Variation and Instability in Dialect-Based Embedding Spaces

Jonathan Dunn
Department of Linguistics and
New Zealand Institute for Language, Brain and Behaviour
University of Canterbury
Christchurch, New Zealand
jonathan.dunn@canterbury.ac.nz

Abstract

This paper measures variation in embedding spaces which have been trained on different regional varieties of English while controlling for instability in the embeddings. While previous work has shown that it is possible to distinguish between similar varieties of a language, this paper experiments with two follow-up questions: First, does the variety represented in the training data systematically influence the resulting embedding space after training? This paper shows that differences in embeddings across varieties are significantly higher than baseline instability. Second, is such dialect-based variation spread equally throughout the lexicon? This paper shows that specific parts of the lexicon are particularly subject to variation. Taken together, these experiments confirm that embedding spaces are significantly influenced by the dialect represented in the training data. This finding implies that there is semantic variation across dialects, in addition to previously-studied lexical and syntactic variation.

1 Dialects and Embedding Spaces

This paper investigates the degree to which embedding spaces are subject to variation according to the regional dialect or variety that is represented by the training data. The experiments train character-based skip-gram embeddings on gigaword corpora representing four regional dialects of English (North America, Europe, Africa, and South Asia). While there is a robust tradition of discriminative modelling of dialects and varieties within NLP (Zampieri et al., 2017, 2018, 2019; Gaman et al., 2020; Chakravarthi et al., 2021; Aepli et al., 2022), there has been much less work on the influence which the dialectal composition of the training data (upstream) has on embedding spaces after training (downstream).

The basic idea in this paper is to train five iterations of character-based skip-gram embeddings on dialect-specific corpora in order to measure both variation (across dialects) and instability (within dialects); this is visualized in Figure 1. In order to find out whether specific parts of the lexicon are especially influenced by the dialect represented in the training data, the lexicon used for comparing embedding spaces is annotated for frequency, concreteness, part-of-speech, semantic domain, and age-of-acquisition.

If the specific dialect represented in the training corpus has no influence on embedding spaces, then variation across regions will be the same as variation within regions. In other words, we must control for instability (operationalized as variation across embeddings from the same dialect) to avoid false positives. However, if the dialect represented in the training data does have an influence on embedding spaces after training, then there will be a clear distinction between variation across dialects and instability within dialects.

The contribution of this paper is to show (i) that dialectal variation in character-based embedding spaces is significantly stronger than the noise caused by background instability and (ii) that this variation remains concentrated in certain parts of the lexicon. To accomplish this, we model the impact of dialect-specific training corpora on embeddings by controlling for background instability and organizing the experiments around the lexical attributes of frequency, concreteness, part-of-speech, semantic domain, and age-of-acquisition.

We begin by reviewing related work on dialectal variation and embedding stability (Section 2), before describing the main experimental questions (Section 3), the data (Section 4), and the methods (Section 5). We then compare variation within and between dialect-specific embeddings (Section 6) before modelling the influence of lexical factors on such dialectal variation (Section 7). Taken together, these experiments confirm that regional dialect or variety has a significant influence on embedding spaces that far exceeds baseline instability.

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 67–77

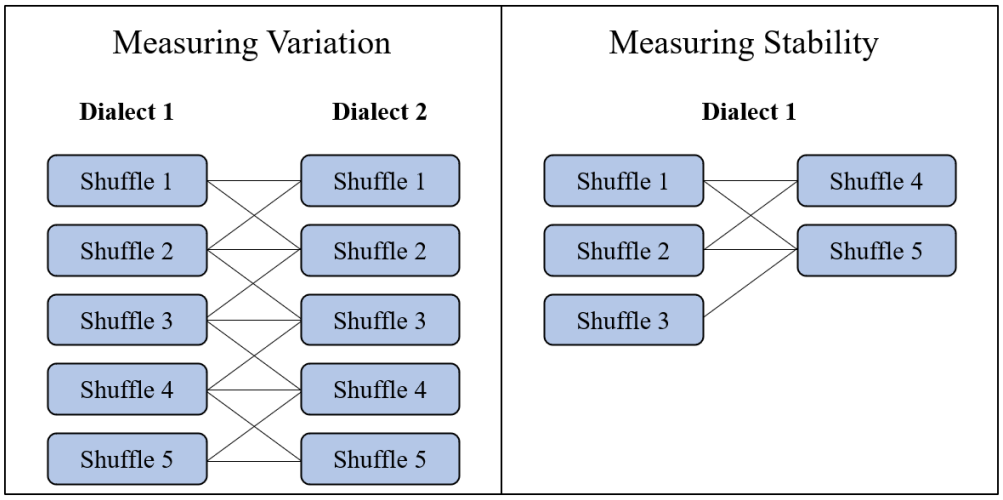

Figure 1: Overview of Comparison Methodology: Variation between dialects is estimated by sampling ten unique pairs of embeddings, where each embedding represents a shuffled version of a dialect-specific corpus. The baseline instability is estimated by sampling ten unique pairs of embeddings from different shuffled versions of a single dialect-specific corpus. Non-geographic factors like time period and random seed are held constant.

2 Related Work

This section discusses previous work in which models trained on data from different varieties (upstream) become significantly different after training (downstream). It also presents previous work on instability in embedding spaces.

Geography and Dialects. The creation of large geo-referenced corpora has made it possible to model variation across dialects, where unique locations represent unique dialect regions. Previous work has described geo-referenced corpora derived from web pages and social media (Davies and Fuchs, 2015; Dunn, 2020). Other work has evaluated the degree to which such corpora represent dialectal patterns found using more traditional methods (Cook and Brinton, 2017; Grieve et al., 2019), and the degree to which these corpora capture population movements triggered by events like the COVID-19 pandemic (Dunn et al., 2020). Further work has shown that geographic corpora from distinct sources largely agree on their representation of national dialects (Dunn, 2021). Building on these corpora, recent work has modelled both lexical variation (Wieling et al., 2011; Donoso and Sánchez, 2017; Rahimi et al., 2017) and syntactic variation (Dunn, 2018, 2019b; Dunn and Wong, 2022) in English as well as in other languages (Dunn, 2019a).

To what degree does dialectal variation influence semantic representations like skip-gram embeddings in addition to lexical and syntactic features? Previous work has shown that there is a significant difference between generic web-based embeddings and web-based embeddings trained using corpora sampled to represent actual population distributions; this difference was observed across 50 languages (Dunn and Adams, 2020). While these previous results lead us to expect dialectal variation across embeddings, there are two remaining questions: First, to what degree is this variation caused by dialectal differences as opposed to random instability? Second, is dialectal variation spread equally across the lexicon, equally influencing nouns and verbs, abstract and concrete, frequent and infrequent words?

Instability in Embeddings. A related line of work focuses on sources of instability in embedding spaces. It has been shown that many embeddings are subject to random fluctuation across different cycles of shuffling and retraining (Hellrich et al., 2019). Such instability has been investigated using word similarities (Antoniak and Mimno, 2018), showing that smaller corpora are subject to greater instability. In this line of work, two embeddings are compared by measuring the overlap in nearest neighbors for a target vocabulary. It has been shown, for example, that even high-frequency words can be unstable (Wendlandt et al., 2018) and that instability is related to properties of a language like the amount of inflectional morphology (Burdick et al., 2021). Other work has focused on the impact of time on embeddings, with variation leading to change (Cassani et al., 2021).

Circle	Region	Country	N. Words, Web	N. Words, Tweets
Inner-Circle	North American	Canada	250 mil	250 mil
		United States	250 mil	250 mil
Inner-Circle	European	Ireland	250 mil	250 mil
		United Kingdom	250 mil	250 mil
Outer-Circle	African	Nigeria	262 mil	100 mil
		Kenya	1 mil	100 mil
		Gabon	100 mil	100 mil
		Uganda	37 mil	100 mil
		Mali	100 mil	100 mil
Outer-Circle	South Asian	India	250 mil	250 mil
		Pakistan	250 mil	250 mil

Table 1: Source of Data by Region, Country, and Register

Other recent work has shown that register variation (Biber, 2012) has a significant impact on embedding similarity across a diverse range of languages (Dunn et al., 2022). This general approach to comparing embedding spaces focuses on aligned vocabulary (using nearest neighbors) rather than aligned embeddings because of instability in such alignment methods themselves (Gonen et al., 2020). As shown by this previous work, the comparison of nearest neighbors provides a robust method for detecting variation across embedding spaces.

This work on instability in embeddings is important because we need to distinguish between (i) variation across dialects and (ii) random fluctuations in embedding representations themselves. In other words, given the finding that embeddings trained on corpora representing different dialects are significantly different, how much of this is noise caused by random instability?

3 Experimental Questions

This paper focuses on two questions: First, are there significant differences in embeddings trained from corpora representing different dialects when accounting for baseline instability in the embeddings? Second, if so, are these dialectal differences specific to a certain part of the vocabulary, such as words belonging to a specific semantic domain?

The basic idea here is to compile four gigaword corpora representing English as used in North America, Europe, Africa, and South Asia. These areas represent different dialect regions. For example, while there are smaller differences between American English and Canadian English, these two dialects are more similar to one another than to other national dialects like Irish English. For example, work on syntactic variation has shown that American and Canadian English, at least in digital contexts, are closely related while UK and Irish English form a separate closely related pair (Dunn, 2019a). Based on the distribution of errors within a confusion matrix, other work has shown that Indian and Pakistani English are likewise more similar to one another than to other dialects (Dunn, 2018).

The dialects included represent both inner-circle and outer-circle varieties. The concept of inner-circle vs outer-circle is based on the historical stages of European colonization (Kachru, 1982). This distinction within the World Englishes paradigm is meant to capture the perceived prestige differences of these dialects rather than to make a distinction between dialects and varieties as linguistic objects. For example, inner-circle populations tend to have a higher socio-economic status and better access to digital technologies, leading to their status as prestige varieties. Both groups can be considered dialects. In some cases speakers of outer-circle varieties could be considered second-language learners; however, regardless of a distinction between native and non-native speakers, the production found in outer-circle varieties remains robust and predictable over time. Thus, we treat both inner-circle and outer-circle varieties as dialects of equal standing but maintain the terminology from the World Englishes paradigm in order to provide a bridge to work in sociolinguistics.

We first train embeddings on each dialect-specific corpus and then measure variation across a lexicon that is annotated for concreteness, age-of-acquisition, semantic domain, part-of-speech, and frequency. We train five sets of embeddings for each dialect-specific corpus, each based on a

random reshuffling of the corpus. This allows us to measure the difference between variation (across dialects) and baseline instability (within dialects).

We work with skip-gram embeddings (SGNS: Mikolov et al. 2013) as implemented in the fastText framework (Bojanowski et al., 2017). In particular, we use the skip-gram variant with negative sampling ($n = 50$) trained for 20 epochs with a learning rate of 0.05 and 100 dimensions. The character n-gram sizes range from 3 to 6, with a maximum of 2 million n-gram buckets allowed. Because previous work has shown that different random seeds can cause instability (Gonen et al., 2020), we control for such instability by using the same random seed for each set of embeddings. Thus, variation caused by random seed and by training parameters is taken into account in this experimental set-up.

Several considerations support the use of non-contextual skip-gram embeddings for these experiments. In the first case, the focus here is on semantic variation rather than lexical or syntactic structures and the long-distance co-occurrences captured by the skip-gram task are taken as better representations for such semantic variation. In the second case, the inclusion of low-resource dialects like African English means that the amount of training data available is limited and insufficient for training robust contextual embeddings. Given the dual goals of focusing on semantics while also including low-resource dialects, skip-grams provide the most practical type of embedding for answering these particular experimental questions.

4 Data

The data used here represents different geographic locations which, in turn, represent different dialects. The data itself is drawn from two registers, web pages and tweets, both derived from the *Corpus of Global Language Use* (Dunn, 2020). The experiments train character-based embeddings for these four different regional dialects, as shown in Table 1. Each corpus contains 1 billion words, equally divided between registers (web pages and tweets). Thus, for example, the inner-circle North American corpus contains 500 million words of tweets, equally divided between Canada and the United States. The African web corpus has additional constraints because there is less data per country. As shown in Table 1 this corpus combines five countries into a single regional data set. The even split between web pages and tweets is maintained.

5 Methods

For each regional variety of English, we train embeddings using the fastText framework with the parameter settings described above. Previous work has shown that this family of embeddings can be unstable; in this context, *instability* means that the same training corpus could result in multiple sets of nearest neighbors over different iterations (Hellrich et al., 2019). We control for this by randomly shuffling each corpus and retraining the embeddings five times. Because all comparisons are between two sets of embeddings, we thus obtain ten observations (unique comparisons) to represent each condition, as visualized in Figure 1. We use the same random seed and the same parameters across all sets of embeddings to control for other sources of variation.

Vocabulary Features. The vocabulary for the embedding space is derived from semantic and psycholinguistic resources that provide categorizations for specific lexical items. This source of vocabulary allows us to compare stability and variation across different sub-sets of the lexicon.

Concreteness	**N.**	**POS**	**N.**
1.0 to 2.0	2,426	Adjective	4,130
2.0 to 3.0	5,619	Adverb	189
3.0 to 4.0	4,167	Name	139
4.0 to 5.0	4,599	Noun	9,827
-	-	Verb	2,322
-	-	Other	205
Total	**16,812**	**Total**	**16,812**

Table 2: Distribution of Vocabulary Items Across Concreteness Categories and Parts-of-Speech

The first source of lexical annotations is a participant-based study of concreteness (Brysbaert et al., 2014). This source provides concreteness ratings between 1 and 5 for each lexical item, with higher values reflecting more concrete and lower values reflecting more abstract judgements from participants. This source also provides the most common part-of-speech for each lexical item. The distribution of the vocabulary across concreteness ratings and parts-of-speech is shown in Table 2. An example of an abstract word (1.0 to 2.0) is *belief*; less abstract (2.0 to 3.0) is *famished*; more concrete (3.0 to 4.0) is *galaxy*; and most concrete (4.0 to 5.0) is *fire*. Within parts-of-speech, most words are categorized as adjectives, adverbs, nouns, or verbs.

Because different vocabulary items are generally

Category	N.	Conc	AoA
General & Abstract	2,384	2.4	10.5
Body & Individual	1,268	3.8	9.8
Arts & Crafts	114	3.8	9.8
Emotion	765	2.3	9.9
Food & Farming	586	4.2	8.6
Government & Public	761	2.9	10.9
Housing & Home	336	4.2	8.7
Money & Commerce	531	3.2	10.5
Entertainment	459	3.9	8.7
Life & Living Things	594	4.3	8.3
Movement & Travel	897	3.5	9.1
Numbers & Measures	795	2.8	9.7
Materials & Objects	1,806	3.7	9.0
Education	118	3.3	9.8
Communication	943	3.2	9.9
Social Actions	1,959	2.7	10.2
Time	474	2.7	9.2
World & Environ.	298	3.9	8.6
Psychological	1,255	2.4	9.8
Science & Tech	161	3.3	11.4
Names & Grammar	307	2.9	7.5
Total	**16,812**	**3.1**	**9.7**

Table 3: Distribution of Vocabulary Items Across Semantic Domains with Concreteness and Age-of-Acquisition Information for Each Domain

learned at different stages of language acquisition, we also include age-of-acquisition ratings for the vocabulary (Kuperman et al., 2012). These ratings are collected via MechanicalTurk but validated against ground-truth age-of-acquisition ratings collected in a laboratory setting. For instance, words like *mom*, *water*, and *yes* are reported to be learned during a child's second year. But words like *constrain*, *confound*, and *thyme* are reported to only be learned at the age of twelve. If more socially-conditioned words are subject to more variation, we might expect, then, that vocabulary learned later in life is subject to more variation as a result. Note that both sets of participant-based ratings (age-of-acquisition and concreteness) depend on inner-circle participants. Thus, these experiments are focused on variation in embedding spaces rather than variation in participant-based lexical features.

The next source of lexical annotations is the UCREL Semantic Analysis system (Piao et al., 2015) which provides a high-level semantic domain for each vocabulary item. For example, there are 586 items belonging to the domain FOOD AND FARMING and 761 to the domain GOVERNMENT AND PUBLIC. The inventory of semantic domains is shown in Table 3 along with the average concreteness and average age-of-acquisition for each. There is a clear relationship between semantic domain and concreteness: for example, the domain that includes PSYCHOLOGICAL STATES is highly abstract at 2.4 while the domain that includes FOOD AND FARMING is highly concrete at 4.2. In the same way, some semantic domains are acquired early (like NAMES AND GRAMMAR at 7.5 years of age) and others much later (like SCIENCE AND TECHNOLOGY at 11.4 years of age).

Word	Stability	Overlap		
	NA	EU	AF	SA
shag	0.53	0.00	0.00	0.03
daft	0.59	0.00	0.00	0.13
posh	0.66	0.00	0.00	0.05
proprietor	0.52	0.10	0.06	0.12
queue	0.63	0.10	0.08	0.08
abolish	0.80	0.22	0.23	0.28
bicker	0.61	0.22	0.03	0.20
isolationist	0.79	0.32	0.17	0.30
justice	0.82	0.32	0.24	0.22
reminisce	0.79	0.42	0.02	0.39
weeping	0.78	0.42	0.38	0.38
dictatorship	0.88	0.68	0.48	0.53
totalitarian	0.88	0.69	0.42	0.51
ten	0.93	0.77	0.57	0.69
twelve	0.94	0.77	0.62	0.70

Table 4: Examples With Different Levels of Overlap, North America Compared to All Other Varieties

In addition to these participant-based and semantic-based annotations, each lexical item also belongs to a frequency strata. This is calculated using the entire corpus across all regions and reported in occurrences per 1 million words.

Calculating Overlap. The stability and similarity of word representations are calculated using the overlap of nearest neighbors (Burdick et al., 2021). Given two sets of embeddings (i.e., North America and Europe) we iterate over each word in the lexicon. First, we retrieve the k nearest neighbors using cosine similarity. Second, we calculate the overlap between the two sets of nearest neighbors. For example, if all ten out of ten words appear in both embeddings as nearest neighbors, the overlap is 100%. If only five words out of ten appear as neighbors, the overlap is 50% (five shared words

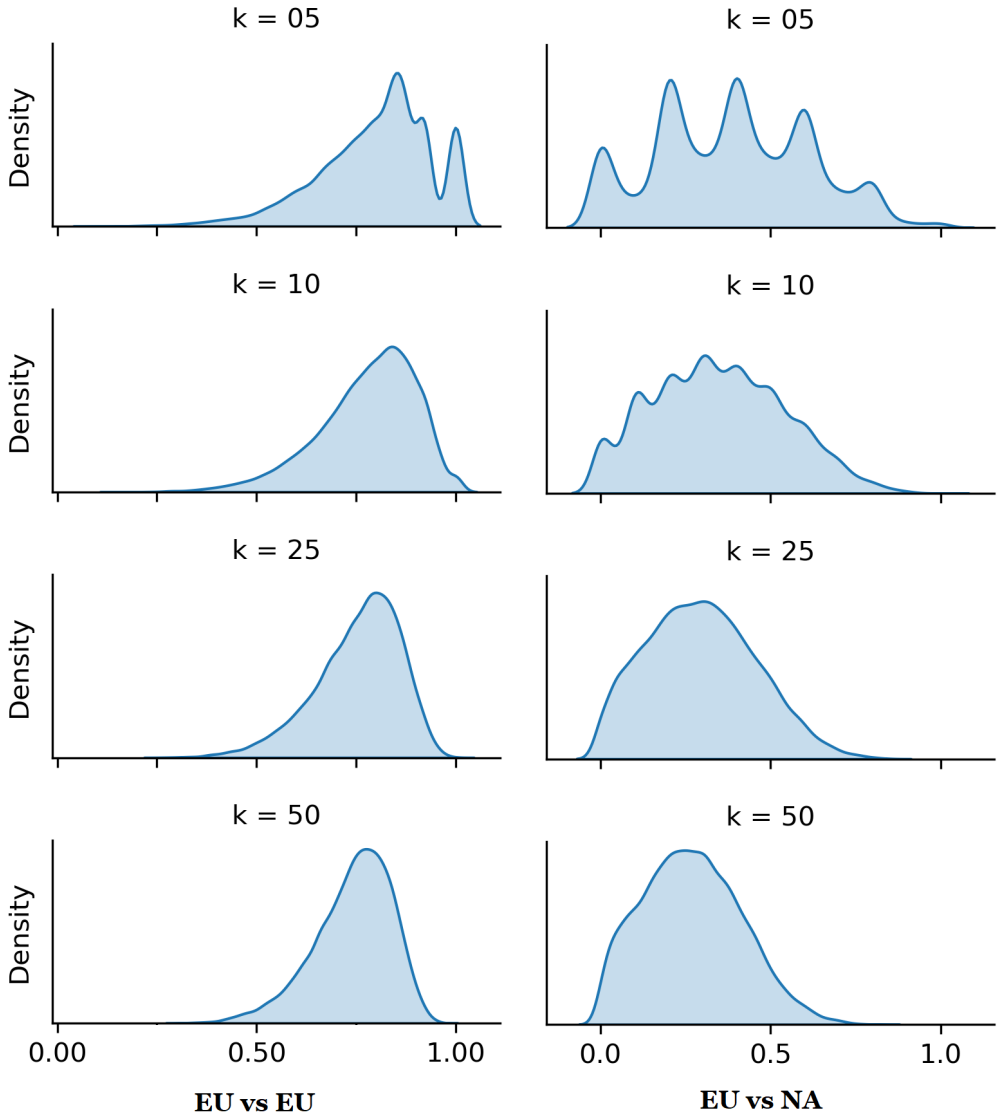

Figure 2: Distribution of Overlap Values Across Settings of k, for Europe-Europe and Europe-America Comparisons

out of 10 possible shared words). This provides a word-specific measure of overlap. This method aligns the vocabulary space rather than aligning the embedding spaces; this approach is taken because alignment methods have previously been shown to be unstable (Gonen et al., 2020) and thus less suitable for identifying variation across dialects.

A selection of example levels of overlap is shown in Table 4, with the North American embeddings compared with all other dialects. The smallest amount of overlap is shown for words like *daft* and *posh* which are used in different senses across these dialects. Culture-specific words like *isolationist* and *justice* provide a mid-level of overlap, with a similar sense but different references across dialects. Finally, a further cultural influence is shown for political words like *dictatorship*, which are more similar in inner-circle dialects than in outer-circle dialects. These examples show the range of overlap levels that are observed.

We measure overlap with values for k of 5, 10, 25, and 50. The distribution of overlap values is shown in Figure 2 for the European and North American model (on the right) and for the European and European model (on the left). Thus, the distributions on the right are across dialects and those on the left are within the same dialect. The impact of k is shown in the plots, with $k = 5$ at top and $k = 50$ at bottom. Smaller values of k lead to ragged distributions simply because the number of possible overlap values is limited. How much impact does the choice of k have on the results? We can see that higher values lead to finer estimates of the distribution of overlap, but overall the val-

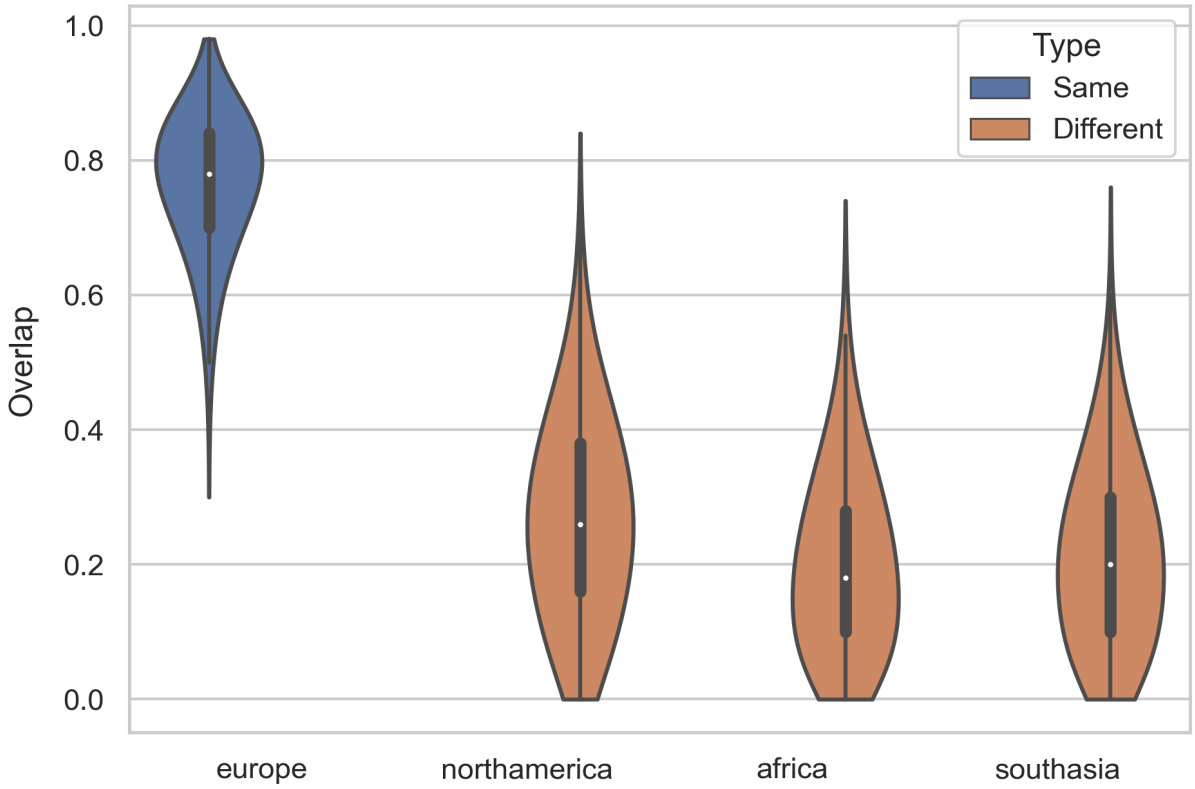

Figure 3: Distribution of Within-Dialect vs Between-Dialect Overlap Values for Europe

ues are much the same once k is above 10. For example, there is a significant Pearson correlation between overlap values at $k = 25$ and $k = 50$ (0.937 on the right and 0.879 on the left, in both cases with $p < 0.001$). We use $k = 50$ for the rest of the analysis, but the choice of k (above 10) has minimal impact on the results. We also see that the overlap within the same dialect (on the left) is greater than the overlap between different dialects (on the right). The figures for all distributions are available in the supplementary material.[1]

6 Overlap Within vs Between Dialects

The first experiment evaluates whether the variation between dialects remains meaningful when compared with baseline instability within a single dialect. The overlap measure described above compares the similarity between two sets of embeddings. We visualize the within vs between dialect condition in Figure 3 for Europe, with each type of comparison a separate violin plot. In blue we see the within-dialect overlap in which we compare European embeddings to other European embeddings. In orange we see between-dialect overlap for each of the other three regions. There is a clear distinction here between variation within the same dialect (baseline instability) and between different dialects (actual variation).

While we measured overlap between ten unique pairs of embeddings for each condition, this figure shows only the first pair for each. The supplementary materials contain the figures for all comparisons. The conclusion remains the same: the variation in embeddings across dialects is not simply a result of instability alone. There is a clear visual distinction between within-dialect and between-dialect overlap in all cases. We test for significance using a paired t-test: for example, are the values for Europe-Europe comparisons actually different from the values for Europe-Africa comparisons? For each comparison, we randomly choose a single pair of embeddings to test (i.e., so that we compare Europe and Africa only once). In each case the difference is significant with $p < 0.001$.

Thus, there is a visually clear and statistically significant difference between baseline instability and variation across dialects. We quantify the magnitude of this difference in Table 5 using a Bayesian estimate of the mean difference across the entire vocabulary and all pairs of embeddings. Within-

[1] https://jdunn.name/2023/03/27/variation-and-instability-in-dialect-based-embedding-spaces/

	EU	NA	AF	SA
EU	**74.1%**	26.8%	19.7%	21.5%
NA	–	**74.0%**	20.3%	22.0%
AF	–	–	**71.5%**	20.7%
SA	–	–	–	**72.8%**

Table 5: Bayesian Estimates of Overall Overlap Within and Between Dialects at 95% Confidence Interval, Across All Comparisons, $k = 50$

dialect overlap ranges from 71.5% to 74.1%, providing a baseline for instability. Between-dialect overlap ranges from 19.7% to 26.8%, showing that the dialect represented by the training corpus has a large influence on downstream embeddings.

There is a slight effect for inner-circle and outer-circle dialects: North America (NA) and Europe (EU) are more similar to one another than to Africa (AF) or South Asia (SA). Compared to the distinction between variation and baseline instability, however, this effect is relatively minor. The outer-circle varieties also have slightly lower stability than the inner-circle varieties.

7 Lexical Factors

We have shown that there is a significant difference in embedding spaces depending on the dialect represented in the training data, a difference that is much greater than baseline instability within dialects (as simulated by shuffling and retraining on the same corpora). This section explores dialectal variation in embedding spaces further by focusing on the impact of the lexical factors described in Section 5. We ask whether this kind of variation is distributed equally across the lexicon or whether it is concentrated in particular types of vocabulary.

We model the relationship between lexical attributes and overlap using a linear mixed effects regression model, with one model for each dialect. Within each model, the region of comparison is a fixed effect: for example, we model variation within the European embeddings using their overlap with North America, Africa, and South Asia as fixed effects. For random effects we include all lexical attributes. We represent each region using the average overlap across all ten pairs of embeddings, using $k = 50$ as before. The means of different regions are independent in the sense that each vocabulary item is modelled independently from corpora representing that region.

The coefficients and p-values for each lexical attribute are shown in Table 6 for all attributes that are significant for at least one dialect ($p < 0.01$). Positive categorical factors are shown above and negative factors below. Columns show results from the four dialect-specific models. While some factors are significant in one dialect but not another, no attributes have opposite effects across dialects (i.e., indicate more variation in one dialect but less variation in another).

Within semantic domains, vocabulary involving BODY AND INDIVIDUAL (e.g., *pain* and *ache*) are more stable across dialects, as are FOOD AND FARMING (e.g., *celery* and *sushi*) and SCIENCE AND TECHNOLOGY (e.g., *biologist* and *geologist*). These terms are less socially-conditioned in the sense that they refer to tangible objects or to specially-defined fields (like biology) that transcend cultural boundaries. On the other hand, vocabulary from semantic domains HOME AND HOUSING (e.g., *guest* or *pew*), MOVEMENT AND TRAVEL (e.g., *turnpike* or *curbside*), and NAMES AND GRAMMAR (e.g., *northwestern* or *roman*) are subject to more variation. These words are more socially-conditioned in the sense that they presume socially-defined concepts: a *guest* requires a definition of family units and a *pew* is a part of the concept CHURCH. Within parts-of-speech, function words (e.g., *of* or *and*) and adverbs (e.g., *hardly* and *exactly*) are much more stable. And named entities (e.g., *Flint*) are much less stable. Verbs are more important to the model than nouns.

Of the three scalar attributes, frequency has a significant effect but the coefficient is so small it is negligible. Concreteness is significant in every region, with more abstract words (e.g., *surreal* and *sanctimonious*) being more stable while more concrete words (e.g., *cookie* and *bug*) are less stable. In this case, the specific instances (the referents) of these more concrete terms are likely to be quite different across dialects (*cookies* are different in different places). Age-of-acquisition is significant in three out of four regions, but it has only a relatively small effect, with words acquired at a younger age being more stable. For instance, *mother* and *grandmother* (learned at age 2) are quite stable while *ethos* and *polarization* (learned at age 15) are subject to variation. The full regression results and the stability/variability values for the entire lexicon are available in the supplementary materials.[2]

[2] https://jdunn.name/2023/03/27/variation-and-instability-in-dialect-based-embedding-spaces/

Positive Factors		**Europe**		**N. America**		**Africa**		**South Asia**	
Subject to Less Variation		*coef.*	*p*	*coef.*	*p*	*coef.*	*p*	*coef.*	*p*
Domain	Body, Individual	3.97	0.000	4.36	0.000	3.82	0.000	4.56	0.000
Domain	Science, Tech	3.10	0.000	4.19	0.000	2.19	0.000	3.25	0.000
Domain	Food, Farming	2.67	0.000	2.98	0.000	1.87	0.000	3.25	0.000
Domain	Emotion	1.44	0.000	1.67	0.000	0.58	0.002	1.08	0.000
Domain	Arts, Crafts	1.37	0.004	–	–	–	–	1.21	0.008
Domain	Govt., Public	1.28	0.000	1.87	0.000	1.57	0.000	1.60	0.000
Domain	Entertainment	0.84	0.001	0.75	0.004	–	–	–	–
Domain	World, Environ.	–	–	0.91	0.003	–	–	1.22	0.000
Domain	Psychological	–	–	0.65	0.000	–	–	0.46	0.005
Domain	Social Actions	–	–	0.49	0.001	–	–	–	–
POS	Verb	2.58	0.000	2.48	0.000	2.71	0.000	2.26	0.000
POS	Function	12.97	0.000	10.35	0.000	12.14	0.000	10.43	0.000
POS	Adverb	8.83	0.000	7.23	0.000	8.47	0.000	6.55	0.000
Negative Factors		**Europe**		**N. America**		**Africa**		**South Asia**	
Subject to More Variation		*coef.*	*p*	*coef.*	*p*	*coef.*	*p*	*coef.*	*p*
Domain	Communication	-0.59	0.002	–	–	-0.77	0.000	-0.59	0.001
Domain	Money, Com.	-1.07	0.000	-0.88	0.000	–	–	-0.67	0.004
Domain	Life, Living	-1.12	0.000	–	–	-1.72	0.000	–	–
Domain	Materials, Objects	-1.14	0.000	-0.91	0.000	-1.42	0.000	-0.63	0.000
Domain	Movement, Travel	-1.94	0.000	-1.50	0.000	-2.14	0.000	-1.28	0.000
Domain	Housing, Home	-2.19	0.000	-2.38	0.000	-2.39	0.000	-1.54	0.000
Domain	Name, Grammar	-2.24	0.000	–	–	-2.10	0.000	–	–
POS	Names	-5.81	0.000	-6.40	0.000	-4.67	0.000	-6.19	0.000
POS	Noun	–	–	-0.34	0.001	–	–	-0.40	0.000
Scalar Factors		**Europe**		**N. America**		**Africa**		**South Asia**	
Lower Ratings=Less Variation		*coef.*	*p*	*coef.*	*p*	*coef.*	*p*	*coef.*	*p*
Empirical	AoA	-0.54	0.000	-0.53	0.000	-0.56	0.000	-0.48	0.000
Empirical	Concreteness	-1.66	0.000	-1.72	0.000	-1.69	0.000	-1.74	0.000

Table 6: Coefficients and P-Values from a Linear Mixed Effects Regression Model Using the Mean Overlap Across Dialects as the Dependent Variable. Non-Significant Effects are Not Shown.

8 Discussion and Conclusions

These experiments have shown that embedding spaces are subject to variation according to the dialect represented by the training data. This variation is significantly greater than noise caused by baseline instability in the embeddings themselves. This finding confirms the importance of regional dialects in NLP: while previous work has shown the impact of dialect on lexical and syntactic representations, this paper shows that such variation also extends to semantic representations.

Previous work has focused on distinguishing between dialects or on directly modelling variation over space and time. This paper has taken a different approach by training otherwise comparable models on corpora representing different dialects, controlling for other sources of variation like parameter settings and random seeds. The results show that the dialects represented in the training context have significant downstream impacts on common semantic representations (embeddings). These findings raise important questions for future work. First, is the influence of dialect consistent across languages or is this a result of the colonial history of a few languages like English? Second, do contextual embeddings also manifest this type of variation or is it confined to non-contextual skip-gram embeddings? Third, would a larger inventory of dialect-specific embeddings change the distribution of variation within the lexicon or is this a stable effect? Regardless of such further questions, these experiments show that dialect has a downstream effect on semantic representations, expanding previous work on lexical and syntactic representations.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2018. Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1):9–37.

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.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46:904–911.

Laura Burdick, Jonathan K. Kummerfeld, and Rada Mihalcea. 2021. Analyzing the Surprising Variability in Word Embedding Stability Across Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5891–5901. Association for Computational Linguistics.

Giovanni Cassani, Federico Bianchi, and Marco Marelli. 2021. Words with consistent diachronic usage patterns are learned earlier: A computational analysis using temporally aligned word embeddings. *Cognitive Science*, 45(4):e12963.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11. Association for Computational Linguistics.

Paul Cook and Laurel J. Brinton. 2017. Building and Evaluating Web Corpora Representing National Varieties of English. *Language Resources and Evaluation*, 51(3):643–662.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1):1–28.

Gonzalo Donoso and David Sánchez. 2017. Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, volume 4, pages 16–25. Association for Computational Linguistics.

Jonathan Dunn. 2018. Finding variants for construction-based dialectometry: A corpus-based approach to regional CxGs. *Cognitive Linguistics*, 29(2):275–311.

Jonathan Dunn. 2019a. Global syntactic variation in seven languages: Towards a computational dialectology. *Frontiers in Artificial Intelligence: Language and Computation*.

Jonathan Dunn. 2019b. Modeling global syntactic variation in english using dialect classification. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 42–53. Association for Computational Linguistics.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation.*, 54:999–1018.

Jonathan Dunn. 2021. Representations of language varieties are reliable given corpus similarity measures. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties, and Dialects*, pages 28–38. Association for Computational Linguistics.

Jonathan Dunn and Ben Adams. 2020. Geographically-balanced gigaword corpora for 50 language varieties. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2528–2536. European Language Resources Association.

Jonathan Dunn, Tom Coupe, and Ben Adams. 2020. Measuring linguistic diversity during covid-19. In *Proceedings of the Workshop on NLP and Computational Social Science*, pages 1–10. Association for Computational Linguistics.

Jonathan Dunn, Haipeng Li, and Damien Sastre. 2022. Predicting embedding reliability in low-resource settings using corpus similarity measures. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, pages 6461–6470. European Language Resources Association.

Jonathan Dunn and Sidney Wong. 2022. Stability of syntactic dialect classification over space and time. In *Proceedings of the International Conference on Computational Linguistics*.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14. International Committee on Computational Linguistics (ICCL).

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555. Association for Computational Linguistics.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence*, 2:11.

Johannes Hellrich, Bernd Kampe, and Udo Hahn. 2019. The Influence of Down-Sampling Strategies on SVD Word Embedding Stability. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 18–26. Association for Computational Linguistics.

Braj Kachru. 1982. *The Other tongue: English across cultures.* University of Illinois Press, Urbana-Champaign.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44:978–990.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *arXiv preprint*, volume arXiv:1301.3781.

Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'egidio, and Paul Rayson. 2015. Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1274. Association for Computational Linguistics.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. In *Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 209–216.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102. Association for Computational Linguistics.

Martijn Wieling, John Nerbonne, and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS One*, 6:9.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16. Association for Computational Linguistics.