. RS also clearly clusters *ad*, *gq* and *ph* and puts all the other varieties at a similar level. The Spearman rank correlation between the number of segments used to train the embeddings and the flattened version of the hierarchical clustering output is 0.8 for GH, 0.2 for RS and -0.1 for EV. The limitations with GH and RS are in agreement with the observations of the previous sections.

In Figure 5, we analyse in detail EV for CEREAL and Twitter for the highest resourced varieties. A visual representation on a map is in Appendix D. None of these trees groups the varieties according to their geographical position or the linguistic zones described by RAE (cf. Figure 1). It is worth pointing out that phonetic differences are in principle not observable with word embeddings on textual data but might leave traces on Twitter embeddings as a result of misspellings. This translates into Argentina (*ar*) and Uruguay (*uy*) —countries where *zheísmo* is present— lying apart in the CEREAL dendrogram, but not in the Twitter one.[6] Trends related to grammar are more evident. The right-hand side of the plots group together varieties without *voseo*: in the case of CEREAL, the exceptions are Uruguay (*uy*) and Dominican Republic (*do*) which should be swapped according to this characteristic; in the case of Twitter, Spain (*es*) sneaks in the region with *voseo*.

Different substrates are in general not observed. Contrary to *voseo* and the grammatical differences it implies, different substrates or neologisms are not global: Quechuan languages were/are spoken in what is today Argentina, Bolivia, Chile, Colombia, Ecuador and Peru; but in Bolivia there are 36 other languages such as Quechua but also Aimara, Chiquitano, etc.

8 Conclusions

Spanish is not a monolithic language. Five centuries of distinct but related evolution across territories have created a rich set of varieties. We study these varieties from a data-based perspective, building specific embeddings with textual data for 24 countries. We then relate the similarities and differences among embedding spaces to the divergences among varieties.

Divergences are subtle in comparison to divergences among languages. Because of this, we explore three common isomorphism metrics in quasi-isomorphic settings. Our results show that EV is the best performing metric in the controlled scenario (data perturbation). GH does not perform far, but subsequent experiments with the variety of embeddings show that it is the metric that depends the most on the amount of training data. RS rapidly degrades when we depart from the controlled experiments and it is less sensitive to the variations.

Lots of characteristics of the language coexist in written documents. The indigenous language substrate and other borrowings, grammatical characteristics such as *voseo*, and verbal tense changes are manifested in word embeddings. *Voseo* showed to be the strongest feature and its imprint is clearly seen in the relational trees we build from the departures from isomorphism obtained with EV. Informal (and sometimes incorrect) text used to create Twitter embeddings also reflects distinctive phonetic traits such as *zheísmo*.

[6] Both countries also share an Italian substrate.

Limitations

We have done an exhaustive exploration of the behaviour of the isomorphism metrics when the same language (Spanish from Spain) is used. The effect of the training domain and data size has been explored before in bilingual settings (Vulić et al., 2020). In this work, we do not systematically quantify the effect that the different sizes in the training data per variety imply, further than removing the varieties with less data according to the conclusions in Vulić et al. (2020). Differences in the amount of data can also imply differences in the domain (especially when few data are available) and these variations have to be taken into account when drawing conclusions.

Acknowledgements

This work has been supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding and by the LT-Bridge Project (GA 952194).

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1–9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Antonio Alatorre. 1989. *Los 1001 años de la lengua española*. El Colegio de México / Fondo de Cultura Económica, Mexico City, Mexico.

Marta Albelda Marco and María Estellés. sd. Corpus ameresco. [online] Last accessed on 13.03.2024.

Ulrich Ammon. 2010. World languages: Trends and futures. In *The Handbook of Language and Globalization*, chapter 4, pages 101–122. John Wiley & Sons, Ltd.

Carlos Benavides. 2003. La distribución del voseo en hispanoamérica. *Hispania*, 86(3):612–623.

Ulysse Le Bihan. 2011. *Italianismos en el habla de la Argentina: herencia de la inmigración italiana*. Ph.D. thesis, Unniversiteter i Oslo.

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.

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.

Analía Cuadrado Rey. 2020. *El italiano en la fraseología actual del español hablado en Argentina*. Edizioni Ca' Foscari.

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2020. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2377–2390, Online. Association for Computational Linguistics.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2021. Tracing source language interference in translation with graph-isomorphism measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 375–385, Held Online. INCOMA Ltd.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*. SIL International, Dallas, TX.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2024. Elote, Choclo and Mazorca: on the Varieties of Spanish. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico. Association for Computational Linguistics.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

Pedro Henríquez Ureña. 1921. Observaciones sobre el español en América. *Revista de filología española*, 8:357.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open

source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

John M. Lipski. 2012. *Geographical and Social Varieties of Spanish: An Overview*, chapter 1. John Wiley & Sons, Ltd.

Humberto López Morales. 1998. La aventura del español en América.

Kelly Marchisio. 2023. *Multilinguality from Static Embedding Spaces: Algorithmic, Geometric, and Data Considerations*. Ph.D. thesis, Johns Hopkins University.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022. Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

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.

Ralph Penny. 2002. *A History of the Spanish Language*. Cambridge University Press.

Pilar Prieto and Paolo Roseano. 2013. Atlas interactivo de la entonación del español. [online] Last accessed on 13.03.2024.

Real Academia Española RAE. 2024. Corpus del Español del Siglo XXI (CORPES). [online] Last accessed on 13.03.2024.

Cecilio A. Robelo. 1904. *Diccionario de aztequismos*.

Eeva Sippola. 2021. *Morphosyntactic Variation in Spanish: Global and American Perspectives*, pages 209–232. Cambridge University Press, United Kingdom.

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.

Carlos Soler Montes. 2015. El modelo de lengua en el aula de ELE: Adecuación de la variedad lingüística desde un punto de vista pluricéntrico. In *La enseñanza de ELE centrada en el alumno*, pages 1237–1244. Universidad Carlos III de Madrid/ Asociación para la Enseñanza de Español como Lengua Extranjera.

Cecelia Staggs. 2019. A perception study of rioplatense spanish. *McNair Scholars Research Journal*, 14(1):11.

Eric S. Tellez, Daniela Moctezuma, Sabino Miranda, Mario Graff, and Guillermo Ruiz. 2023. Regionalized models for Spanish language variations based on Twitter. *Language Resources and Evaluation*, pages 1–31.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE . *Journal of Machine Learning Research*, 9:2579–2605.

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.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

A Calibration of the Isomorphism Metrics

Tables 3 to 8 show the detailed results for the experiments in Section 6. In all cases, tables report the results for a given isomorphism metric (either RS, EV or GH) using the 5 types of data points and 5 different number of points defined in Section 4. Metrics are evaluated on pairs of embeddings spaces $\{E_i, E_j\}$ all of them belonging to Spanish from Spain under the two training conditions of Section 6: model perturbation (Tables 3, 4 and 5) and data perturbation (Tables 6, 7 and 8).

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.989 ± 0.000	0.989	0.988	0.988	0.989	0.990	0.989	0.990	0.989	0.990	0.989
	500	0.982 ± 0.001	0.981	0.981	0.981	0.981	0.984	0.981	0.985	0.981	0.984	0.981
	1000	0.979 ± 0.002	0.978	0.977	0.978	0.978	0.982	0.978	0.983	0.978	0.982	0.977
	2500	0.976 ± 0.002	0.974	0.974	0.975	0.974	0.979	0.974	0.981	0.975	0.980	0.974
	5000	0.974 ± 0.003	0.973	0.972	0.974	0.972	0.979	0.973	0.981	0.973	0.980	0.972
Random	100	0.000 ± 0.008	0.014	-0.005	0.002	0.000	0.009	0.016	0.002	0.000	0.001	-0.013
	500	0.000 ± 0.002	0.000	-0.004	0.002	-0.004	0.002	0.003	0.000	-0.001	0.000	0.001
	1000	0.000 ± 0.000	-0.001	-0.001	0.000	-0.001	0.000	0.001	0.000	-0.001	0.000	0.000
	2500	0.000 ± 0.000	-0.001	-0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
	5000	0.000 ± 0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random BiDict	100	0.959 ± 0.002	0.960	0.957	0.955	0.957	0.963	0.958	0.964	0.957	0.963	0.958
	500	0.959 ± 0.002	0.959	0.958	0.956	0.958	0.963	0.959	0.965	0.959	0.963	0.958
	1000	0.959 ± 0.002	0.959	0.958	0.956	0.958	0.963	0.959	0.965	0.958	0.963	0.958
	2500	0.960 ± 0.002	0.960	0.958	0.957	0.959	0.963	0.959	0.965	0.958	0.963	0.958
	5000	0.960 ± 0.002	0.960	0.959	0.957	0.959	0.964	0.959	0.965	0.959	0.964	0.959
Numbers	100	0.997 ± 0.000	0.997	0.997	0.998	0.997	0.998	0.997	0.998	0.997	0.998	0.997
	500	0.994 ± 0.000	0.995	0.994	0.994	0.995	0.996	0.994	0.996	0.996	0.994	0.994
	1000	0.993 ± 0.001	0.993	0.993	0.992	0.993	0.994	0.992	0.995	0.993	0.995	0.993
	2500	0.988 ± 0.001	0.988	0.988	0.987	0.988	0.990	0.988	0.990	0.988	0.990	0.988
	5000	0.985 ± 0.001	0.986	0.985	0.984	0.985	0.987	0.985	0.987	0.985	0.987	0.985
NEs	100	-0.003 ± 0.013	-0.018	0.013	0.000	0.001	0.013	0.006	-0.011	-0.017	0.007	-0.027
	500	0.002 ± 0.007	0.002	0.002	-0.003	-0.010	0.011	0.015	0.007	0.006	0.007	-0.008
	1000	0.002 ± 0.006	0.000	0.004	-0.005	-0.010	0.002	0.015	-0.002	0.007	0.005	0.004
	2500	0.000 ± 0.003	0.001	0.000	-0.003	0.000	-0.001	0.004	-0.005	0.007	0.000	-0.002
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 3: Complete results for the RS metric with combinations of 5 embedding spaces build from the same partition of the corpus but trained using different seeds.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3E5	E4E5
Frequent	100	2 ± 1	2.3	2.8	2.6	3.3	1.8	2.2	1.6	0.9	1.0	1.5
	500	2 ± 1	4.5	2.8	4.3	3.8	2.4	1.5	1.7	1.6	1.5	2.0
	1000	3 ± 1	3.7	2.8	3.6	5.8	3.5	3.5	2.5	2.0	5.9	4.4
	2500	3 ± 1	3.0	3.6	3.2	4.9	2.3	2.6	2.6	2.5	3.3	5.9
	5000	5 ± 2	6.8	7.9	7.8	9.4	6.7	1.7	1.6	8.6	1.7	6.9
Random	100	3 ± 1	3.5	3.1	4.9	3.2	2.3	5.1	4.3	2.5	1.7	2.2
	500	5 ± 3	3.4	3.5	4.8	5.7	1.9	3.1	8.7	3.9	10.0	11.4
	1000	6 ± 2	10.1	4.7	3.4	4.6	11.4	5.3	9.6	5.2	3.9	3.8
	2500	10 ± 3	10.5	8.1	8.5	5.8	6.4	18.1	10.6	11.0	9.0	16.7
	5000	14 ± 7	6.6	7.4	23.0	9.6	13.7	22.8	4.8	27.6	15.1	17.4
Random BiDict	100	1 ± 1	2.6	1.4	0.9	1.1	2.4	2.1	1.7	1.6	1.1	0.7
	500	3 ± 1	3.7	3.0	1.5	5.7	3.5	4.0	4.1	4.0	4.3	4.4
	1000	4 ± 1	3.1	3.8	6.3	5.2	4.5	3.5	4.1	8.1	6.3	3.9
	2500	7 ± 3	9.8	8.1	4.3	9.8	2.6	10.9	2.6	8.6	4.0	11.1
	5000	5 ± 2	5.6	10.5	5.5	7.8	5.2	5.1	4.4	6.5	5.6	2.3
Numbers	100	3 ± 1	5.2	1.4	1.9	5.7	5.7	3.6	2.6	2.0	5.7	3.0
	500	4 ± 1	3.4	1.3	3.1	4.6	4.0	7.7	3.6	2.3	3.9	6.9
	1000	5 ± 2	9.6	4.3	9.1	6.2	5.2	4.7	3.9	5.7	3.0	2.7
	2500	9 ± 5	11.8	13.4	13.4	20.9	1.9	4.6	7.2	6.2	8.4	8.3
	5000	7 ± 1	6.4	4.3	8.6	8.5	8.9	9.3	8.4	6.9	8.3	3.2
NEs	100	3 ± 1	2.1	4.3	3.7	2.1	3.9	4.3	2.7	8.0	4.9	3.0
	500	7 ± 3	7.5	10.8	11.3	8.9	7.7	9.7	6.9	3.7	2.4	2.5
	1000	6 ± 3	2.0	6.5	5.7	7.3	8.2	6.4	6.4	10.3	13.5	2.7
	2500	7 ± 3	12.7	7.3	4.9	5.5	14.8	7.4	7.8	6.3	6.8	2.7
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 4: Complete results for the EV metric with combinations of 5 embedding spaces build from the same partition of the corpus but trained using different seeds.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.02 ± 0.00	0.027	0.031	0.038	0.033	0.025	0.037	0.025	0.020	0.017	0.036
	500	0.02 ± 0.00	0.025	0.021	0.021	0.017	0.025	0.023	0.030	0.020	0.022	0.018
	1000	0.02 ± 0.00	0.025	0.026	0.020	0.018	0.025	0.018	0.026	0.017	0.022	0.015
	2500	0.01 ± 0.00	0.018	0.020	0.016	0.015	0.021	0.024	0.023	0.019	0.018	0.013
	5000	0.01 ± 0.00	0.017	0.019	0.014	0.015	0.021	0.018	0.010	0.015	0.012	0.013
Random	100	0.17 ± 0.12	0.086	0.073	0.244	0.127	0.072	0.317	0.054	0.318	0.053	0.371
	500	0.15 ± 0.06	0.100	0.256	0.160	0.273	0.171	0.128	0.189	0.107	0.029	0.118
	1000	0.16 ± 0.07	0.084	0.245	0.080	0.270	0.160	0.100	0.185	0.202	0.052	0.226
	2500	0.11 ± 0.03	0.123	0.136	0.125	0.068	0.079	0.075	0.125	0.127	0.205	0.077
	5000	0.07 ± 0.01	0.044	0.076	0.046	0.095	0.052	0.064	0.089	0.086	0.086	0.077
Random BiDict	100	0.03 ± 0.01	0.018	0.021	0.039	0.021	0.030	0.042	0.029	0.061	0.026	0.035
	500	0.02 ± 0.00	0.021	0.032	0.017	0.022	0.016	0.019	0.020	0.029	0.034	0.035
	1000	0.02 ± 0.00	0.031	0.032	0.029	0.027	0.020	0.036	0.017	0.039	0.030	0.038
	2500	0.02 ± 0.00	0.022	0.017	0.016	0.027	0.020	0.021	0.019	0.014	0.022	0.031
	5000	0.01 ± 0.00	0.028	0.026	0.018	0.018	0.013	0.019	0.015	0.015	0.012	0.014
Numbers	100	0.05 ± 0.05	0.030	0.043	0.055	0.038	0.015	0.025	0.027	0.039	0.208	0.035
	500	0.02 ± 0.00	0.023	0.027	0.025	0.027	0.031	0.032	0.021	0.025	0.019	0.022
	1000	0.02 ± 0.00	0.020	0.025	0.028	0.019	0.014	0.026	0.019	0.040	0.024	0.020
	2500	0.02 ± 0.00	0.031	0.025	0.019	0.019	0.022	0.028	0.023	0.025	0.020	0.028
	5000	0.02 ± 0.00	0.019	0.022	0.024	0.024	0.020	0.017	0.021	0.032	0.032	0.026
NEs	100	0.10 ± 0.03	0.151	0.080	0.093	0.062	0.125	0.170	0.135	0.121	0.105	0.041
	500	0.07 ± 0.03	0.049	0.073	0.072	0.080	0.061	0.032	0.107	0.079	0.154	0.076
	1000	0.09 ± 0.03	0.120	0.053	0.068	0.134	0.094	0.132	0.031	0.063	0.102	0.146
	2500	0.03 ± 0.00	0.029	0.045	0.033	0.029	0.039	0.027	0.017	0.040	0.035	0.025
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 5: Complete results for the GH metric with combinations of 5 embedding spaces build from the same partition of the corpus but trained using different seeds.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.860 ± 0.056	0.839	0.837	0.768	0.882	0.901	0.871	0.943	0.787	0.945	0.828
	500	0.314 ± 0.032	0.333	0.312	0.264	0.331	0.329	0.329	0.347	0.277	0.357	0.263
	1000	0.131 ± 0.014	0.151	0.123	0.119	0.147	0.132	0.129	0.138	0.109	0.152	0.114
	2500	0.038 ± 0.000	0.040	0.039	0.033	0.045	0.042	0.037	0.041	0.030	0.043	0.031
	5000	0.015 ± 0.001	0.017	0.015	0.014	0.017	0.016	0.013	0.015	0.013	0.018	0.014
Random	100	0.002 ± 0.015	0.014	-0.010	0.009	0.026	0.004	-0.007	0.003	-0.033	0.013	0.008
	500	0.000 ± 0.001	-0.001	0.000	-0.002	-0.004	0.001	0.000	0.001	0.000	-0.003	0.003
	1000	0.000 ± 0.000	0.000	0.000	0.000	0.000	0.001	-0.001	0.000	0.000	0.000	0.000
	2500	0.000 ± 0.000	-0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	5000	0.000 ± 0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random	100	0.884 ± 0.008	0.870	0.894	0.894	0.882	0.889	0.892	0.879	0.889	0.883	0.869
BiDict	500	0.882 ± 0.004	0.880	0.881	0.878	0.886	0.881	0.885	0.878	0.891	0.889	0.879
	1000	0.883 ± 0.002	0.885	0.885	0.883	0.883	0.883	0.883	0.878	0.887	0.886	0.881
	2500	0.883 ± 0.001	0.884	0.883	0.882	0.886	0.884	0.883	0.880	0.885	0.885	0.881
	5000	0.883 ± 0.000	0.882	0.884	0.885	0.885	0.883	0.884	0.883	0.885	0.884	0.884
Numbers	100	0.604 ± 0.087	0.809	0.687	0.631	0.533	0.639	0.516	0.539	0.549	0.534	0.604
	500	0.116 ± 0.012	0.133	0.119	0.134	0.104	0.131	0.098	0.112	0.113	0.100	0.120
	1000	0.061 ± 0.008	0.049	0.068	0.073	0.071	0.067	0.057	0.058	0.049	0.060	0.067
	2500	0.037 ± 0.007	0.042	0.045	0.036	0.055	0.036	0.040	0.034	0.028	0.037	0.026
	5000	0.022 ± 0.003	0.021	0.028	0.022	0.026	0.021	0.029	0.017	0.019	0.021	0.021
NEs	100	-0.001 ± 0.018	0.029	0.008	-0.005	-0.007	-0.005	0.003	-0.029	-0.011	-0.027	0.030
	500	-0.003 ± 0.007	-0.002	0.006	0.008	-0.008	-0.015	-0.008	-0.008	-0.002	-0.011	0.002
	1000	-0.002 ± 0.004	-0.008	0.004	0.0	-0.008	-0.006	-0.005	0.000	-0.003	-0.005	0.003
	2500	-0.002 ± 0.002	0.000	0.003	-0.004	-0.004	0.000	-0.003	0.001	-0.002	-0.006	-0.005
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 6: Complete results for the RS metric with combinations of 5 embedding spaces build from 5 different random partitions of the CEREAL corpus.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	2 ± 1	2.7	3.0	3.4	2.9	0.9	1.2	1.9	0.7	1.5	2.0
	500	3 ± 1	6.7	4.0	5.5	1.8	2.8	4.4	4.6	3.0	2.2	3.7
	1000	2 ± 1	3.6	3.5	3.8	5.8	1.8	1.6	2.6	2.2	2.6	1.6
	2500	4 ± 1	3.8	4.4	1.5	5.3	1.6	5.1	7.6	5.9	6.1	6.6
	5000	5 ± 4	1.1	3.3	2.8	11.8	2.1	2.3	10.9	3.8	11.1	9.0
Random	100	5 ± 1	6.1	3.5	5.6	2.8	9.7	5.9	5.0	5.6	3.9	6.8
	500	5 ± 2	8.0	5.4	4.7	4.8	3.9	10.3	3.1	4.6	1.5	5.6
	1000	5 ± 1	3.7	6.4	3.8	5.4	5.4	4.2	5.5	6.4	8.8	5.1
	2500	7 ± 1	3.3	6.4	8.1	6.4	6.9	7.4	6.6	11.2	9.6	6.6
	5000	14 ± 5	19.1	13.1	6.6	8.6	10.3	16.1	25.3	6.9	21.1	14.1
Random	100	3 ± 2	2.9	2.2	8.0	2.8	7.6	1.7	2.8	1.3	1.7	5.5
BiDict	500	4 ± 1	2.0	5.6	4.1	3.3	6.6	2.6	3.3	2.2	4.7	6.1
	1000	6 ± 4	10.9	15.0	7.5	1.6	1.8	9.6	3.9	4.0	5.1	10.3
	2500	7 ± 2	5.2	7.5	13.3	9.5	9.1	6.8	8.3	9.1	4.6	5.0
	5000	8 ± 4	6.0	9.1	4.5	4.9	5.1	21.1	6.2	9.5	10.4	6.0
Numbers	100	3 ± 1	3.2	7.1	3.5	1.6	2.6	2.0	3.1	2.0	7.4	4.3
	500	5 ± 2	1.8	4.6	8.7	5.0	5.3	8.6	5.7	8.8	3.0	5.3
	1000	9 ± 6	3.6	3.2	20.1	8.0	2.8	15.3	4.6	17.0	6.9	16.3
	2500	12 ± 4	12.3	17.3	15.2	19.0	14.2	14.1	6.6	5.1	11.6	14.4
	5000	11 ± 3	16.2	6.8	9.3	15.8	12.6	10.1	6.1	13.7	11.3	15.3
NEs	100	2 ± 1	4.2	1.1	1.8	1.1	3.9	5.5	3.5	1.7	2.0	3.3
	500	3 ± 1	3.5	4.0	3.1	2.8	7.3	5.3	5.6	3.4	1.9	1.9
	1000	7 ± 2	4.0	11.1	5.9	11.0	8.0	7.9	6.8	5.7	4.9	8.5
	2500	7 ± 2	8.5	10.1	14.6	7.4	4.8	9.6	5.3	5.3	4.5	7.2
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 7: Complete results for the EV metric with combinations of 5 embedding spaces build from 5 different random partitions of the CEREAL corpus.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.03 ± 0.01	0.026	0.058	0.023	0.024	0.038	0.028	0.034	0.035	0.062	0.027
	500	0.02 ± 0.00	0.029	0.022	0.028	0.020	0.022	0.025	0.023	0.021	0.026	0.022
	1000	0.02 ± 0.00	0.019	0.015	0.028	0.020	0.018	0.021	0.022	0.020	0.026	0.022
	2500	0.01 ± 0.00	0.013	0.015	0.019	0.020	0.015	0.015	0.020	0.021	0.021	0.021
	5000	0.01 ± 0.00	0.018	0.018	0.014	0.020	0.016	0.016	0.017	0.018	0.021	0.014
Random	100	0.15 ± 0.06	0.037	0.072	0.209	0.134	0.086	0.215	0.133	0.228	0.206	0.198
	500	0.19 ± 0.10	0.373	0.299	0.188	0.332	0.107	0.184	0.075	0.133	0.074	0.155
	1000	0.13 ± 0.05	0.065	0.153	0.176	0.073	0.129	0.123	0.070	0.253	0.146	0.132
	2500	0.13 ± 0.03	0.143	0.091	0.188	0.160	0.129	0.116	0.067	0.186	0.146	0.103
	5000	0.12 ± 0.04	0.090	0.068	0.163	0.116	0.085	0.160	0.096	0.166	0.102	0.195
Random BiDict	100	0.05 ± 0.02	0.036	0.031	0.045	0.106	0.033	0.088	0.029	0.040	0.051	0.057
	500	0.03 ± 0.01	0.044	0.039	0.028	0.030	0.027	0.080	0.031	0.052	0.025	0.028
	1000	0.03 ± 0.01	0.069	0.023	0.034	0.036	0.031	0.025	0.031	0.037	0.035	0.036
	2500	0.03 ± 0.01	0.036	0.032	0.043	0.022	0.019	0.065	0.031	0.024	0.023	0.037
	5000	0.02 ± 0.00	0.016	0.019	0.024	0.028	0.025	0.027	0.030	0.023	0.022	0.024
Numbers	100	0.06 ± 0.04	0.031	0.040	0.125	0.030	0.028	0.124	0.022	0.120	0.028	0.114
	500	0.05 ± 0.00	0.050	0.040	0.054	0.043	0.060	0.043	0.047	0.052	0.055	0.071
	1000	0.02 ± 0.00	0.024	0.030	0.031	0.033	0.033	0.033	0.020	0.026	0.038	0.023
	2500	0.02 ± 0.00	0.024	0.026	0.022	0.034	0.027	0.033	0.030	0.023	0.031	0.036
	5000	0.05 ± 0.01	0.044	0.021	0.070	0.073	0.040	0.067	0.069	0.065	0.059	0.041
NEs	100	0.08 ± 0.02	0.063	0.123	0.053	0.057	0.129	0.055	0.121	0.098	0.101	0.066
	500	0.07 ± 0.03	0.106	0.113	0.081	0.045	0.023	0.069	0.107	0.075	0.114	0.052
	1000	0.07 ± 0.02	0.102	0.058	0.123	0.066	0.046	0.077	0.058	0.096	0.040	0.076
	2500	0.05 ± 0.02	0.067	0.027	0.030	0.043	0.084	0.064	0.052	0.041	0.054	0.041
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 8: Complete results for the GH metric with combinations of 5 embedding spaces build from 5 different random partitions of the CEREAL corpus.

B Qualitative Behaviour of the Embedding Spaces

Figure 6 shows the 10-top nearest neighbours for three different varieties of Spanish words that would translate into *corn*: `elote`, `choclo`, and `mazorca` (cf. Section 7). The results with respect to three embedding spaces —cl, es and mx— show the differences associated to the three concepts. For instance, `elote` goes from inexistence in cl to all the way into a neighbourhood of regional ingredients and dishes in mx, passing through a concept more associated to foreign cuisine in es.

Figure 6: t-SNE projections (van der Maaten and Hinton, 2008) for the neighbouring spaces for the word `corn`, used as `choclo` in Chile (cl), `mazorca` in Spain (es) and `elote` in Mexico (mx).

C Extended Isomorphism Results on the Variety-Specific Spanish Embeddings

Figures 7, 8 and 9 show the extended results for the experiments in Section 7. Here, in addition to MFW 100 for EV (best configuration reported in the main text), we show the results for the top-2 best configurations for the three isomorphism metrics, RS, EV and GH: RS on MFW 100 and random BiDict 100, EV on BiDict 100, and GH on the MFW 5,000 and random BiDict 5,000 (in cases where 5,000 points are not available we use the maximum number of available points). In all cases, figures represent the scores for a given isomorphism metric using the embeddings computed on CEREAL and on Twitter data.

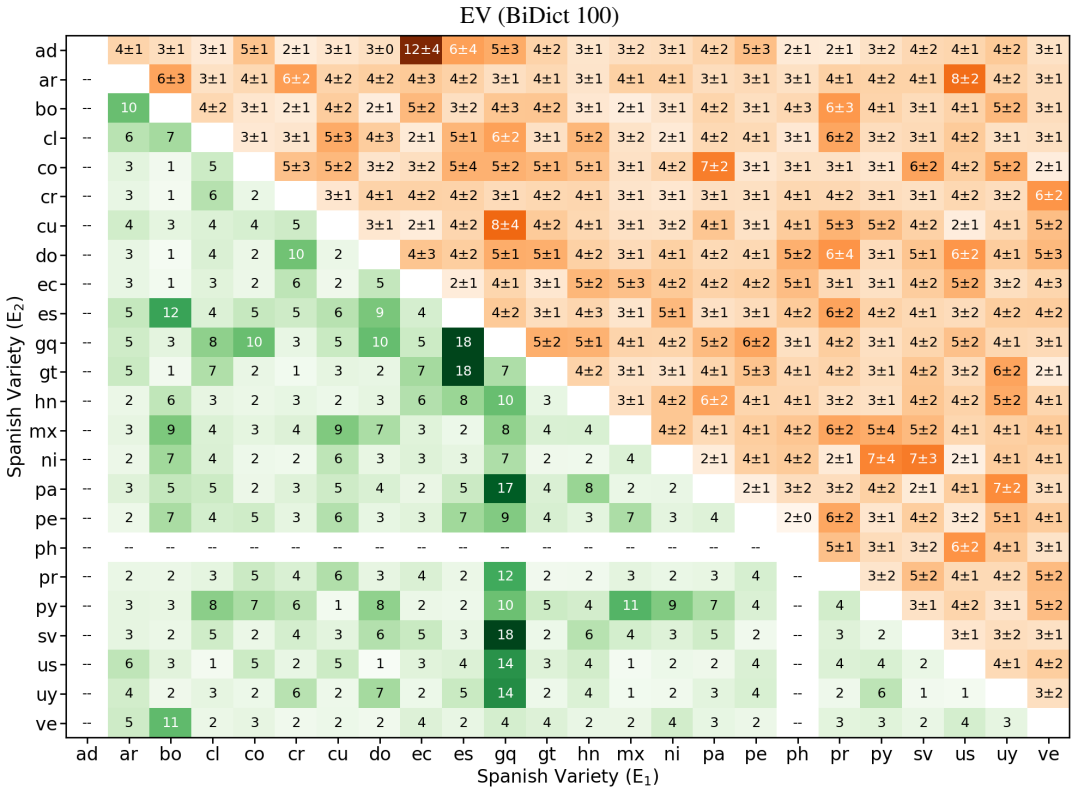
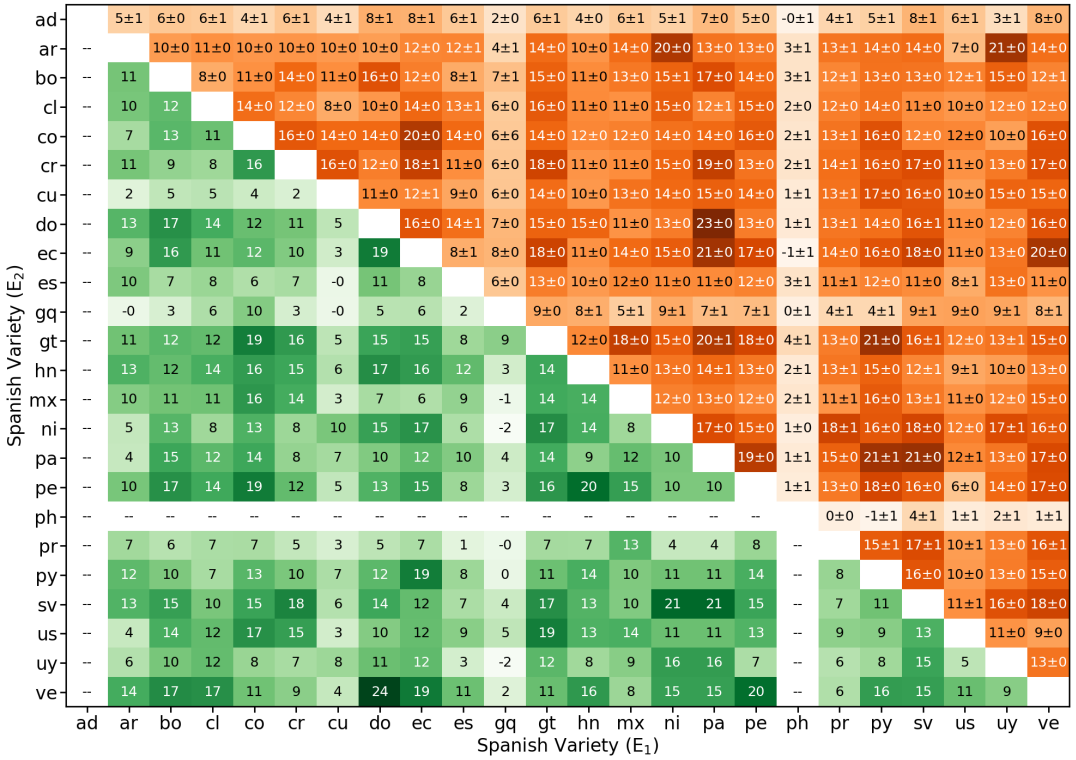


Figure 7: EV with random BiDict 100 words for the 24 Spanish varieties. Top–right triangles (orange) correspond to the results with the CEREAL corpus and bottom–left triangle (green) to the Twitter corpus.

(a) RS (MFW 100)



(b) RS (BiDict 100)

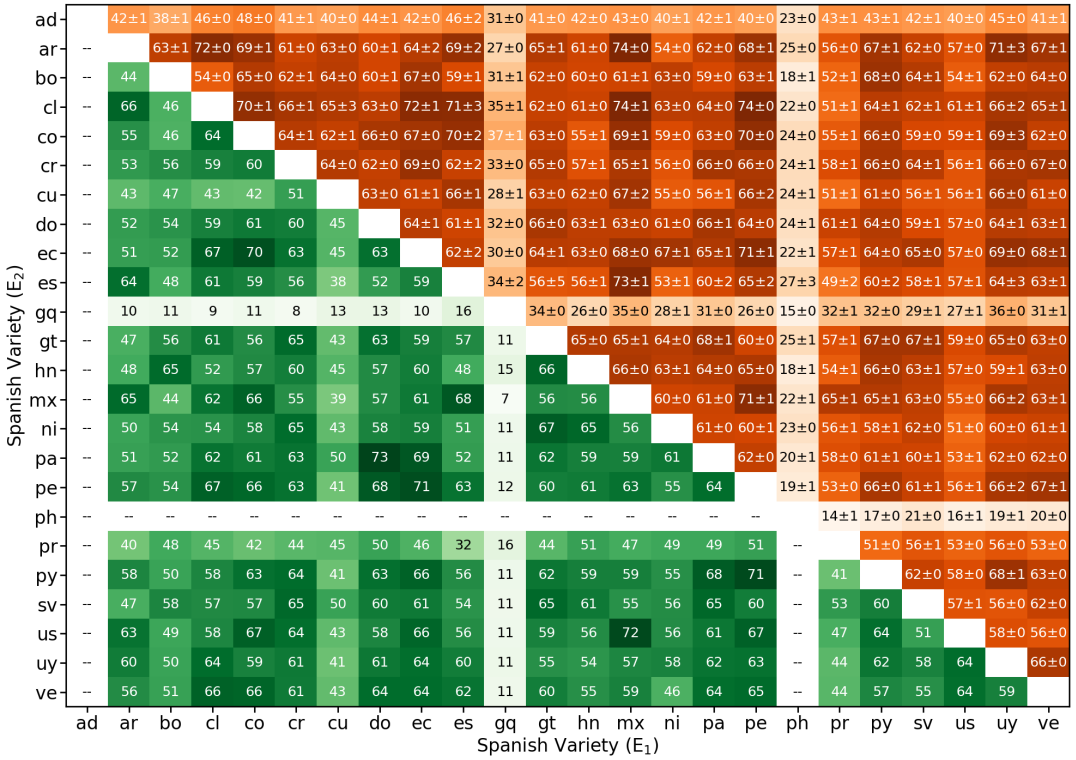
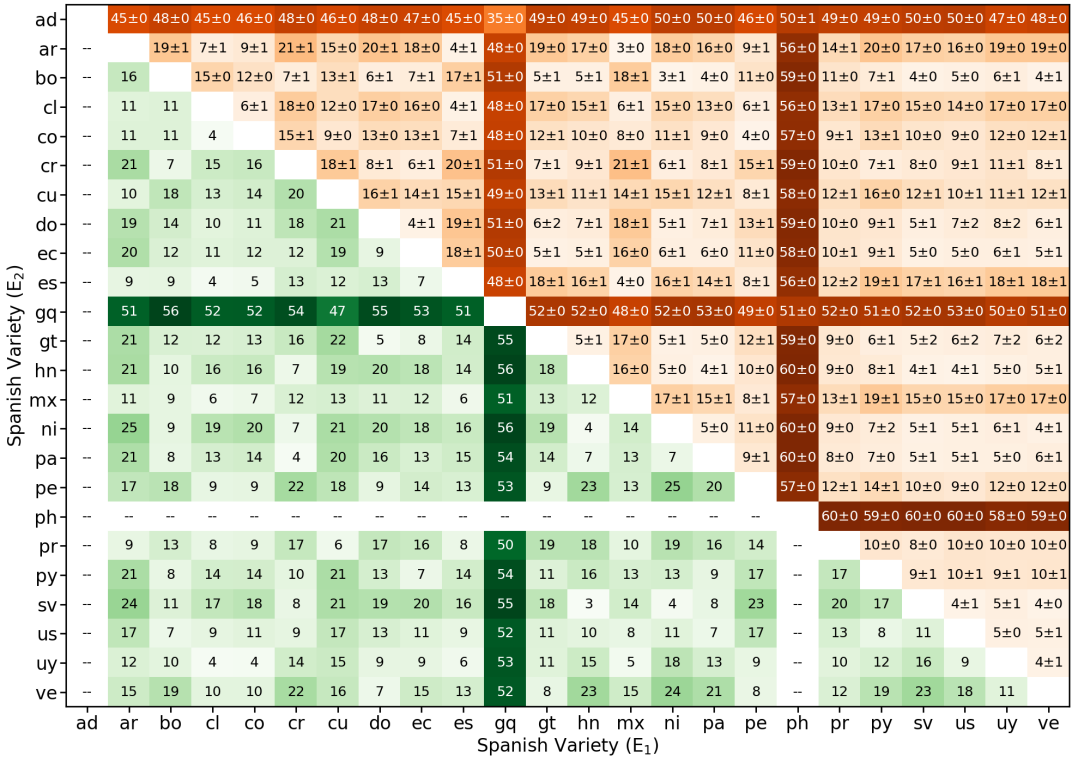


Figure 8: RS with (a) 100 most frequent words (MFW) and (b) BiDict entries multiplied by 100 for better readability. Top–right triangles (orange) correspond to the mean results with the CEREAL corpus and bottom–left triangle (green) to the Twitter corpus.

(a) GH (MFW 5,000)



(b) GH (BiDict 5,000)

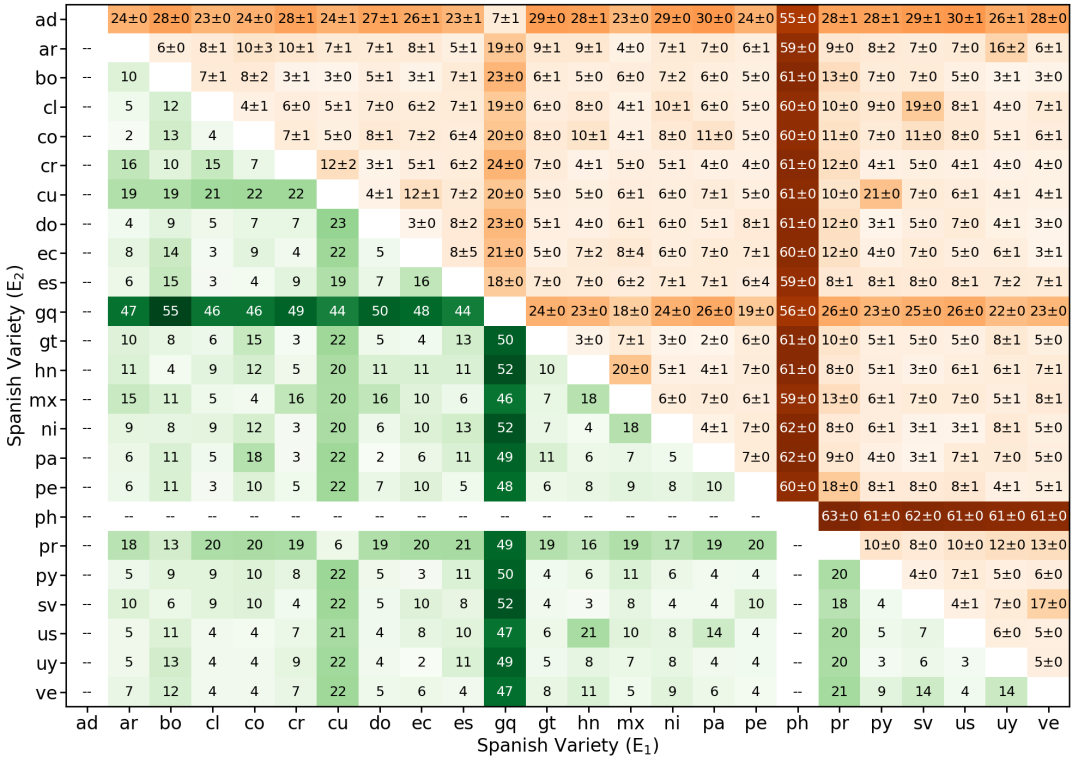


Figure 9: (GH with (a) 5,000 most frequent words (MFW) and (b) GH with 5,000 random BiDict words for the 24 Spanish varieties. Results are multiplied by 100 for better readability. Top–right triangles (orange) correspond to the mean results with the CEREAL corpus and bottom–left triangle (green) to the Twitter corpus.

D Extended Analysis on Phylogenetics

D.1 Visual Representation

Figure 10 focuses on EV and represents the regions that can be drawn on the basis of the resulting clusters. A comparison against the Spanish linguistic zones as defined by RAE (see Figure 1) reveals some divergences. Among them, Central America not necessarily being tied to Mexico as well as Colombia and Venezuela, which here appear differentiated.

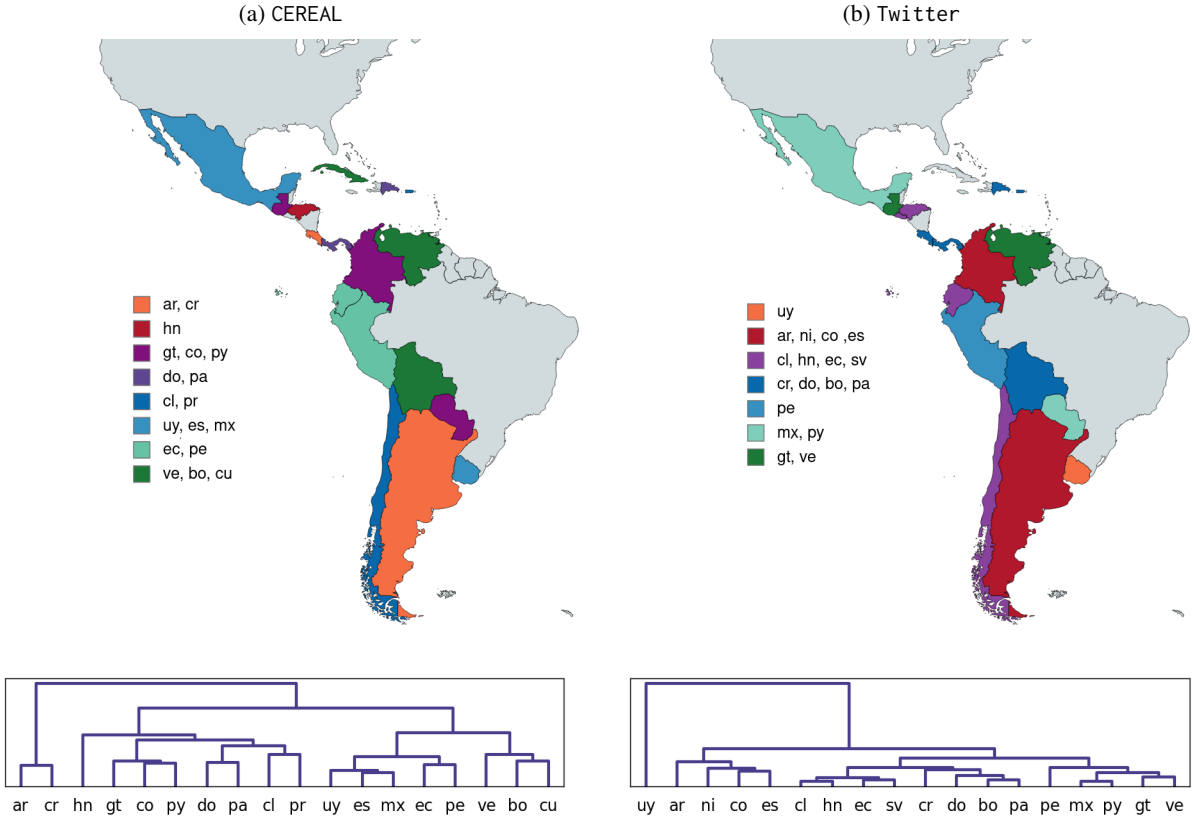


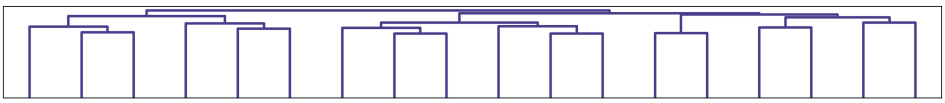
Figure 10: Geographical representation of the Spanish varieties clustered according the EV (MFW 100) score; *es* is omitted from the plot for visibility reasons, but it is included in the legend together with the family it groups with. Plots are done with MapChart (`https://www.mapchart.net`).

D.2 Extended Results on the Hierarchical Clustering Experiments

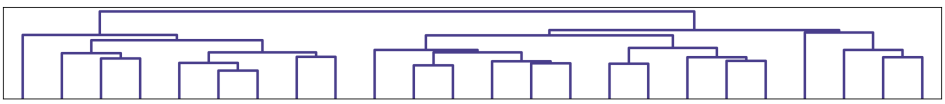
As in the previous sections, we show, for completeness, results for the top-2 best configurations for the 3 isomorphism metrics: RS, EV and GH. Figure 11 depicts the phylogenetic (relational) trees obtained from scores on the embeddings built with `CEREAL` for the 2nd best performing configurations (1st one is in the main text): RS on random BiDict 100, EV on BiDict 100, and GH on random BiDict 5,000. We compare the trees for 24 varieties and the subset of the 17 highest resourced varieties. Figure 12 shows the top-2 configurations for the scores derived from the embeddings computed on `Twitter` data.

(a) RS (BiDict 100), all varieties

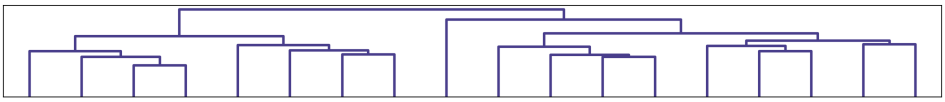
(b) RS (BiDict 100), high-resourced varieties



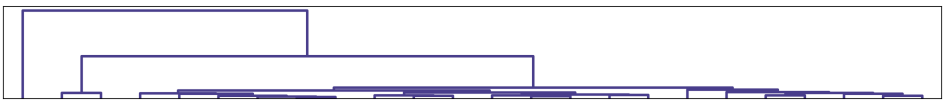
(c) EV (BiDict 100), all varieties



(d) EV (BiDict 100), high-resourced varieties



(e) GH (BiDict 5,000), all varieties



(f) GH (BiDict 5,000), high-resourced varieties

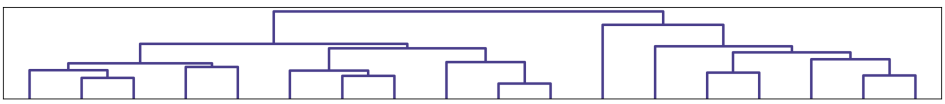
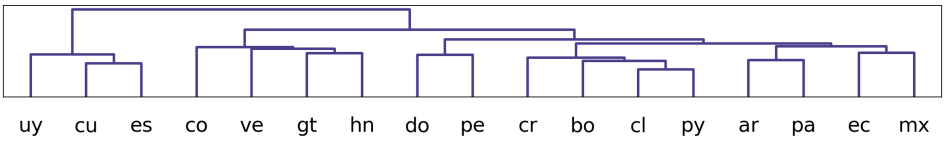
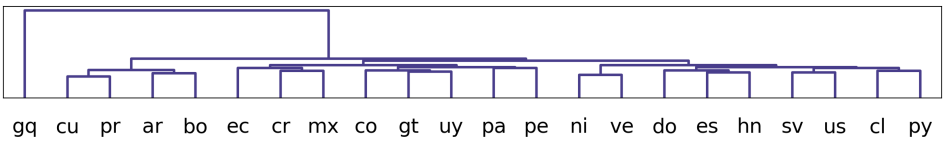


Figure 11: Hierarchical clustering on the outputs of the isomorphism measures obtained in Section 7 for the embeddings computed using the CEREAL corpus.

(a) RS (MFW 100)

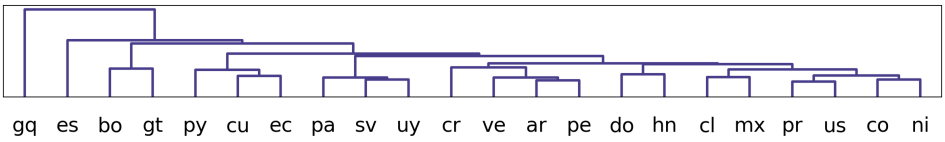


(b) RS (BiDict 100)

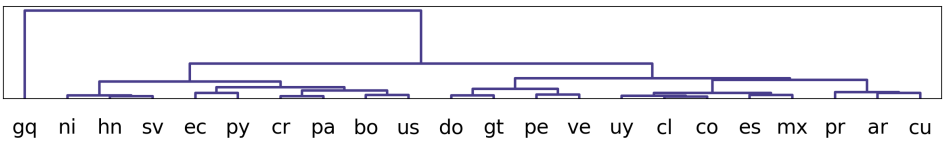


(c) EV (MFW 100)

(d) EV (BiDict 100)



(e) GH (MFW 5,000)



(f) GH (BiDict 5,000)

Figure 12: Hierarchical clustering on the outputs of the isomorphism measures obtained in Section 7 with Twitter embeddings.