Elote, Choclo and Mazorca: on the Varieties of Spanish

Cristina España-Bonet
DFKI GmbH, Saarland Informatics Campus
Saarbrüken, Germany
cristinae@dfki.de

Alberto Barrón-Cedeño
Universitá di Bologna
Forlì, Italy
a.barron@unibo.it

Abstract

Spanish is one of the most widespread languages: the official language in 20 countries and the second most-spoken native language. Its contact with other languages across different regions and the rich regional and cultural diversity has produced varieties which divert from each other, particularly in terms of lexicon. Still, available corpora, and models trained upon them, generally treat Spanish as one monolithic language, which dampers prediction and generation power when dealing with different varieties. To alleviate the situation, we compile and curate datasets in the different varieties of Spanish around the world at an unprecedented scale and create the CEREAL corpus. With such a resource at hand, we perform a stylistic analysis to identify and characterise varietal differences. We implement a classifier specially designed to deal with long documents and identify Spanish varieties (and therefore expand CEREAL further). We produce varietal-specific embeddings, and analyse the cultural differences that they encode. We make data, code and models publicly available.

1 Introduction

Spanish was heavily expanded by overseas European conquest starting in the early-16th century. One of the main axes of the colonisation was the imposition of the Spanish culture. As a result, nearly 70% of the current population in Latin America is Catholic and Spanish is a global language (Ammon, 2010) and the second largest native language over the world (+450M speakers), only after Chinese (Eberhard et al., 2023). Different from the latter, Spanish is a sea-bound language and it is the official language in 20 countries across America, Africa and Europe, spreading over 11.7M km^2. The diversity in terms of culture, geography and languages in contact (e.g., Basque in Northern Spain, Nahuatl in Central Mexico, Quechua in the Andes) have produced well-differentiated varieties[1] permeated by the lexicon and semantics from the relevant local languages.

Even with the apparent diversity, Spanish is often considered a homogeneous language in most data-driven natural language processing (NLP) tasks and applications. A first reason is the difficulty to get data for all the varieties, also taking into account the lack of information about the origin of the data for lots of the existing corpora. A second reason is a popular motto in the NLP community: *the more data the better*. Following this mantra, data in Spanish is usually mixed together and, as a consequence, one might increase the performance on downstream tasks at the cost of erasing cultural differences. Such differences can be reflected in textual corpora by the topics present, the linguistic forms used or even the writing style. At the corpus scale, the differences are evident, but in individual texts, specially short ones, differences might be non-existent. Because of this reason, a text might belong to more than one variety but, the longer the text, the more likely its origin country shines. The register is relevant as well. For instance, newspapers tend to follow standard norms. Under this point of view, the jungle of the Web is a good source to get diverse data. Still, one should acknowledge the presence of collaborative texts (i.e. the Wikipedia) and media outlets for which no clear variety can be assigned.

We develop a long-document classifier to extract and release CEREAL and CEREALex, *Corpus del Español REAL (extended)*, a new corpus from online data five times larger than the NOW corpus, to the best of our knowledge the largest available corpus of Spanish varieties up to date. CEREALex contains 35 B words in 41 M documents, 36 M of which include information about the country they were

[1] We refer to varieties of Spanish following Hudson (1996): "a *variety* is a set of linguistic items with similar social (including geographical and cultural) distribution."

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3689–3711

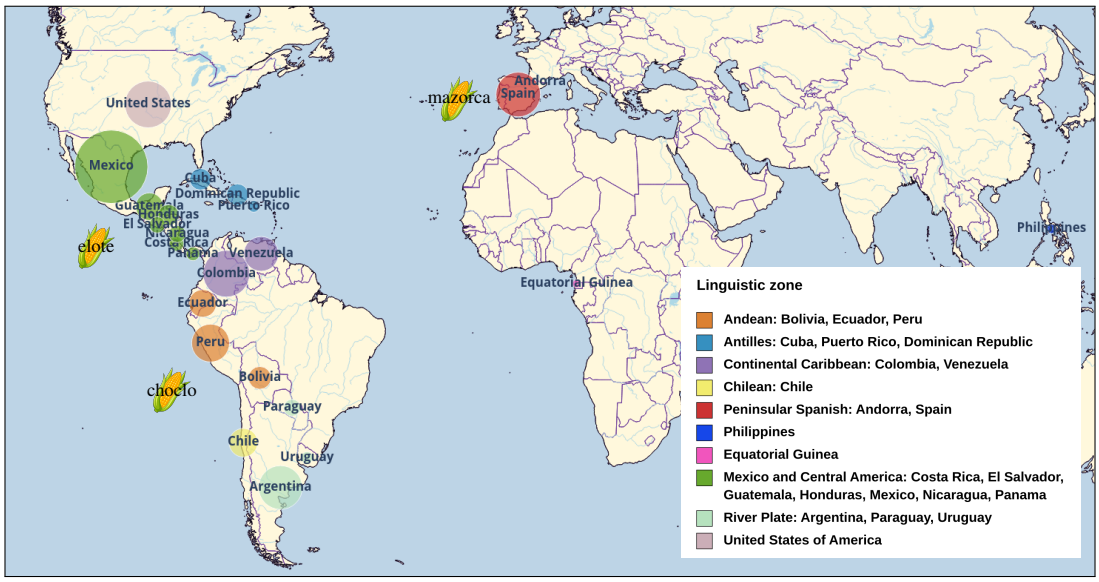


Figure 1: Common geographic (country-based) Spanish linguistic zones as described by the *Real Academia Española*. The size of the bubbles is proportional to the number of Spanish speakers. The three words associated to corn included in the title of this work appear on the sea close to the regions where they are most used.

published at covering 24 countries: Latin American countries, Spain, Andorra, Equatorial Guinea, the Philippines and the US (see Section 4). After the analysis of the corpus (Section 3), including a wide stylistic analysis, we create embeddings for each variety and explore if they are able to reflect the cultural aspects of their country by inducing bilingual lexicons and reproducing human biases (Section 5). Data, code and models are freely available.[2]

2 Related Work

Geolocated Corpora in Spanish. The *Real Academia Española* (RAE) created the *Corpus del Español del Siglo XXI* (CORPES) (RAE, 2023). CORPES allocates 30% of the contents to forms from Spain and 70% to forms from America. The material produced in America is further classified into the linguistic zones shown in Figure 1 which also includes the Philippines, Equatorial Guinea and Spain itself. Written materials come from books (40%), newspapers (40%), Internet (7.5%), and others (2.5%). The remaining 10% contains oral language. CORPES v1.0 features 365 k documents, totalling over 395 M orthographic forms; 256 M for the American varieties. Unfortunately, CORPES is only browsable and is not fully downloadable. The *Corpus del Español NOW*, is another corpus whose construction is rooted on diversity.[3] It is larger than CORPES —with 7 B words— and covers 21 countries. In this case, the proportion between continents is 78% America vs 22% Europe. NOW is downloadable, upon the payment of a fee.

Another effort is that of Gonçalves and Sánchez (2014), who collect 0.75 M geolocated tweets posted over a two-year span. They queried the Twitter API with seed words representing concepts that are expressed in different ways across countries; e.g., corn is `elote` in Mexico and most of Central America (from the Nahuatl `elotitutl`), `choclo` in most of South America (from the Quechua `chuqllu`) and `mazorca` in Colombia, Cuba, and Spain (from the Arabic `masúrqa`). Recently, Tellez et al. (2023) collected 800 M tweets (11 B tokens) over a three-year span, 2016 to 2019, posted from 26 countries (including a few where Spanish is not an official language). With them, they create 26 country-specific word embeddings, plus BERT-like language models for the countries with the largest amounts of data. The models are publicly available, but the data is not.[4]

Smaller-scale corpora include those for the series of shared tasks on Discriminating Between Similar Languages (DSL), organised within the Workshop on NLP for Similar Languages, Varieties and Dialects. The DSL Collection (Tan et al., 2014) contains individual sentences in Spanish from Ar-

[2] https://cereal-es.github.io/CEREAL
[3] https://www.corpusdelespanol.org
[4] https://ingeotec.github.io/regional-spanish-models/

ccTLD	Country	CEREAL docs.	CEREAL segments	Training$^{\#class}$	Validation$^{\#class}$ & Tests$^{\#class}$	Classified$^{\#class}$	CEREALex docs.
ad	Andorra	1.5k	13k	$0^{3,4,5}$	$0^{3,4,5}$	NA	1.5k
ar	Argentina	2.0M	21M	$30k^{3,4}/500k^{5}$	$63^{3,4}/1k^{5}$	$NA^{3,4}/2.7M^{5}$	4.7M
bo	Bolivia	75k	976k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	75k
cl	Chile	1.1M	12.1M	$30k^{3}/500k^{4,5}$	$64^{3}/1k^{4,5}$	$NA^{3}/2.1M^{4}/1.9M^{5}$	2.2M
co	Colombia	650k	8.3M	$30k^{3,4,5}$	$63^{3,4,5}$	NA	650k
cr	Costa Rica	59k	826k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	59k
cu	Cuba	116k	1.9M	$30k^{3,4,5}$	$63^{3,4,5}$	NA	116k
do	Dominican Rep.	14k	1.2M	$30k^{3,4,5}$	$63^{3,4,5}$	NA	14k
ec	Ecuador	158k	1.6M	$30k^{3,4,5}$	$63^{3,4,5}$	NA	158k
es	Spain	5.7M	70.5M	$500k^{3,4,5}$	$1k^{3,4,5}$	$17.0M^{3}/15.5M^{4}/15.8M^{5}$	21.4M
gq	Eq. Guinea	801	4k	$628^{3,4,5}$	$0^{3,4,5}$	NA	801
gt	Guatemala	51k	562k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	51k
hn	Honduras	60k	657k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	60k
mx	Mexico	2.4M	20.9M	$500k^{3,4,5}$	$1k^{3,4,5}$	$5.0M^{3}/4.2M^{4}/2.9M^{5}$	5.7M
ni	Nicaragua	37k	406k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	37k
pa	Panama	39k	449k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	39k
pe	Peru	442k	5.1M	$30k^{3,4,5}$	$63^{3,4,5}$	NA	442k
ph	Philippines	112	1.5k	$91^{3,4,5}$	$0^{3,4,5}$	NA	112
pr	Puerto Rico	12k	128k	$11k^{3,4,5}$	$0^{3,4,5}$	NA	12K
py	Paraguay	66k	776k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	66k
sv	El Salvador	41k	402k	$30k^{3,4,5}$	$63^{3,4,5}$	NA	41k
us	United States	22k	378k	$19k^{3,4,5}$	$0^{3,4,5}$	NA	22k
uy	Uruguay	154k	1.8M	$30k^{3,4,5}$	$63^{3,4,5}$	NA	154k
ve	Venezuela	109k	1.2M	$30k^{3,4,5}$	$63^{3,4,5}$	NA	109k
All with ccTDL		13.5M	151.1M	NA	NA	NA	NA
Unknown		27.7M	337.4M	NA	NA	$7.8M^{3}/5.8M^{4}/4.4M^{5}$	4.9M
Total		NA	NA	$1.5M^{3}/2.0M^{4}/2.5M^{5}$	$3k^{3}/4k^{4}/5k^{5}$	NA	36.3M

Table 1: Distribution of documents and segments after deduplication. Summary of the training, validation and test data. Column *Classified* shows the number of documents in CEREAL labelled with our docTransformer classifier. The superindexes refer to the amount of instances included in the settings considering 3, 4, or 5 classes.

gentina, Mexico, Peru, and Spain.

Text Classification of Long Documents. Since the appearance of BERT (Devlin et al., 2019) and the change of paradigm in NLP it involved, most classification tasks achieve state-of-the-art results by fine-tuning a language model (Sun et al., 2019; Adhikari et al., 2019). BERT-like models are based on self-attention, which scales quadratically with the input length. As a result, training is usually constrained to input texts of up to 512 tokens. Longer inputs can be processed with more efficient architectures, such as Linformer (Wang et al., 2020), Big Bird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020). These architectures accept at least 4096 input tokens thanks to the usage of sparse attention mechanisms that scale linearly. However, it has been noticed that the models above do not always improve the results of a simple truncation of the input to 512 tokens or a sentence average (Park et al., 2022; Sannigrahi et al., 2023).

Other approaches involve the identification of key sentences in a document and their concatenation (Ding et al., 2020) or a hierarchical transformer with a nested sentence encoder and document encoder (Zhang et al., 2019). Closer to our work, Pappagari et al. (2019) split the input document before sending it to a base language model and, afterwards, feed the outputs into another transformer. Our model is effective and simpler in the sense that we do not need a complementary module besides the base language model.

Cultural NLP. Multilinguality and a fair treatment of all the languages is considered as one of the biggest challenges in current NLP (Joshi et al., 2020). To date, little work on NLP has focused on cross/multi-cultural aspects because of the additional challenges involved, and because of a lack of awareness from the researchers (Hershcovich et al., 2022). The workshop on Cross-Cultural Considerations in NLP[5] represents a step to promote awareness. A few studies are beginning to appear that analyse how multilinguality changes or affects cultural norms in word embeddings and language models (España-Bonet and Barrón-Cedeño, 2022; Haemmerl et al., 2023; Mukherjee et al., 2023;

[5] https://sites.google.com/view/c3nlp/home

	Tokens	punct.	Ratio of type-token	legomena	Entropy	Yule's K	SzigrisztPazos /INFLEZ	Laplacian Energy	Clustering
ar	578 ± 479	2.3 ± 0.9	51 ± 10	37 ± 11	6.9 ± 0.7	152 ± 59	59 ± 13	0.46 ± 0.05	0.08 ± 0.04
bo	672 ± 496	2.3 ± 0.9	46 ± 12	31 ± 15	7.0 ± 0.6	159 ± 60	58 ± 13	0.45 ± 0.05	0.08 ± 0.03
cl	621 ± 520	2.3 ± 0.7	50 ± 11	36 ± 12	7.0 ± 0.7	155 ± 65	56 ± 15	0.46 ± 0.05	0.08 ± 0.04
co	248 ± 531	1.5 ± 0.8	53 ± 12	36 ± 13	6.5 ± 0.7	181 ± 69	54 ± 16	0.47 ± 0.06	0.05 ± 0.04
cu	772 ± 596	2.5 ± 1.0	50 ± 10	38 ± 11	7.2 ± 0.6	149 ± 68	56 ± 12	0.46 ± 0.06	0.08 ± 0.03
es	702 ± 556	2.4 ± 0.9	48 ± 11	34 ± 11	7.0 ± 0.7	155 ± 67	59 ± 20	0.45 ± 0.06	0.09 ± 0.04
mx	594 ± 477	2.3 ± 0.7	50 ± 11	36 ± 12	7.0 ± 0.6	149 ± 51	56 ± 14	0.46 ± 0.05	0.08 ± 0.04
pe	562 ± 466	2.4 ± 0.7	50 ± 11	36 ± 11	6.9 ± 0.7	155 ± 59	57 ± 16	0.46 ± 0.05	0.08 ± 0.04

Table 2: Style, richness and readability metrics for the dialectal variations in the CEREAL corpus. The numbers are an average over the (up to) 30 k instances for each variation. Appendix A shows all 20 metrics for all varieties.

Ramesh et al., 2023; Kabra et al., 2023; Naous et al., 2023). España-Bonet and Barrón-Cedeño (2022) showed that multilingual neural models run short at exhibiting cultural biases that are inherent to members of different cultures. Their experiments on 9 languages —including Spanish— show that monolingual static embeddings exhibit human biases at different levels across languages and that multilinguality attenuates further the effect of the bias. For properly inclusive NLP, cultural biases should be taken into account.

Languages can evolve into several dialects or varieties because of being spoken in different countries; this is a clear case where varieties reflect cultural differences and it is the case for Spanish. Treating it as a single unit dilutes these cultural nuances. Collecting data in order to have all the cultures equally represented is one of the strategies identified by Haemmerl et al. (2023) for truly cross-cultural NLP.

3 Country-Specific Spanish Texts

CEREAL Corpus. We use Open Super-large Crawled Aggregated coRpus (OSCAR) version 22.01, a multilingual corpus obtained by filtering Common Crawl (Ortiz Suárez et al., 2019; Abadji et al., 2021). We download the dump for Spanish, with data from all Spanish-speaking countries and extract for each document its unique ID and source URL. We use the country code top-level domain (ccTLD) from the URL to label the texts according to the country it belongs to (see Figure 1). For documents without a ccTLD, we use the tag `mix`. We do not process the text in any way, other than marking paragraphs with tag <NS>.

Table 1 shows statistics. After deduplication, we obtain 41 M documents, 33% of which include a ccTLD. Still, 28 M documents have an unknown origin. It is worth noting that almost half of the documents with ccTLD come from Spain. The rest are distributed across the other 23 countries, predominantly in Latin America. Corpora like CORPES and NOW try to keep the proportion between Spain and Latin America closer to the number of speakers, but when resorting to the available digital data to create a corpus this is not the case. Argentina, Chile, Mexico and Spain all have more than 1 M documents. An accurate identification of the origin of the remaining 28 M documents would allow these varieties to have a representation comparable to languages such as Arabic, Greek or Korean in OSCAR.[6] This is the goal in Section 4, where we introduce our **CEREALex corpus** with automatic annotations for country of origin.

Segment-Level CEREAL. In order to deal with NLP tasks that do not need document-level data, we produce a corpus at the segment level using the <NS> tags for document segmentation. Deduplication reduces the size of the corpus from 391 M to 151 M sentences for the original data with country information and from to 916 M to 337 M for the part without origin information. Details in Table 1.

3.1 Stylistic Analysis

Different stylistic metrics are computed to infer the background of a person/group of people behind text, such as gender (Rangel and Rosso, 2019), native language (Volansky et al., 2013) or dialect (Lui and Cook, 2013; van der Lee and van den Bosch, 2017). In the same vein, we compute a number of such features to give light on whether the commonalities and divergences across variations are generally reflected in large volumes of data. We select a random sample of up to 30 k documents per variety (see in Table 1 varieties with less docu-

[6] https://huggingface.co/datasets/oscar-corpus/OSCAR-2201

ments), and compute 20 features that account for basic statistics and text complexity including textual richness and readability.[7]

Table 2 shows 9 of those features for representatives from the linguistic zones in Figure 1. Appendix A shows all 20 metrics for all dialect variations. As the absolute numbers of tokens reflect, there is a tendency for both Cuban and Peninsular Spanish documents to be longer: both 700+ tokens (45 sentences) long on average. Still, the difference with respect to other varieties is not significant. Colombian Spanish is the one with the shortest ratio of punctuation marks and also the one with the shorter documents in terms of tokens. It also has the largest word length, which is typically associated to a higher conceptual complexity. The rest of the statistics metrics, including word lengths, converge into very similar values for all varieties.

To assess the richness of vocabulary, we compute the type–token ratio, the ratio of hapax legomena and dislegomena, Shannon's entropy (Shannon, 1948) and Yule's K (Yule, 1944). Generally-large type-token ratios (∼50 on average) and the abundance of hapax legomena (36+) indicate rich texts where the authors use a varied vocabulary in all cases, with Bolivian Spanish lying below. Still, high values of these metrics can be an artefact of short texts. Shannon's entropy and specially Yule's K are length-independent and show small variations across varieties. According to these measures, Colombian is the least vocabulary-rich variety. The Szigriszt-Pazos index (Szigriszt Pazos, 1992) —an adaptation of the Flesch index for English (Kincaid et al., 1975) for Spanish— allocates all varieties into a "standard" level (Szigriszt Pazos, 1992, p. 265), with the lowest values for the Central American varieties reflecting higher levels of difficulty and the largest value for Equatorial Guinea, positioning it as the least difficult to read.

Finally, we compare two graph-based complexity measures: normalised Laplacian energy (Cavers et al., 2010) and average clustering coefficient (Newman, 2010). As with previous metrics, differences among varieties are not statistically significant, with Peninsular Spanish in the highest range of complexity and Colombian in the lowest.

4 Classification of Long Documents

The documents in OSCAR have an average length of 900 words, significantly surpassing the 512 token limit from most LM-based classifiers which, despite this limit, achieve state-of-the-art results in many classification tasks. However, the differences across language varieties might not be present in short fragments, making it crucial to consider the whole contents of a document, regardless of length. Therefore, we implement `docTransformer`,[8] a multi-class classifier on top of XLM-RoBERTa large (Conneau et al., 2020) that considers the whole document, as shown in Figure 2. The extension to document-level is simple: the `[CLS]` token is indeed an average of the `[CLS]` tokens for each of the n fragments in the document. This is done online, which allows for the full document to take part in the classification and the backpropagation. The best practice to split the document is explored in the following.

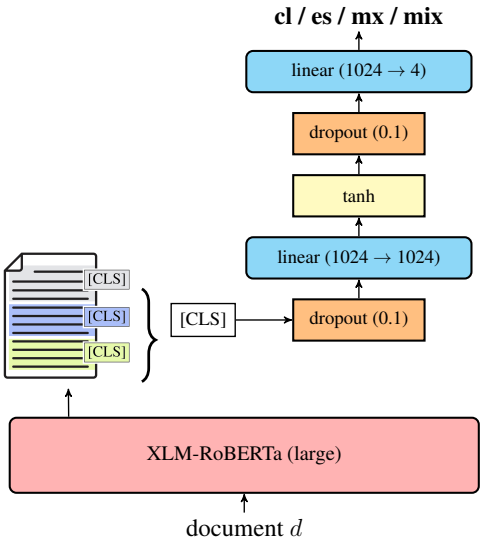


Figure 2: `docTransformer`: fine-tuning for multi-class classification. After encoding, d is split into n parts and the average of the `[CLS]` token for each part is used for the classification head.

Experimental Settings. We conduct 3 tasks:

3C 3 classes: `es`, `mx` and `mix`
4C 4 classes: `cl`, `es`, `mx` and `mix`
5C 5 classes: `ar`, `cl`, `es`, `mx` and `mix`

The tasks consider the top-k languages in terms of number of documents (see Table 1). We select balanced subsets of the `CEREAL` corpus for training and validation. First, we consider documents with lengths ranging from 80 to 3 k words and randomly select 500 k documents per class (`ar`, `cl`, `es`, `mx`). To build class `mix`, we select up to 30 k documents

[7] https://github.com/cristinae/stylometrics

[8] https://github.com/cristinae/docTransformer

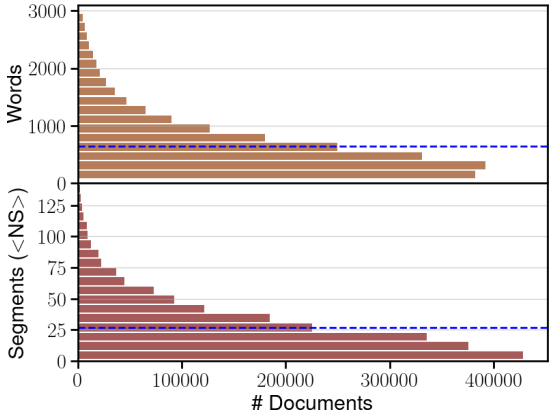


Figure 3: Statistics for the training documents in task 4C. The average is indicated with a dashed line.

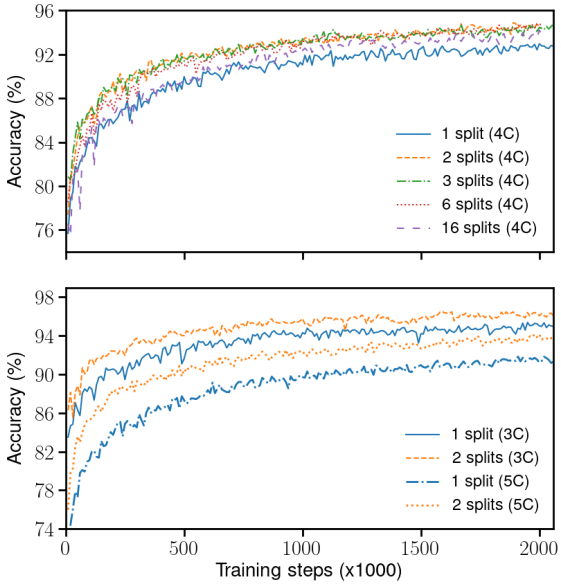


Figure 4: Evolution of the accuracy in validation over 2 M training steps for the three tasks 3C, 4C and 5C.

from each variety not included in the other classes, summing up to 510 k documents. We build the validation set similarly, with 1 k documents per class and 63 per language variety in `mix`.

For the test set, we use a subset of the corpus with unknown origin. To assign the origin, we collect sports, fashion, travel, gastronomy and news sites without a country identifier in the URL for the 24 varieties, and extract the documents available in `CEREAL` with the `mix` class. We build datasets with the same proportions as the validation sets for the 3 classification tasks. Table 1 shows statistics about the training, validation and test sets. Appendix B lists the sites used to create the test set.

Figure 3 depicts the data distribution for the 4C training set. Notice that only half of the documents fit the 512 token limit. The average number of segments is 27, but a few surpass 100. Both length and number of segments are relevant for the classifier.

Training Models. We train the classifiers along three iterations for 3C and two iterations for 4C and 5C to explore a similar number of instances in the three cases. We first explore the relevance of going beyond the 512 input token limit of RoBERTa with 5 models in 4C. After encoding, we split an input document d into 2, 3, 6, or 16 parts. The baseline is a single split with the first 512 tokens. With the 2-splits model, one considers up to 1,024 tokens, with 3 splits up to 1,536, and so on. We cover fully every d in the training set with 6 splits (max. 3,072 tokens per document). With the model with 16 splits, we are interested in checking if the length of the input fragments is relevant: the model with 6 and the model with 16 splits cover the same amount of text because the documents are limited to 3,000 words, but in the latter the `[CLS]` tokens summarise information of less context (specifically, max. 62 words vs 167 in the 6-splits model).

The splitting of d is done at the fragment level using the <NS> tags. If d has less or the same number of segments as the desired splits, each split contains a segment. If d has more divisions, we balance the number of segments across splits. If d is longer than the number of splits allowed, we first balance the number of segments and then these fragments are limited to the 512 limit each, so information is lost homogeneously across d and not gathered at the end, as the baseline system would do.

Figure 4 (top) shows the performance evolution for the 5 models for 4C as measured by accuracy. Table 3 shows the final accuracy per model. The model with only 2 splits is systematically 2 points above the baseline all along the training. At the beginning of the training, models with more splits are always below the 2-splits version, specially those with 6 and 16 splits, but towards the end of the training all the document-level models tend to converge. Increasing the number of splits is somehow equivalent to increasing the batch size, more text is seen at every batch. So, for small corpora with long documents, the document-level extension is always preferred. For large corpora but with relevant information all along the document, as it is our case, the extension also shows significant improvements. Given the length of our documents, just 2 splits are enough to significantly improve over the baseline; splitting further the data is not

# Splits	Validation	Test
1	93.17±0.24	87.43±0.33
2	94.79±0.08	89.65±1.06
3	94.69±0.07	87.63±0.32
6	94.76±0.04	86.82±0.45
16	94.34±0.13	89.29±1.44

Table 3: Sampling mean of the accuracy (%) over 3 runs with 95% CI standard error for the 5 models in task 4C.

	3C		5C	
	1 split	2 splits	1 split	2 splits
Val.	95.2±0.2	96.5±0.1	92.3±0.3	94.3±0.1
Test	90.3±1.0	89.0±1.4	86.6±0.6	88.8±0.8

Table 4: Accuracy (%) over 3 runs on validation and test for the models trained on the 3C and 5C tasks.

always beneficial in test. Notice that the model with 2 splits fully covers around 75% of the documents.

After these observations, we evaluate the 3C and 5C tasks on the baseline and the 2-split model. We experiment a drop in performance in test (Table 4). Confusion matrices (included in Appendix C) show that the loss in quality comes because some of the Mexican articles are miss-classified as `es` or `mix`.[9]

`CEREALex`, Classifying the Unknown. Finally, we apply the 3C, 4C and 5C classifiers with 2 splits to the 28 M documents of the *unknown* part of `CEREAL`. Table 1 (Classified column) reports the number of classified documents per task and class. We create `CEREALex` determining the ultimate classification through a majority vote across the three tasks in case of intersection of classes. That is, all `ar` articles are considered, `cl`, `es` and `mx` are considered as such if they appear in two classification results, otherwise `mix` is kept. The final corpus contains 36 M documents with annotated origin.

4.1 Salient Phrases per Variety

We use Layer Integrated Gradients (Sundararajan et al., 2017) to determine which words or phrases contribute more to the classification. We compute the attribution of each subword with the integrated gradients and define the attribution of a word as the sum of its subunits. We consider the 1D distribution of attributions along the words of a document after removing punctuation, extract the words on local maxima which surpass a threshold and build phrases in case the neighbouring words (±2 in our experiments) also surpass a threshold. After manual inspection of individual documents, we establish the threshold for identifying the head of a phrase as the 98% percentile and lower it to 68% for neighbouring words.

At the document level, considering phrases is informative. However, at the corpus level the addition of phrases increases the sparsity, as it is more difficult for phrases to be common in several documents. For example, a document about the Mexican dish `mole` has as the highest attributed phrases: `'El'`, `'herencia gastronómica'`, `'República Mexicana es el'`, `'Mazateca El'` and `'Cañada El'`. That is, an article, the keyword that detects that the topic is food, and 3 Mexican locations in this order. In the four occasions the article `El` is highlighted, it comes before word `mole`, showing that the classifier focuses on the left context of the keyword. A document about `alebrijes` (Mexican sculptures of fantastical creatures) has: `'México'`, `'viernes al Rockefeller Center'`, `'encomienda'`, `'Semana'`, `'coincidirá'`, `'mexicana Las'` and `'México el'`. In this case, besides locations, we have two words (`encomienda` and `coincidirá`) that might be used in a specific way in the Mexican variety. `'Semana de México'` is not extracted as a phrase but as two words because the attribution to `de` is very low; this is common to lots of other prepositional phrases. It is interesting that the word with the lowest (negative) attribution is `España`, as the term appears as a positive clue in the `es` class. These examples are fully depicted in Appendix C.3, together with a list of top attributed phrases.

At the corpus level, we look at the top attributed words/phrases per class in the test set. In the 4C experiment we obtain mostly stopwords (e.g., `que`, `en`, `de`, `por`) and locations (e.g., `Chile`, `España`, `Chiapas`, `Nicaragua`). Other than that, we observe "Clever Hans" by the classifier, as it seems to learn boilerplate text common to a class (e.g., `'Inicio Nacionales'` can be the header of a newspaper site) which cannot explain differences among varieties.

5 Cultural Aspects in Embeddings

We analyse the quality and effect of variety-specific word embeddings built from the `CEREAL` corpus. We estimate fastText (Bojanowski et al., 2017) em-

[9] We consider it an effect of using articles from mx.hola.com. *Hola* is a magazine based in Spain with several other editions in Latin America. Some articles might therefore be written in/by Spanish journalists and shared across editions.

beddings as described in Appendix D. For bilingual lexicon induction, we use fastText pretrained English embeddings.[10] For comparison purposes, we use the Twitter embeddings from Tellez et al. (2023).

5.1 Bilingual Lexicon Induction (BLI)

We study the relevance of the underlying embeddings to induce English–Spanish lexicons for 21 Spanish varieties. At the heart of BLI are crosslingual word representations. We do not focus here on the method to build the bilingual space, but use a standard VecMap method (Artetxe et al., 2018)[11] to construct the spaces. Both for our and the Twitter models, we do the mappings between the English and the 21 Spanish embedding spaces with VecMap, using a semi-supervised configuration that uses identical words as seed dictionary.

Data Settings. We build the evaluation of bilingual dictionaries grounded on VARILEX-R, a database with linguistic data for 21 Spanish varieties (Chacón García, 2016; Ueda and Moreno Fernández, 2016). Linguistic concepts are instantiated by words or phrases in all the varieties, as many as the participants to the project provided by solving different kinds of cloze tests. All concepts are aligned to an English word or phrase. Since we work with word embeddings, we consider only single-token entries, eliminating articles and specifications (e.g., `el pijama` (Spain) and `la pijama` (Mexico) both turn into `pijama` and `palomitas (de maíz)` becomes `palomitas`). Appendix E details statistics of the final resource.

Experiments. We evaluate the quality of the embeddings with respect to the accuracy on the VARILEX BLI test using CSLS retrieval as implemented by VecMap. Figure 5 shows the results for the languages with the largest training sets, those that have been enlarged with our classifier —Argentinian, Chilean, Mexican, and Spanish—, and for the combination of the 24 varieties (`all`). The diagonal marks the cases where we use the embedding space of the same variety of the dictionary we want to induce. One would expect the diagonal to have the highest values for a given dictionary, and this is always the case with `CEREALex` even when considering all the data together to build the embeddings. With `CEREAL` and `Twitter`, which

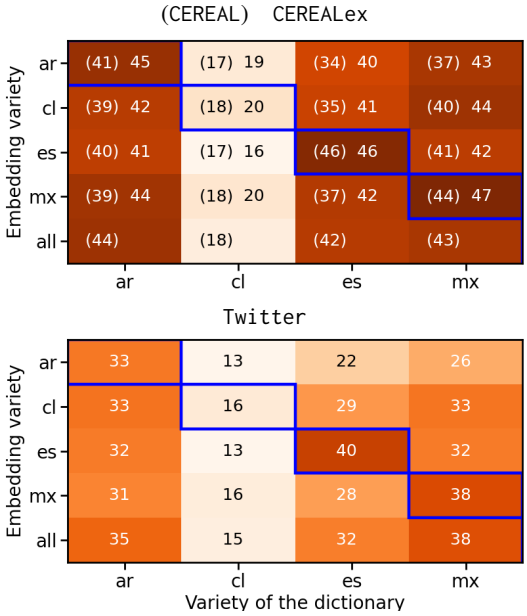


Figure 5: Accuracy (%) in BLI for the high-resourced varieties for the 3 embedding models used. The top heatmap shows results for `CEREAL` (in parenthesis) and for `CEREALex`.

use less training data per variety, this is not always the case and using the combined corpus instead of the corpus per variety does not represent a bad alternative. The conclusions are in line for `CEREAL`, `CEREALex` and `Twitter`, but the performance using embeddings from Twitter is notably diminished, despite the dediacritisation of the vocabulary.

Note that Chilean aligns with the same trends as the other varieties, but at a lower performance. Consulting with Chilean citizens regarding the quality of the dictionary, we learned that its vocabulary might not be used globally, but may reflect the linguistic preferences of a specific localised or minority group. In general, lower values column-wise might indicate a lower quality of the dictionary, while lower values row-wise might indicate a low quality of the embeddings. This trend is observed when we consider all the varieties and do the 22x21 comparison (see the accuracies in Appendix E).

5.2 Human Biases in Embeddings

We resort to the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) as a measure for cultural biases in word embeddings. WEAT is inspired by the Implicit Association Test (IAT) in psychology (Greenwald et al., 1998), which measures differences in response time when subjects are requested to pair items and attributes that they find similar and when pairing items and attributes

[10] https://fasttext.cc/docs/en/crawl-vectors.html

[11] https://github.com/artetxem/vecmap

	es	mx	combo
CEREAL-CAWEAT1	$1.47^{+0.17}_{-0.30}$	$1.08^{+0.30}_{-0.81}$	$1.19^{+0.22}_{-0.50}$
CEREAL-CAWEAT2	$1.39^{+0.15}_{-0.44}$	$1.27^{+0.29}_{-0.46}$	$1.50^{+0.08}_{-0.42}$
CEREALex-CAWEAT1	$1.50^{+0.08}_{-0.38}$	$1.10^{+0.28}_{-1.00}$	–
CEREALex-CAWEAT2	$1.39^{+0.10}_{-0.54}$	$1.31^{+0.27}_{-0.70}$	–
Twitter-CAWEAT1	$1.29^{+0.10}_{-0.39}$	$0.93^{+0.41}_{-1.06}$	$1.09^{+0.26}_{-0.55}$
Twitter-CAWEAT2	$1.01^{+0.33}_{-0.38}$	$1.16^{+0.35}_{-0.40}$	$1.25^{+0.29}_{-0.19}$

Table 5: Size effect (median and 95% CIs) of human biases as measured by CA-WEAT1 and CA-WEAT2, in the Mexican (mx), Peninsular Spanish (es) and the union of 24 variants embeddings. In the latter we use CA-WEAT lists for 5 variants (combo).

that they find different. For word embeddings, one does not measure response times, but distances in embedding spaces. Only two of the WEAT tests available represent human biases: one measuring associations between flowers/insects and pleasant/unpleasant attributes (WEAT1) and one measuring associations between musical instruments/weapons and pleasant/unpleasant attributes (WEAT2). The other tests represent social biases, such as sexism or racism, and we cannot expect them to be comparable across cultures. The seminal work by Caliskan et al. (2017) provides lists in English for WEAT1 and WEAT2 among other tests. Here, we follow the approach in (España-Bonet and Barrón-Cedeño, 2022) to collect cultural aware lists (CA-WEAT) for the language varieties under study as explained below.

Data Settings. We ask 30 volunteers from Bolivia, Colombia, Cuba, Ecuador, Mexico and Spain to provide lists with 25 flowers, insects, weapons, instruments, and both pleasant and unpleasant concepts in Spanish in analogy to the original CA-WEAT lists. After cleaning, we discard 4 lists with noise or missing words and keep 12 Mexican, 9 Spanish, 2 Colombian, 2 Ecuadorian and 1 Bolivian. In order to deal with lists which only problem is a word —either missing, repeated or out of context—, we implement and apply a matrix factorisation model for collaborative filtering (Takács et al., 2008) whose objective is recommending the most likely missing item for a given native speaker on the basis of the full decision history by all volunteers, regardless of their background. After this cleaning process, we have many more lists in Mexican and Peninsular Spanish than in the other varieties. Therefore, we group the lists into three sets: 12 Mexican (mx), 9 Spanish (es) and a combination with 2 Mexican, 2 Spanish, 2 Colombian, 2 Ecuadorian and 1 Bolivian (combo).

Experiments. We calculate the effect size of the bias as explained in Appendix F in our variety-specific and in the Twitter embeddings (es, mx and all models). For the statistical analysis of the results, we provide the median over the lists of a variety and 95% confidence intervals (CI) using order statistics. As Table 5 shows, the observed biases are not universal and they do depend on both the domain of the data (embeddings) and the language variety. In general, biases in Common Crawl are stronger than in Twitter. Results with CEREAL and CEREALex are not statistically significantly different. If we look at the language dimension, biases in the mx embeddings are less pronounced than in es embeddings except for the CA-WEAT2 case on Twitter (instruments and weapons). When we put all the varieties together (the all embedding model) and use the combo set, no straightforward trend is observed. This is an indication that when joining different varieties together, biases or cultural aspects can be erased, amplified or mixed.

6 Conclusions

Spanish is a language with large variations across territories. Variations are reflected in written documents as the high accuracies of our classifiers show. The amount of data we gather in CEREAL and CEREALex turns Argentinian, Chilean, Colombian, Mexican and Peninsular Spanish into high-resourced varieties that do not need extra data to achieve competitive results in tasks such a bilingual lexicon induction. For other varieties, the addition of data from the richest counterparts is still needed.

Ongoing work is being devoted to study the topology of our variety-specific embedding spaces and to derive data-based phylogenetic trees for Spanish. Although currently the amount of documents in CEREAL that remains with unknown origin is relatively small (5 M documents out of 41 M), we plan to train classifiers for the lowest-resourced varieties. Beside creating larger corpora, further work should also focus on the development of variety-specific curated resources for NLP tasks evaluation.

Corpora, code and models are available for the research community in hope for fostering research and development of technology that reflects cultural variety beyond single languages.

Limitations

In this work we build a corpus from the Web. This inherently introduces a technological bias to the corpus, as it excludes social groups with limited access to technology. The disparity among Spanish-speaking countries in this aspect becomes apparent with the lack of a correlation between a country's population and the amount of online text.

An unintended bias in the analysis might appear due to the origin of the authors. The authors are speakers of Peninsular and Mexican Spanish. We also consulted with natives from Bolivia and Chile. Citizens from other countries have not been involved in the project.

The classification method that we use for Spanish would also work for other languages such as English, French, Portuguese or German, but it is not a generic method. Our goal in this work is to provide a high-quality corpus for several Spanish varieties. Within each variety there are also differences which are not tackled here, but we hope this corpus is useful to dig into large cultural differences.

Some of the experiments performed depend on the quality of additional resources such as bilingual lexicons and CA-WEAT lists. These resources face the same problems as the corpus itself, data is more challenging to collect for certain cultures compared to others. While low-quality issues in bilingual dictionaries were identified only in Chilean, the number of CA-WEAT lists is only significant for Mexican and Peninsular Spanish. Because of this, the distribution of the effect size over the lists is not normal. Some lists are outliers either because of non-common words in the variety, typos or multiword expressions which translate into OOVs. Having more lists at least for the higher-resourced varieties (i.e. those with good embeddings) would help in confirming the conclusions.

Acknowledgements

This work has been supported by the LT-Bridge Project (GA 952194) and by the German Federal Ministry of Education and Research (BMBF) under funding code 01IW20010 (CORA4NLP).

The authors are grateful to Roberto Asín for helping in the interpretation of the experiments with Chilean data.

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. In *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.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. DocBERT: BERT for document classification. *arXiv preprint arXiv:1904.08398*.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):5012–5019.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

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.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Michael S. Cavers, Shaun M. Fallat, and Steve Kirkland. 2010. On the normalized Laplacian energy and general Randić index R-1 of graphs. *Linear Algebra and its Applications*, 433:172–190.

Carmen Chacón García. 2016. *Análisis demolingüístico del léxico hispánico: estudio aplicado a las nociones específicas del plan curricular del Instituto Cervantes*. Phd thesis, UNED. Universidad Nacional de Educación a Distancia, Madrid, Spain.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. CogLTX: Applying BERT to long texts. In *Advances in Neural Information Processing Systems*, volume 33, pages 12792–12804. Curran Associates, Inc.

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.

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

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, 74(6):1464–1480.

Katharina Haemmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, Toronto, Canada. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Richard A. Hudson. 1996. *Sociolinguistics*, 2nd edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel. Technical report, Institute for Simulation and Training, University of Central Florida.

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.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15, Brisbane, Australia.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having beer after prayer? Measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Mark Newman. 2010. *Networks: An Introduction*. Oxford University Press, Oxford.

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

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical Transformers for Long Document Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Hyunji Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 702–709, Dublin, Ireland. Association for Computational Linguistics.

Real Academia Española RAE. 2023. Corpus del Español del Siglo XXI (CORPES). [online] Accessed on 27.11.2023.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Francisco Rangel and Paolo Rosso. 2019. Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in Twitter. In *Working Notes of CLEF 2019–Conference and Labs of the Evaluation Forum*, CLEF '2019, Lugano, Switzerland.

Sonal Sannigrahi, Josef van Genabith, and Cristina España-Bonet. 2023. Are the best multilingual document embeddings simply based on sentence embeddings? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2306–2316, Dubrovnik, Croatia. Association for Computational Linguistics.

Shlomo S. Sawilowsky. 2009. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Francisco Szigriszt Pazos. 1992. *Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad*. Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain.

Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. 2008. Matrix factorization and neighbor based algorithms for the Netflix Prize problem. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, page 267–274, New York, NY. Association for Computing Machinery.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

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.

Hiroto Ueda and Francisco Moreno Fernández. 2016. VARILEX-R: Variación léxica en español del mundo / Datos revisados. [online] Accessed on 4.12.2023.

Chris van der Lee and Antal van den Bosch. 2017. Exploring lexical and syntactic features for language variety identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 190–199, Valencia, Spain. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

C. Udny Yule. 1944. *The statistical study of literary vocabulary*. Cambridge University Press.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

A Style Characteristics in Spanish Varieties

We report the scores given by 20 style-related metrics in Tables 6 and 7. Table 6 gathers the results related to the basic text statistics and Table 7 to more complex measures for text richness and readability.

B Multi-Variety Test Sets

Table 8 shows the websites included in the test sets used in the three classification tasks: 3C, 4C and

	Number of sents.	Number of tokens*	Ratio of digits	Ratio of upper-case	Ratio of punct.	Ratio of function words	Length of the sents. (words)	Length of the tokens (chars.)	Syllables per word	Ratio of * words short (<4chars)	Ratio of * words long (>2syl)
ar	34 ± 30	578 ± 479	1.5 ± 2.7	4 ± 2	2.3 ± 0.9	10.0 ± 2.3	20 ± 9	5.1 ± 0.4	2.04 ± 0.15	42 ± 5	32 ± 5
bo	39 ± 31	672 ± 496	1.2 ± 1.3	4 ± 3	2.3 ± 0.9	9.7 ± 2.2	20 ± 9	5.2 ± 0.4	2.05 ± 0.13	42 ± 3	32 ± 5
cl	37 ± 31	621 ± 520	1.1 ± 1.8	4 ± 2	2.3 ± 0.7	9.9 ± 2.3	20 ± 10	5.2 ± 0.4	2.08 ± 0.16	41 ± 4	33 ± 6
co	16 ± 36	248 ± 531	0.5 ± 1.8	2 ± 2	1.5 ± 0.8	9.9 ± 2.4	16 ± 13	5.6 ± 0.5	2.14 ± 0.17	42 ± 4	36 ± 6
cr	35 ± 29	575 ± 471	1.1 ± 1.4	4 ± 2	2.3 ± 0.7	9.9 ± 2.2	19 ± 19	5.3 ± 0.4	2.11 ± 0.17	41 ± 4	34 ± 6
cu	45 ± 34	772 ± 596	2.0 ± 3.4	3 ± 2	2.5 ± 1.0	9.5 ± 2.2	20 ± 9	5.3 ± 0.4	2.10 ± 0.16	41 ± 4	34 ± 6
do	31 ± 25	541 ± 394	1.1 ± 1.1	4 ± 2	2.3 ± 0.7	9.7 ± 2.0	21 ± 9	5.2 ± 0.3	2.06 ± 0.13	42 ± 3	33 ± 5
ec	33 ± 27	522 ± 407	1.5 ± 2.8	4 ± 3	2.3 ± 0.7	9.6 ± 2.3	19 ± 14	5.3 ± 0.4	2.09 ± 0.18	41 ± 5	33 ± 6
es	45 ± 37	702 ± 556	1.3 ± 2.6	3 ± 2	2.4 ± 0.9	9.8 ± 2.3	18 ± 18	5.2 ± 0.4	2.06 ± 0.16	42 ± 5	32 ± 6
gq	23 ± 17	429 ± 322	0.8 ± 0.7	3 ± 3	2.2 ± 0.6	9.8 ± 2.0	20 ± 6	5.1 ± 0.3	2.04 ± 0.10	41 ± 4	28 ± 5
gt	34 ± 30	560 ± 464	1.5 ± 2.8	4 ± 2	2.3 ± 0.8	9.9 ± 2.3	19 ± 11	5.2 ± 0.4	2.09 ± 0.15	41 ± 4	33 ± 5
hn	31 ± 27	586 ± 513	1.0 ± 0.8	3 ± 2	2.2 ± 0.5	10.1 ± 1.9	22 ± 9	5.1 ± 0.4	2.03 ± 0.15	43 ± 3	32 ± 5
mx	33 ± 29	594 ± 477	1.1 ± 1.9	3 ± 2	2.3 ± 0.7	9.9 ± 2.2	22 ± 13	5.2 ± 0.4	2.05 ± 0.15	42 ± 4	32 ± 6
ni	27 ± 24	520 ± 431	1.6 ± 2.2	4 ± 2	2.2 ± 0.6	9.8 ± 2.1	23 ± 16	5.2 ± 0.4	2.07 ± 0.14	41 ± 4	33 ± 5
pa	35 ± 28	613 ± 497	2.4 ± 4.0	5 ± 10	2.3 ± 0.8	9.3 ± 2.5	25 ± 49	5.3 ± 0.4	2.08 ± 0.16	40 ± 5	34 ± 6
pe	35 ± 30	562 ± 466	1.3 ± 1.9	4 ± 3	2.4 ± 0.7	9.6 ± 2.2	19 ± 12	5.3 ± 0.4	2.08 ± 0.16	41 ± 4	33 ± 6
ph	36 ± 24	770 ± 509	0.6 ± 0.8	3 ± 3	2.0 ± 0.7	9.8 ± 2.1	22 ± 7	5.1 ± 0.3	2.05 ± 0.12	42 ± 4	30 ± 5
pr	25 ± 22	484 ± 369	1.0 ± 0.9	3 ± 1	2.1 ± 0.6	9.8 ± 2.0	22 ± 9	5.1 ± 0.3	2.01 ± 0.11	43 ± 3	30 ± 4
py	30 ± 27	527 ± 458	1.2 ± 1.5	4 ± 3	2.3 ± 0.7	9.7 ± 2.2	20 ± 10	5.2 ± 0.4	2.07 ± 0.15	42 ± 4	33 ± 5
sv	27 ± 26	476 ± 418	1.3 ± 1.7	4 ± 3	2.1 ± 0.8	9.6 ± 2.1	22 ± 11	5.3 ± 0.5	2.09 ± 0.16	41 ± 6	34 ± 6
us	50 ± 33	842 ± 585	0.9 ± 1.0	3 ± 2	2.2 ± 0.6	10.3 ± 2.3	20 ± 18	5.1 ± 0.4	2.02 ± 0.14	42 ± 4	30 ± 5
uy	39 ± 36	623 ± 542	1.2 ± 2.0	3 ± 2	2.4 ± 0.9	10.2 ± 2.3	19 ± 10	5.2 ± 0.4	2.06 ± 0.17	42 ± 4	32 ± 6
ve	36 ± 31	611 ± 519	1.2 ± 1.7	4 ± 3	2.4 ± 0.8	9.7 ± 2.2	20 ± 13	5.3 ± 0.4	2.09 ± 0.15	41 ± 4	34 ± 6

* Without punctuation or digits

Table 6: Stylistics 1. Text statistics used for the stylistic analysis. All numbers are averaged over (up to) 30 k documents per country.

	TTR	Ratio of hapax legomena	Ratio of hapax dislegomena	Entropy	Yule's K	Fernandez Huerta	Szigriszt-Pazos /INFLEZ	Laplacian energy	Clustering
ar	51 ± 10	37 ± 11	7 ± 3	6.9 ± 0.7	152 ± 59	63 ± 13	59 ± 13	0.46 ± 0.05	0.08 ± 0.04
bo	46 ± 12	31 ± 15	8 ± 4	7.0 ± 0.6	159 ± 60	62 ± 13	58 ± 13	0.45 ± 0.05	0.08 ± 0.03
cl	50 ± 11	36 ± 12	7 ± 3	7.0 ± 0.7	155 ± 65	61 ± 14	56 ± 15	0.46 ± 0.05	0.08 ± 0.04
co	53 ± 12	36 ± 13	9 ± 3	6.5 ± 0.7	181 ± 69	59 ± 16	54 ± 16	0.47 ± 0.06	0.05 ± 0.04
cr	50 ± 11	36 ± 11	7 ± 3	6.9 ± 0.7	162 ± 58	59 ± 22	54 ± 22	0.46 ± 0.05	0.08 ± 0.04
cu	50 ± 10	38 ± 11	6 ± 2	7.2 ± 0.6	149 ± 68	61 ± 12	56 ± 12	0.46 ± 0.06	0.08 ± 0.03
do	52 ± 9	38 ± 10	7 ± 2	7.0 ± 0.6	146 ± 44	60 ± 11	56 ± 11	0.46 ± 0.04	0.08 ± 0.03
ec	51 ± 10	38 ± 11	7 ± 3	6.9 ± 0.7	155 ± 55	61 ± 17	56 ± 17	0.46 ± 0.05	0.08 ± 0.04
es	48 ± 11	34 ± 11	7 ± 3	7.0 ± 0.7	155 ± 67	63 ± 20	59 ± 20	0.45 ± 0.06	0.09 ± 0.04
gq	52 ± 9	38 ± 10	6 ± 2	6.8 ± 0.4	145 ± 47	65 ± 9	61 ± 9	0.47 ± 0.04	0.07 ± 0.03
gt	51 ± 10	37 ± 10	7 ± 2	6.9 ± 0.7	152 ± 53	62 ± 14	57 ± 14	0.46 ± 0.05	0.08 ± 0.03
hn	52 ± 10	39 ± 10	7 ± 2	7.0 ± 0.7	142 ± 39	61 ± 12	57 ± 12	0.46 ± 0.05	0.07 ± 0.04
mx	50 ± 11	36 ± 12	7 ± 3	7.0 ± 0.6	149 ± 51	60 ± 15	56 ± 14	0.46 ± 0.05	0.08 ± 0.04
ni	53 ± 10	39 ± 11	7 ± 2	6.9 ± 0.7	149 ± 57	58 ± 18	53 ± 17	0.47 ± 0.05	0.07 ± 0.03
pa	49 ± 10	36 ± 10	7 ± 2	7.0 ± 0.7	156 ± 64	56 ± 55	52 ± 54	0.46 ± 0.05	0.08 ± 0.04
pe	50 ± 11	36 ± 11	7 ± 3	6.9 ± 0.7	155 ± 59	62 ± 16	57 ± 16	0.46 ± 0.05	0.08 ± 0.04
ph	47 ± 9	34 ± 9	6 ± 3	7.2 ± 0.7	135 ± 47	60 ± 10	55 ± 10	0.43 ± 0.05	0.08 ± 0.03
pr	52 ± 9	38 ± 9	7 ± 2	6.9 ± 0.6	148 ± 39	63 ± 11	58 ± 11	0.46 ± 0.05	0.08 ± 0.03
py	51 ± 10	37 ± 11	7 ± 3	6.9 ± 0.6	156 ± 60	60 ± 13	56 ± 13	0.46 ± 0.05	0.08 ± 0.04
sv	51 ± 9	36 ± 9	8 ± 3	6.7 ± 0.7	161 ± 68	59 ± 15	54 ± 15	0.46 ± 0.05	0.08 ± 0.04
us	45 ± 11	32 ± 11	6 ± 2	7.2 ± 0.6	147 ± 47	64 ± 21	59 ± 21	0.44 ± 0.05	0.09 ± 0.03
uy	50 ± 11	36 ± 11	7 ± 3	6.9 ± 0.7	154 ± 55	63 ± 13	59 ± 14	0.46 ± 0.06	0.08 ± 0.04
ve	51 ± 10	37 ± 11	7 ± 3	7.0 ± 0.7	152 ± 65	59 ± 15	55 ± 15	0.46 ± 0.05	0.08 ± 0.03

Table 7: Stylistics 2. Text richness measures used for the stylistic analysis. All numbers are averaged over (up to) 30 k documents per country.

Country	Freq.	Site	Country	Freq.	Site
ar	14	https://bilinkis.com	es	37	https://hemeroteca.vozlibre.com
ar	3	https://elle.clarin.com	es	18	https://losviajesdeclaudia.com
ar	353	https://elplanetaurbano.com	es	11	https://losviajesdedomi.com
ar	171	https://www.clarin.com	es	89	https://vozlibre.com
ar	3	https://www.fondodeolla.com	es	3	https://www.donquijote.org
ar	113	https://www.perfil.com	es	120	https://www.elplural.com
ar	5	https://www.poneteeldelantal.com	es	36	https://www.elrincondesele.com
ar	96	https://www.tycsports.com	es	2	https://www.gironafc.cat
ar	195	https://zuletasintecho.com	es	113	https://www.recetasderechupete.com
ar	47	http://turismo.perfil.com	es	571	https://www.telva.com
bo	63	https://www.lostiempos.com	gt	63	https://www.prensalibre.com
cl	25	http://labrujitadejengibre.blogspot.com	hn	4	https://contracorriente.red
cl	10	https://astroturismochile.travel	hn	59	https://www.tunota.com
cl	161	https://chile.as.com	mx	138	http://escrutiniopublico.blogspot.com
cl	28	https://ecochile.travel	mx	300	https://mx.hola.com
cl	43	https://puconchile.travel	mx	7	https://viajerosvagabundos.com
cl	145	https://rufianrevista.org	mx	300	https://www.diariodemexico.com
cl	35	https://tradenews.chile.travel	mx	154	https://www.motivosamarmx.com
cl	29	https://www.chile.travel	mx	59	https://www.reforma.com
cl	523	https://www.latercera.com	mx	11	https://www.tolucafc.com
cl	1	http://tradenews.chile.travel	mx	24	http://www.marieldeviaje.com
co	14	https://www.mycolombianrecipes.com	mx	7	http://www.yovivolamoda.com
co	49	https://www.semana.com	ni	63	https://www.el19digital.com
cr	63	https://www.nacion.com	pa	63	https://www.prensa.com
cu	52	https://diariodecuba.com	pe	63	https://www.rcrperu.com
cu	1	https://eldiariodecuba.com	py	63	https://www.ultimahora.com
cu	10	https://www.cuba.travel	sv	63	https://www.elsalvador.com
do	1	http://espacinsular.org	uy	1	http://decano.com
do	1	https://12y2.com	uy	37	https://lapalomadiariodigital.com
do	1	https://matense.net	uy	3	https://poracayporalla.com
do	1	https://www.alcarrizosdigital.net	uy	22	https://www.uypress.net
do	56	https://www.diariolibre.com	ve	1	https://es.aleteia.org
do	1	https://www.laprensatraslaverdad.com	ve	1	https://festiverd.com
do	1	http://www.alcarrizosdigital.net	ve	49	https://www.eluniversal.com
do	1	http://www.diariocristal.com	ve	11	https://www.venezuelatuya.com
ec	63	https://www.elcomercio.com	ve	1	http://www.grancine.net

Table 8: List of sites used to build the test set of documents without ccTDL.

5C. For `ar` and `cl`, we list the sites covering 1000 documents; when these varieties are used within the `mix` class, we use a subset of 63 documents. See Table 1 for the complete distribution across varieties.

C `docTransformer` Characteristics

C.1 Hyperparameters

The classifier is a small network on top of RoBERTa that first performs a dropout ($p = 0.1$) on the averaged [CLS] token, followed by a linear layer and a tanh activation function. After a second dropout layer ($p = 0.1$) a linear classification layer projects into the output classes. Back propagation takes place over the whole architecture.

We train the classifiers along three iterations for 3C and two iterations for 4C and 5C. All models have a batch size of 2 documents with gradient accumulation of 8. The batch size was chosen as the largest batch that fit our machines (NVIDIA A100 80GB) for the model with 16 splits. A training on 4 M documents (two iterations on the 4C corpus) takes 12 days in a single GPU. We use a cross-entropy loss, AdamW optimiser and a learning rate that decreases linearly starting at $5 \cdot 10^{-6}$. The learning rate was tuned in the range $10^{-5} - 10^{-6}$.

C.2 Performance and Confusion Matrices

Detailed evaluation performance of the classification models. Figure 6 shows the confusion matrices in validation and test for 4C, that is, the classification task in `cl`, `es`, `mx`, and `mix`. Notice that the main difference between the 2 sets (validation and test) is that in test some of the Mexican documents are classified as corresponding to the `es` or `mix` classes. The same effect is observed in Figure 7, where is evident that most of the lost in performance with respect of the validation data comes because of mis-classification of Mexican documents.

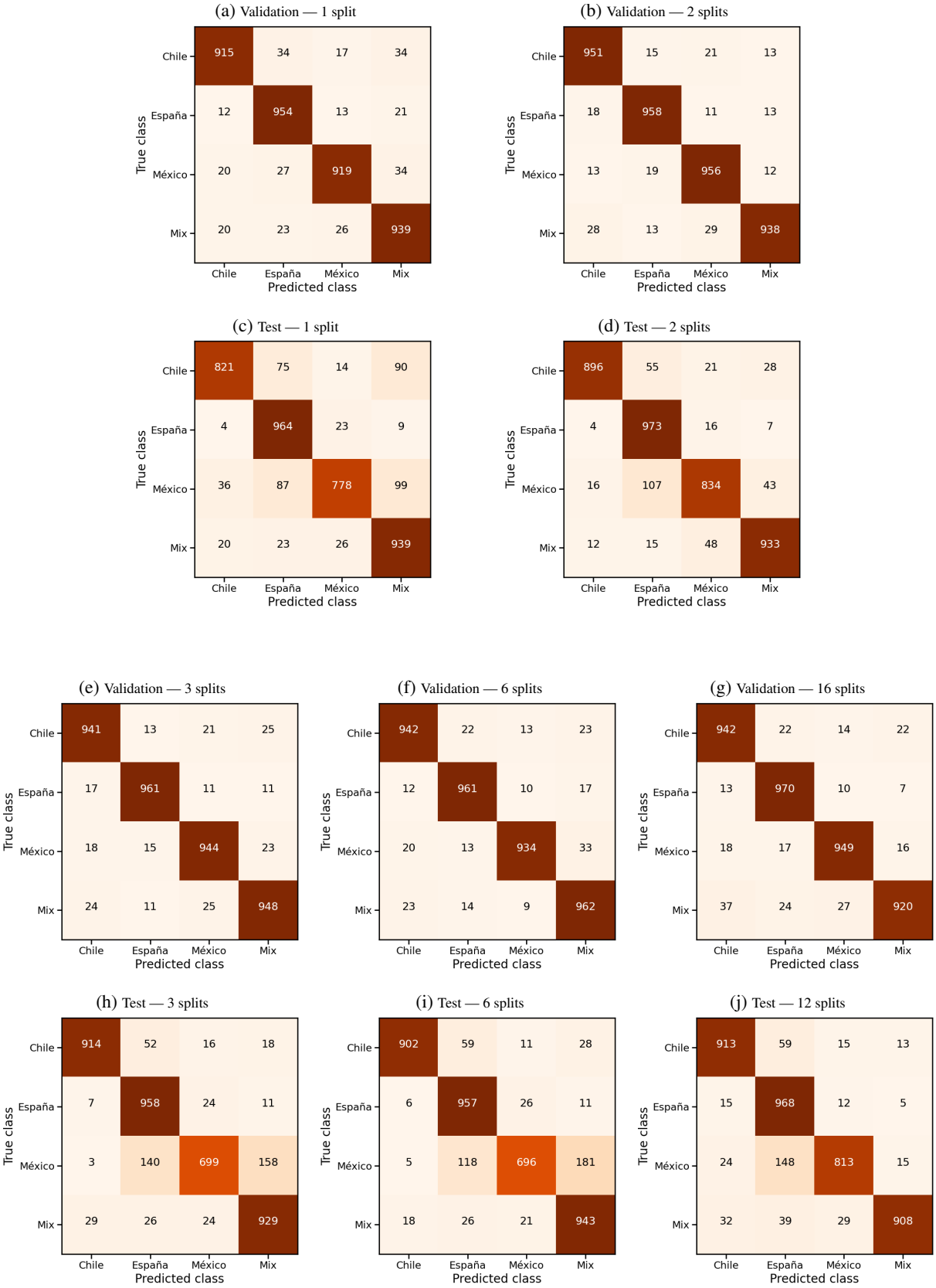


Figure 6: Confusion matrix for the validation and test sets. Test contains a combination of sports, fashion, travel, gastronomy and news sites as summarised in Table 8.

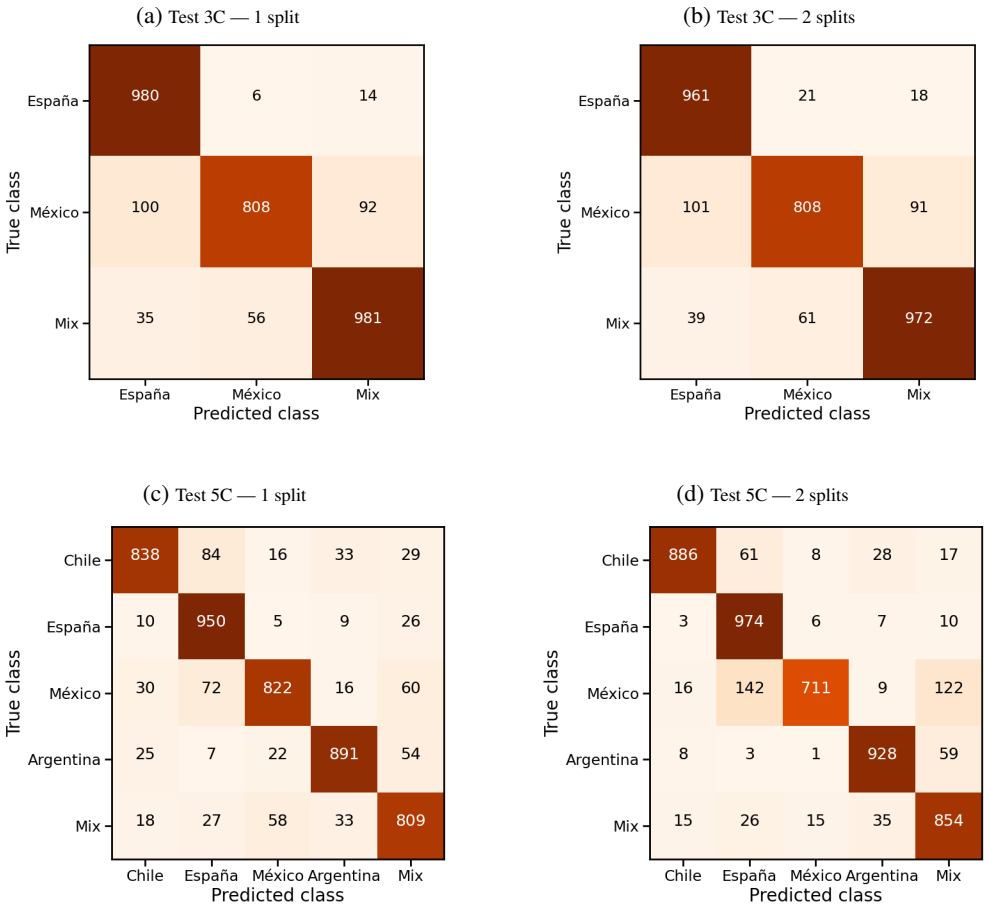


Figure 7: Confusion matrix for the 3C and 5C test sets. As in the 4C case, the test contains a combination of sports, fashion, travel, gastronomy and news sites as summarised in Table 8.

C.3 Explainability with Layer Integrated Gradients

Example documents (Figures 8 and 9) and corpus level top attributions (Table 9).

D Word Embeddings for Varieties

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. We use the default skipgram configuration in fastText (Bojanowski et al., 2017) to train 300-dimensional embeddings for tokens appearing 20+ times. We provide and make publicly available Spanish embeddings for 24 countries (we name the models after their ccTLD code) in addition to a model trained with all the varieties together (named `all`[12]). Our setting is comparable to (Tellez et al., 2023) except for the minimum frequency of in-vocabulary tokens (the default being 5 in their case) and the fact that we keep diacritics. For comparison purposes, we use their and our variety-specific embeddings in the experiments. The statistics of the resources are detailed in Table 10.

E Bilingual Lexicon Induction

We report in Table 11 the statistics for the VARILEX-R extracted bilingual dictionaries both at phrase and word level. The table lists the number of phrases, words and the fertility, where fertility is the ratio between the number of words in Spanish and the unique words in English. Notice that higher fertilities facilitate higher accuracies as there are more chances to retrieve a term.

Our experiments use the word level version of the dictionaries, but both versions are available at the CEREAL-ES site.[13]

Figure 10 depicts the accuracy for all the combinations of the embeddings (21+1) that can be used to infer the dictionary (21). The `all` embeddings are trained with data from the 21 varieties.

F Bias Effect Size in Embeddings

F.1 WEAT and Effect Size Equations

The Word Embedding Association Test (WEAT) (Caliskan et al., 2017) is a bias measurement method for word embeddings. It uses lists of terms (e.g., *flowers*) and attributes (e.g., *pleasant concepts*) that conceal implicit human associations (e.g., *flowers/insects vs. pleasant/unpleasant thoughts*). Mathematically, the association of each term t is defined as its average cosine similarity against the list of target attributes A:

$$assoc(t, A) = \frac{\sum_{a \in A} cos(\mathbf{t}, \mathbf{a})}{|A|}, \quad (1)$$

where $\mathbf{t}$ is the embedding for t and $\mathbf{a}$ is the embedding for an element $a \in A$. The association difference Δ_{assoc} for a term t between attributes A (*pleasant*) and B (*unpleasant*) is then

$$\Delta_{assoc}(t, A, B) = assoc(t, A) - assoc(t, B). \quad (2)$$

Given two sets of terms X and Y (e.g., *flowers* and *insects*), we use Cohen's d as standardised measure of the *effect size* (i.e. the strength of the bias). The effect size is defined as the difference between the two means divided by the standard deviation for all instances in X and Y:

$$d = \frac{\mu\left(\Delta_{assoc}(x, A, B)_{\forall x \in X}\right) - \mu\left(\Delta_{assoc}(y, A, B)_{\forall y \in Y}\right)}{\sigma\left(\Delta_{assoc}(w, A, B)_{\forall w \in X \cup Y}\right)}. \quad (3)$$

Sawilowsky (2009) defined the scale of magnitude for d as very small (< 0.01), small (< 0.20), medium (< 0.50), large (< 0.80), very large (< 1.20), and huge (< 2.00).

F.2 Fine-grained Analysis

Figure 11 shows the detailed effect sizes per individual lists and how they contribute to the mean value of a variety.

[12] We cut down the amount of peninsular Spanish data to that of the second largest variety —Mexican— to build a more balanced dataset.

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

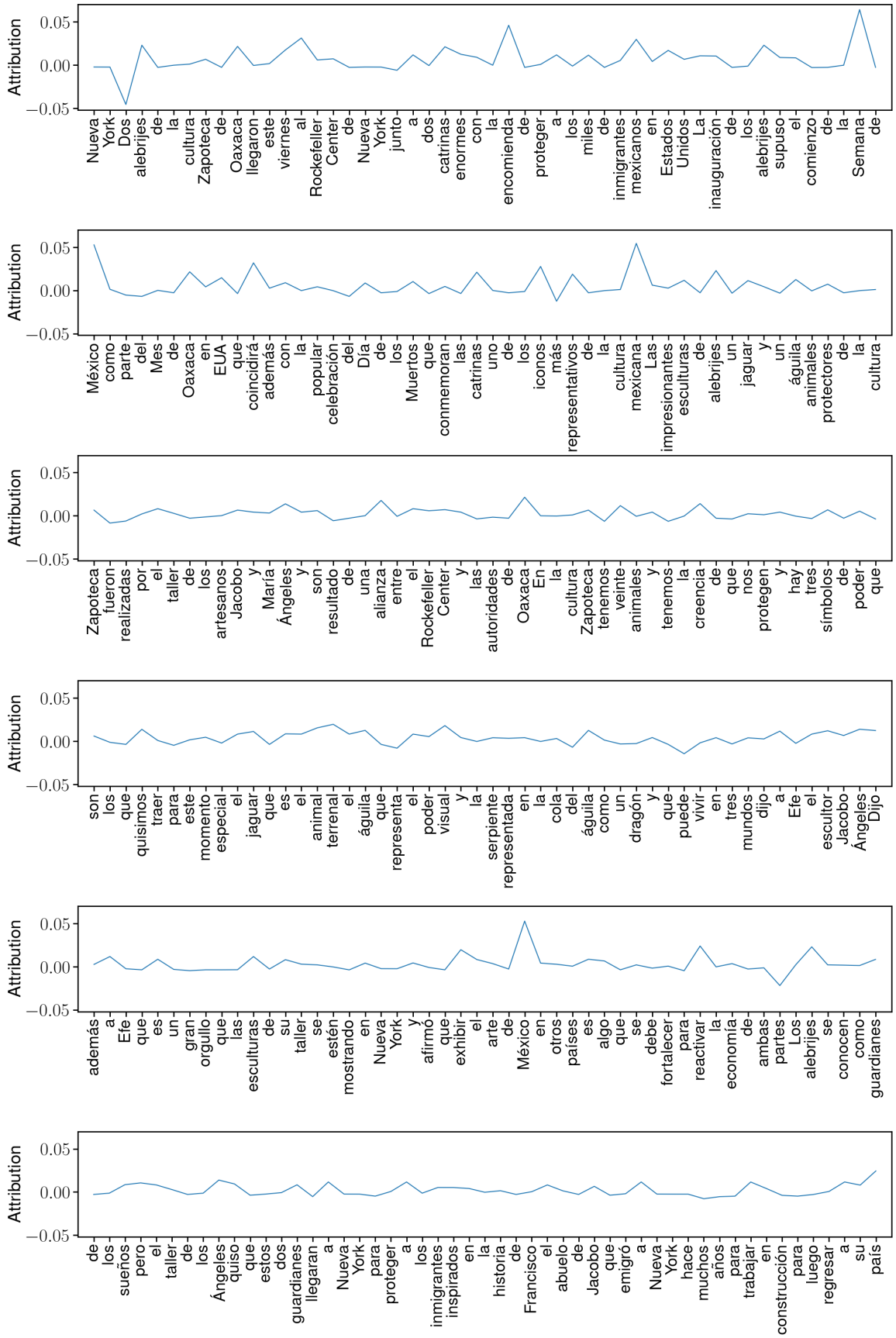


Figure 8: Attribution scores as obtained by the layer integrated gradients method after summing attributions from subunits. The snippet corresponds to the first 300 words of a Mexican article classified as such by the 4C classifier.

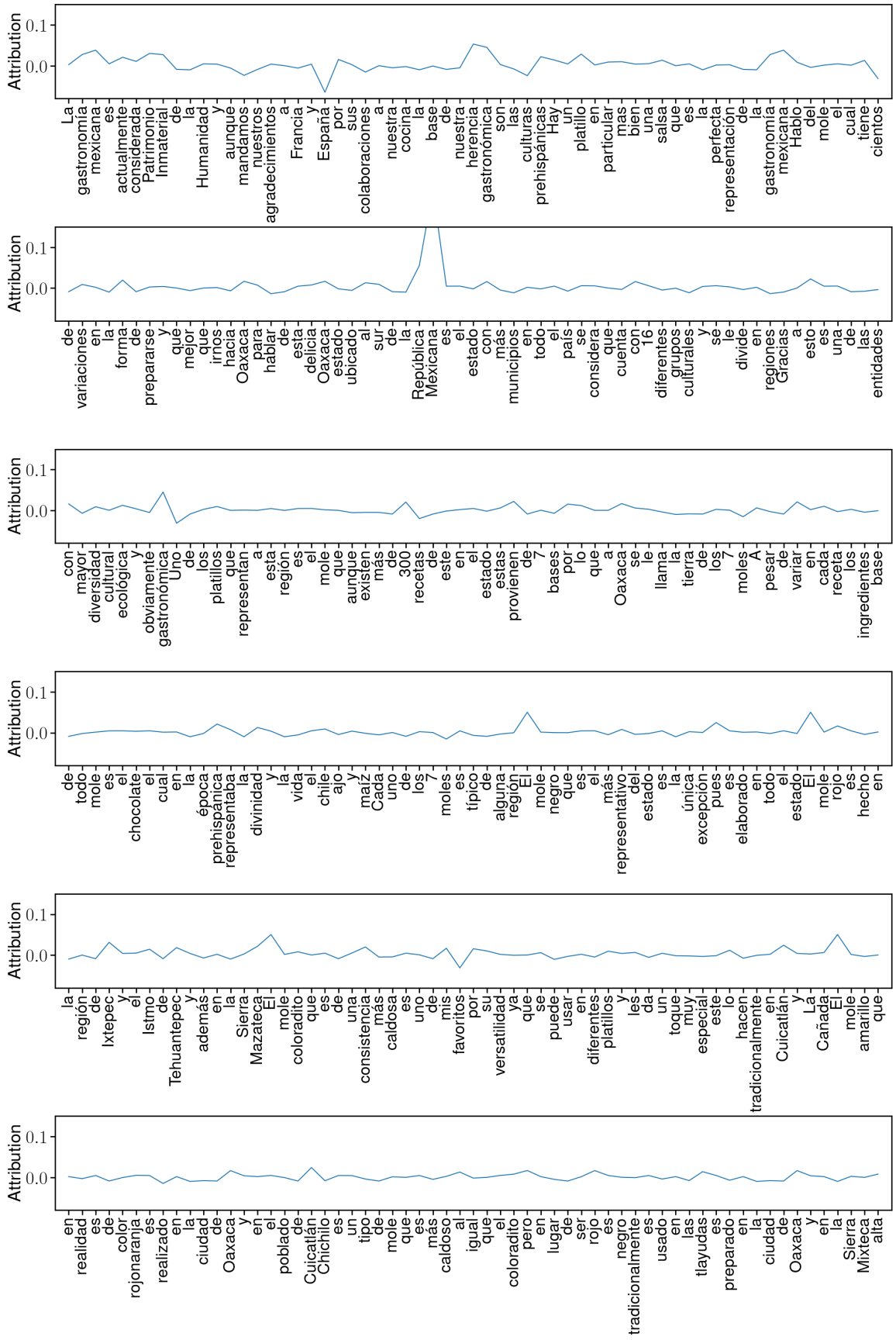


Figure 9: Attribution scores as obtained by the layer integrated gradients method after summing attributions from subunits. The snippet corresponds to the first 366 words of a Mexican article classified as such by the 4C classifier.

cl	es	mx	mix
Chile, 953	que, 373	Chiapas, 780	Nicaragua, 96
que, 374	España, 172	México, 254	en, 93
en, 271	tras, 169	de, 180	Panamá, 82
por, 119	y, 159	que, 110	el, 61
de, 94	Madrid, 129	en Chiapas, 102	Cuba, 60
Qué, 80	Cómo, 107	las, 59	que, 60
Boric, 78	con, 95	Cómpralo, 58	por, 48
Elecciones, 74	los, 94	Ciudad, 53	Guatemala, 43
pero, 71	las, 88	en, 47	Venezuela, 42
La Tercera, 62	obligatorios, 81	impresa, 46	Nacional, 36
en Chile, 58	la, 78	por, 46	su, 33
Tiempo, 57	por, 72	Chiapas a, 43	En, 33
y, 50	Recetas, 69	pues, 43	Municipales Economicas Deportes, 32
Santiago, 49	rebeló, 63	como, 40	Bolivia, 29
La, 47	os, 61	lo, 39	Costa Rica, 28
del, 47	euros, 60	del, 36	Cochabamba, 27
la, 46	lo, 56	y, 35	Ecuador, 27
chileno, 45	vacunación obligatoria, 53	Chiapas en, 34	como, 27
a, 44	Cómo hacer, 47	Chiapas El, 33	Actualidad, 25
Tercera, 44	superó, 45	pero, 31	Foto, 24
La Tercera Secciones, 42	TELVA, 44	para, 29	Colombia, 24
En, 40	y aprendió, 43	En, 29	Argentina, 22
La Serena, 40	Actualidad, 42	mexicana, 27	El Salvador, 21
un, 38	Qué, 41	es, 27	Ministerio, 21
como, 37	para, 41	Casas Chiapas, 24	pero, 21
Tercera Secciones, 37	coalición, 40	Chiapas México, 24	GN, 20
chilena, 36	coalición por, 38	una, 21	Fotogalerías, 20
país, 36	pero, 37	los, 21	Honduras, 20
UNAM, 36	aprovechó, 35	mexicano, 20	El, 19
Kast, 34	QAnon, 34	Tuxtla Gutiérrez Chiapas, 20	Boliviana, 18
chilenos, 33	Si, 33	El, 20	este, 18
al, 33	En, 33	al, 20	Noticias, 17
las, 33	como, 33	la, 20	Inicio Nacionales Municipales Economicas Deportes, 17
los, 32	en, 32	bitácora, 19	es, 17
Deco, 31	impaciente, 31	mil, 19	Municipales Economicas Deportes Internacionales, 16
lectura, 30	URJC, 31	un, 19	Municipales Economicas Deportes Internacionales Coronavirus, 15
Y, 29	recetas, 30	OaxacaChiapasNacional, 19	Ñucanchi, 15
de Chile, 29	Y, 29	fomentar en Chiapas, 19	sesión, 15
Cómo, 29	Comprar, 28	pesos, 18	Perú Inicio Economía, 15
El, 28	podéis, 27	mexicanos, 17	Nacionales Municipales Economicas Deportes, 15
Cuáles, 28	Cuáles, 25	noticias, 16	Honduras Farándula internacional, 15
chino, 27	español reduce, 25	ARTE, 15	en el, 15
comuna, 26	a, 25	La, 15	Investigación Opinión Entretenimiento, 14
feminismo, 26	impaciente invierno provoca, 24	Ciudad de México, 15	Buenos Aires, 14
el, 26	este, 24	su, 15	se, 14
Chile y, 25	te, 24	REPÚBLICA, 14	Perú, 14
insomnio, 24	Recetasderechupetecom, 23	Chiapas En, 13	EL COMERCIO, 14
Síguenos, 23	Telvacom, 23	Síguenos, 13	y, 14
minutos, 23	si, 23	Gutiérrez Chiapas, 13	Panamá y, 13
mapuche, 20	Fundación Jiménez, 22	de la, 13	DIARIO DE CUBA, 13
nacional, 20	PSOE, 21	Su, 12	país, 13
Ministerio, 20	superó un, 21	en Chiapas ANTONY, 12	Morales, 13
Ñublense, 20	Gobierno español reduce, 20	muy, 12	en Panamá, 12
chilenas, 19	Galicia, 20	ppm, 12	Policía, 12
a Chile, 19	español, 19	primavera 2020, 12	Panamá Ellas, 12

Table 9: Top 55 phrases according to the attribution score of Layer Integrated Gradients (phrase, frequency). Data correspond to the classification of the test set with 4C.

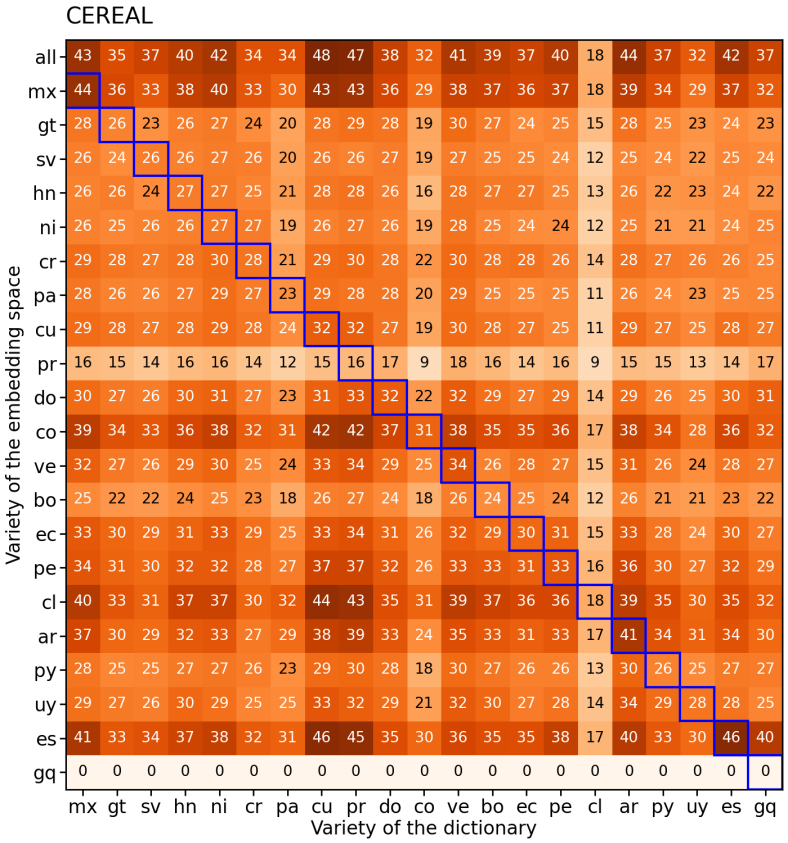
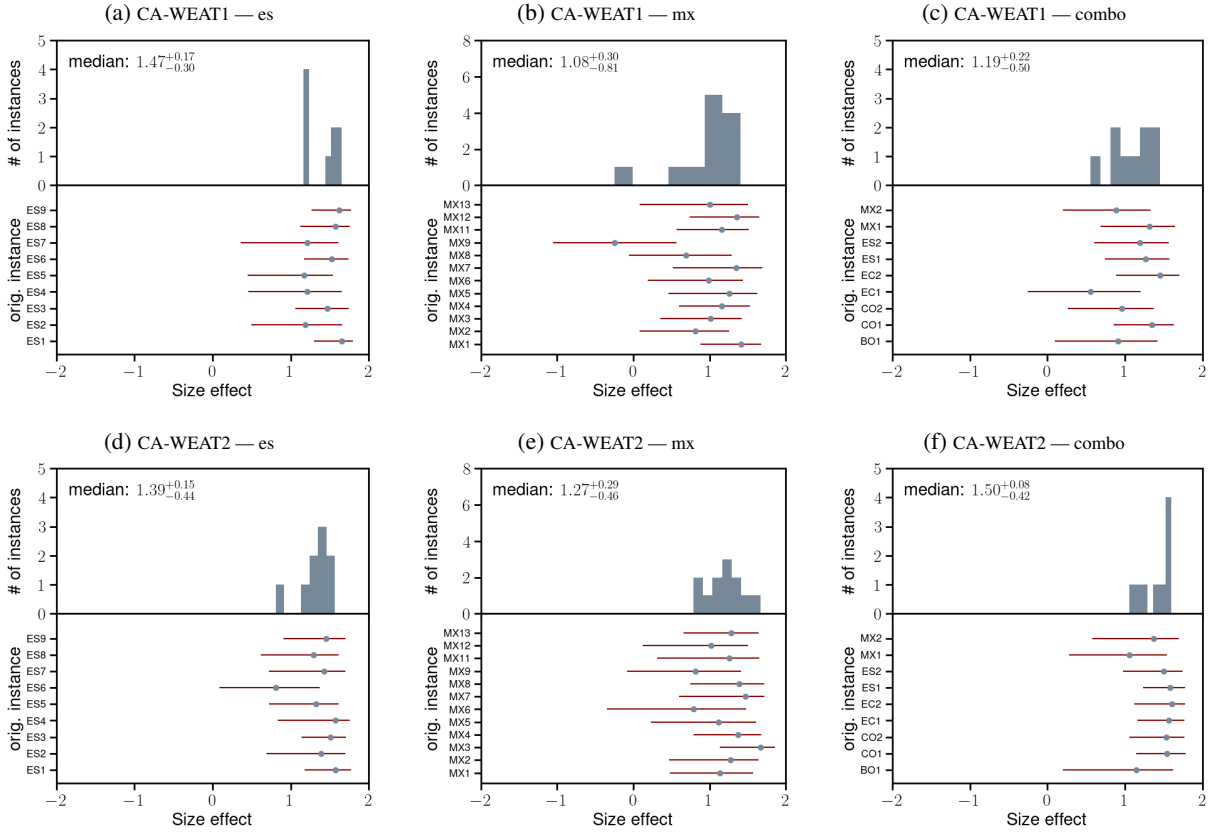


Figure 10: Accuracy (%) for the bilingual lexicon induction task for the 21 varieties with available test dictionaries.

CEREAL



Twitter

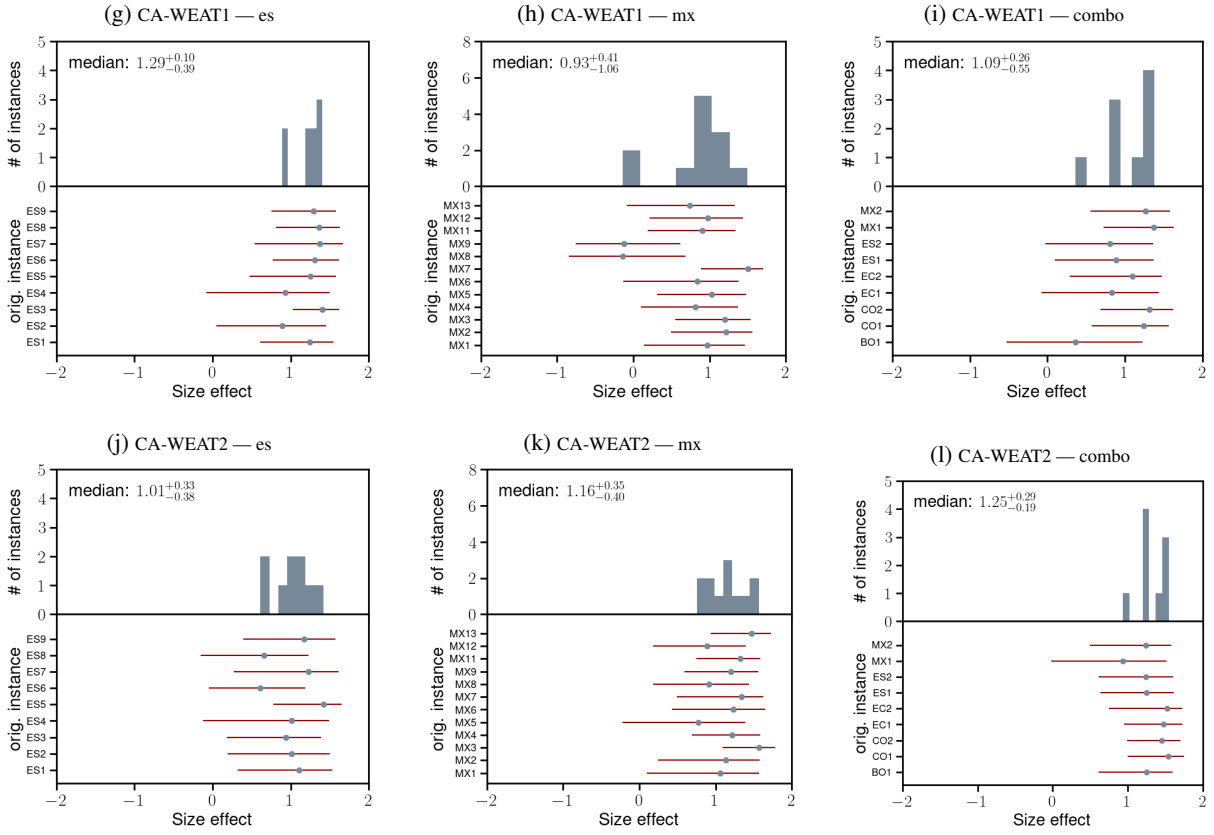


Figure 11: Effect sizes for the individual lists for each variety set (es, mx and combo) in the two embeddings domains (web and twitter), using CA-WEAT1 and CA-WEAT2. Median and confidence intervals are estimated with order statistics at 95% level for CA-WEAT. CIs for each list are estimated by bootstrap resampling also at 95% level.

ccTLD	Segments	Words	Vocab.	VocabTw
ad	13,023	543,047	2,671	–
ar	20,950,705	986,413,066	284,191	673,424
bo	975,429	49,518,821	53,799	47,012
cl	12,079,476	548,257,312	199,493	282,737
co	8,323,794	375,326,751	163,212	324,635
cr	825,513	37,760,657	45,893	103,086
cu	1,919,998	93,368,177	82,275	18,682
do	1,183,336	48,726,587	52,409	108,655
ec	1,624,269	66,662,454	64,312	147,560
es	20,950,705	880,495,659	596,842	571,196
gq	4,050	329,469	1,698	1,167
gt	561,714	23,421,191	35,860	95,252
hn	656,212	24,971,660	35,707	60,580
mx	20,875,244	912,645,564	250,313	438,136
ni	405,935	18,921,537	31,345	68,605
pa	448,974	18,431,387	31,268	111,635
pe	5,066,369	213,937,404	122,884	178,113
ph	1,382	75,761	405	–
pr	128,103	5,619,179	15,062	23,062
py	775,101	33,771,401	46,513	124,162
sv	401,348	17,068,212	29,433	73,833
us	376,839	21,335,770	34,368	292,465
uy	1,804,329	85,809,183	75,491	200,032
ve	1,201,624	55,514,289	59,334	271,924
all	101,553,472	4,518,924,538	736,895	1,696,232

Table 10: Number of segments and words used to compute the variety-specific word embeddings. The last columns show the vocabulary size of these embeddings and the equivalent for the Twitter embeddings in Tellez et al. (2023).

ccTLD	# Phrase entries	# Word entries	English uniques	Fertility (words)
ar	3237	1619	502	3.2
bo	1670	812	478	1.7
cl	2204	999	411	2.4
co	785	453	294	1.5
cr	908	565	354	1.6
cu	3828	1911	499	3.8
do	2042	1014	479	2.1
ec	2078	983	480	2.0
es	3085	1466	490	3.0
gq	1538	683	428	1.6
gt	1782	915	486	1.9
hn	2381	1165	491	2.4
mx	3886	1743	504	3.5
ni	3413	1616	491	3.3
pa	2028	907	427	2.1
pe	1699	924	477	1.9
pr	3231	1678	500	3.4
py	2790	1313	494	2.7
sv	1290	682	403	1.7
uy	1688	766	477	1.6
ve	2432	1195	480	2.5

Table 11: Statistics for the bilingual dictionaries English–Spanish extracted from VARILEX (Ueda and Moreno Fernández, 2016).