Predicting Embedding Reliability in Low-Resource Settings Using Corpus Similarity Measures

Jonathan Dunn, Haipeng Li, Damian Sastre
University of Canterbury, Department of Linguistics
and the New Zealand Institute for Language, Brain and Behaviour
Christchurch, New Zealand
jonathan.dunn@canterbury.ac.nz, haipeng.li@canterbury.ac.nz, dsa105@uclive.ac.nz

Abstract

This paper simulates a low-resource setting across 17 languages in order to evaluate embedding similarity, stability, and reliability under different conditions. The goal is to use corpus similarity measures before training to predict properties of embeddings after training. The main contribution of the paper is to show that it is possible to predict downstream embedding similarity using upstream corpus similarity measures. This finding is then applied to low-resource settings by modelling the reliability of embeddings created from very limited training data. Results show that it is possible to estimate the reliability of low-resource embeddings using corpus similarity measures that remain robust on small amounts of data. These findings have significant implications for the evaluation of truly low-resource languages in which such systematic downstream validation methods are not possible because of data limitations.

Keywords: low-resource languages, embedding similarity, corpus similarity, register variation

1. Validating Low-Resource Embeddings

This paper simulates a low-resource setting for 17 non-English languages in order to evaluate embedding similarity, stability, and reliability under different conditions. While these are actually high-resource languages, we are able to simulate a low-resource setting across different language families, writing systems, and types of morphology by constraining both the amount and the type of data that is made available. We then try to predict the downstream similarity, stability, and reliability of embeddings given upstream properties of the training corpora. The larger goal is to predict the reliability of embeddings in truly low-resource settings in which such experiments are not possible. The key finding of the paper is that there is a strong relationship between the similarity of training corpora and the similarity of embeddings, a relationship that extends across a diverse range of languages and registers.

Embeddings remain a key representation within NLP, as shown by many samples of recent work (Miaschi and Dell'Orletta, 2020; Adelmann et al., 2021). At the same time, however, recent work has also shown that embeddings are surprisingly variable (Wendlandt et al., 2018; Burdick et al., 2021). Such work has shown that high-resource languages like English, with many billions of words available for training, have embeddings that differ by data set (Antoniak and Mimno, 2018), by geographic population (Dunn and Adams, 2020), and even by random iterations on the same data set (Hellrich et al., 2019). The basic implication is that, for instance, English embeddings from web data from South Asia are expected to be quite different from English embeddings from American news articles.

For high-resource languages, such variability is mitigated by the wide availability of in-domain training data for most tasks. But for low-resource languages there is a systematic gap in the kind of training data that is available. For instance, many languages have the Bible (Christodoulopoulos and Steedman, 2015) or related religious literature (Agić and Vulić, 2019) as their largest corpus. The problem is that representations learned from such corpora are likely to be significantly different from those learned from other sources.

How much do representations of low-resource languages depend on the selection of data that happens to be available? To answer this question, this paper uses corpus similarity measures on training data (upstream) to predict differences in trained embeddings (downstream). The basic idea is to model the influence of training data on the variability of embeddings by simulating different low-resource contexts.

This question is important because most languages are relatively low-resource, lacking data sets that contain billions of words. The ability to predict variability in embeddings given training data would enable us to estimate reliability in low-resource languages for which evaluations such as those in this paper are not possible.

2. Experimental Questions

The main contribution of this paper is to evaluate the influence of training corpora on embedding stability for low-resource languages by simulating low-resource and medium-resource settings. We use measures of corpus similarity to determine both (i) relationships between sets of training data and (ii) homogeneity within individual training sets. The basic question is whether we can predict downstream embedding similarity (after training) given upstream corpus similarity (before training). The larger goal is to estimate the reliability of embeddings for low-resource languages, in which systematic

Language	Code	Family	Writing	Morphology
Arabic	ara	Afro-Asiatic	Abjad	Root-Pattern
Indonesian	ind	Austronesian	Alphabet	Agglutinative
Polish	pol	IE:Balto-Slavic	Alphabet	Fusional
Russian	rus	IE:Balto-Slavic	Alphabet	Fusional
German	deu	IE:Germanic	Alphabet	Fusional
Dutch	nld	IE:Germanic	Alphabet	Analytic
Swedish	swe	IE:Germanic	Alphabet	Analytic
Greek	ell	IE:Hellenic	Alphabet	Fusional
Farsi	fas	IE:Indo-Iranian	Abjad	Analytic
French	fra	IE:Romance	Alphabet	Fusional
Italian	ita	IE:Romance	Alphabet	Fusional
Portuguese	por	IE:Romance	Alphabet	Fusional
Spanish	spa	IE:Romance	Alphabet	Fusional
Japanese	jpn	Isolate	Logographic	Agglutinative
Korean	kor	Isolate	Logographic	Agglutinative
Turkish	tur	Turkic	Alphabet	Agglutinative
Finnish	fin	Uralic	Alphabet	Agglutinative

Table 1: Languages Used in Experiments, Sorted By Family, with Writing System and Type of Morphology

evaluations of different downstream embeddings is not possible because of insufficient data. This first section introduces the main experimental conditions and the questions they are used to address.

Source. How does the source of training data impact embedding similarity? We draw training data from three distinct registers: social media, Wikipedia, and web pages. A register is a unique context of production associated with a specific communicative situation (Biber and Conrad, 2009). A long line of research has shown that register has a significant impact on both grammar and the lexicon (Biber, 2012; Biber et al., 2020). Because of the significance of register variation, we expect that embeddings trained from different registers (such as tweets vs Wikipedia articles) will themselves be quite different. While high-resource languages have many registers available for training purposes, low-resource languages often have data from a limited range of registers (e.g., religious or legal documents). This experimental condition, register-specific embeddings, allows us to evaluate whether the context of production has a significant influence on downstream embeddings.

Size. How does the amount of training data impact embedding variability? We evaluate embedding stability over increasing amounts of training data in order to determine whether more data overall is able to compensate for differences in the data. This condition looks at corpora ranging from 10 million words to 100 million words in increments of 10 million. This line of experimentation allows us to simulate low-resource and medium-resource contexts to find out how embeddings change given more training data.

Language Properties. Do specific types of languages have more stable embeddings? The experiments here are conducted across 17 non-English languages as shown in Table 1. These languages represent 10 unique subfamilies, three types of writing system, and four types of morphology. This selection of languages allows us to determine if any of the observed behaviours can be attributed to a specific type of language.

In the next section, we position this current study against related work. We then present the underlying data sets used in the experiments (Section 4) and the methods used for training embeddings and calculating both corpus similarity and embedding similarity (Section 5). We then analyze the impact of different registers (Section 6), the impact of increasing amounts of training data within registers (Section 7), the reliability of low-resource embeddings (Section 8), and the reliability of corpus similarity measures (Section 9). Finally, we consider the implications of this work for natural language processing more broadly (Section 10).

3. Related Work

This section reviews related work on both corpus similarity and embedding stability. First, corpus similarity measures have been used in different applications, such as text classification, information retrieval, and the evaluation of machine translation. Originally defined as a problem within corpus linguistics (Kilgarriff, 1997; Kilgarriff, 2001), many measures have since been proposed (Kilgarriff, 2009; Fothergill et al., 2016; Piperski, 2017; Lu et al., 2020).

One common feature across these approaches is that they are based upon word frequency; in fact, frequency-based approaches have consistently out-performed model-based approaches (Fothergill et al., 2016). More recently, these measures have been used to evaluate fluctuation within and between registers for different language varieties (Dunn, 2021), as well as to classify documents (Nanayakkara and Ranathunga, 2018; Leban et al., 2016) and detect paraphrases in German (Torres-Moreno et al., 2014).

In terms of embedding stability, recent work has used

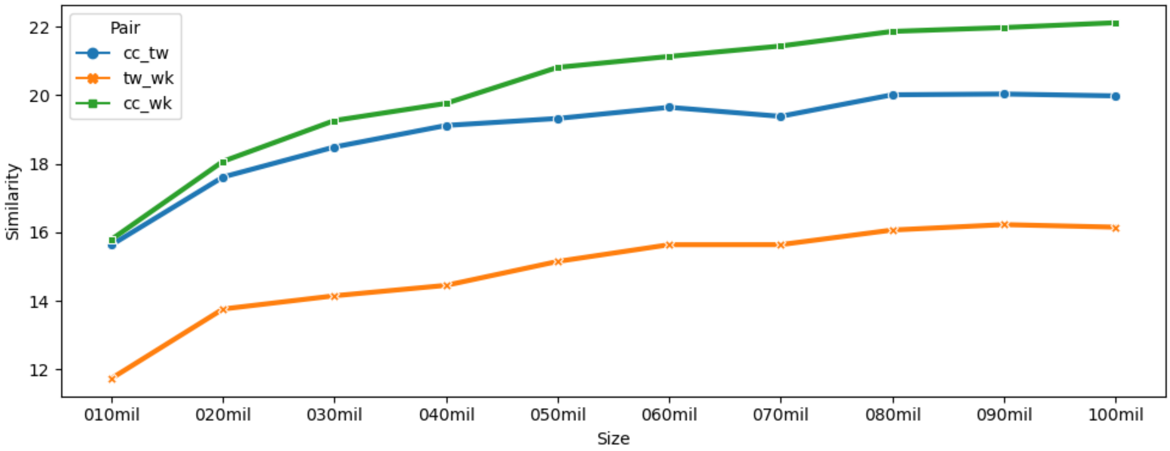

Figure 1: Embedding Similarity Across Registers by Increased Training Data, Arabic

word similarity to investigate variation between embeddings trained on a single corpus (Antoniak and Mimno, 2018), focusing on the training corpus itself as a source of variation. After examining four algorithms and six data sets, results show that corpus size is one of several sources of variability between embeddings. The broader conclusion, shared by the experiments in this paper, is that embeddings represent a specific corpus rather than representing an entire language.

Other work has focused on evaluating whether various factors contribute to the stability of word embeddings and analysing the effects of stability on downstream tasks (Wendlandt et al., 2018). Using a ridge regression model to predict the stability of individual words, these results show that stability within domains is greater than stability across domains. *Domains* in this setting are comparable to registers.

To evaluate the stability of word embeddings derived from a single corpus, (Hellrich et al., 2019) modifies the Singular Value Decomposition algorithm and compares it with other algorithms on three English corpora that ultimately represent distinct registers. The results show that the modified SVD is found to be both reliable and accurate as compared to other algorithms. This study also concludes that stability is positively influenced by corpus size, so that larger sizes lead to higher stability.

Recent work has used linguistic properties to explain the stability of word embeddings across different languages (Burdick et al., 2021). Again using a regression model, this work finds that languages with more complex morphology tend to be less stable than languages with simpler morphology. That finding is not replicated in this present study, although the issue is raised by the case of Arabic in Section 8. Most other work, however, focuses only on English (Antoniak and Mimno, 2018; Wendlandt et al., 2018; Hellrich et al., 2019).

This paper is unique in its cross-linguistic, cross-register, and cross-size experimental design. This approach allows us to determine in a systematic manner which properties of corpora influence embedding stability.

4. Data

The data for these experiments comes from comparable corpora representing 17 languages. The Wikipedia register (WK) is collected from the public Wikimedia dump of March 2020. Languages are identified using the designation provided by Wikipedia. The web register (CC) is collected from the Corpus of Global Language Use (Dunn, 2020), ultimately derived from the Common Crawl. The social media register (TW) is collected from geo-referenced tweets. For both web pages and tweets, languages are identified using the idNet package (Dunn, 2020). A list of languages is provided in Table 1, including the family, type of writing system, and type of morphology for each language. Each source of data (i.e., register) provides a corpus of 100 million words.

5. Methods

The experiments in this paper depend on two measures: similarity between embeddings and similarity between training corpora. Following previous work (Wendlandt et al., 2018), we calculate the similarity between two sets of embeddings by taking the aggregate overlap of nearest neighbors. This is calculated as follows: first, we create an independent corpus for each language, representing different registers from the main experimental data. These background corpora contain movie subtitles, news commentary articles, and Bible translations (Tiedemann, 2012; Christodoulopoulos and Steedman, 2015). Thus, each language is represented by the three registers described above (WK, TW, CC) in addition to these separate out-of-domain corpora. We find the 1,000 most common lexical items in these background corpora. For each of the top lexical items, we then retrieve the ten nearest neighbors from each set of embeddings.

The overlap for each lexical item is the percentage of words which appear as neighbors within both sets of embeddings. For example, if *dog* is a nearest neighbor for *cat* in both tweet-based and web-based embeddings, this indicates a certain similarity between those sets

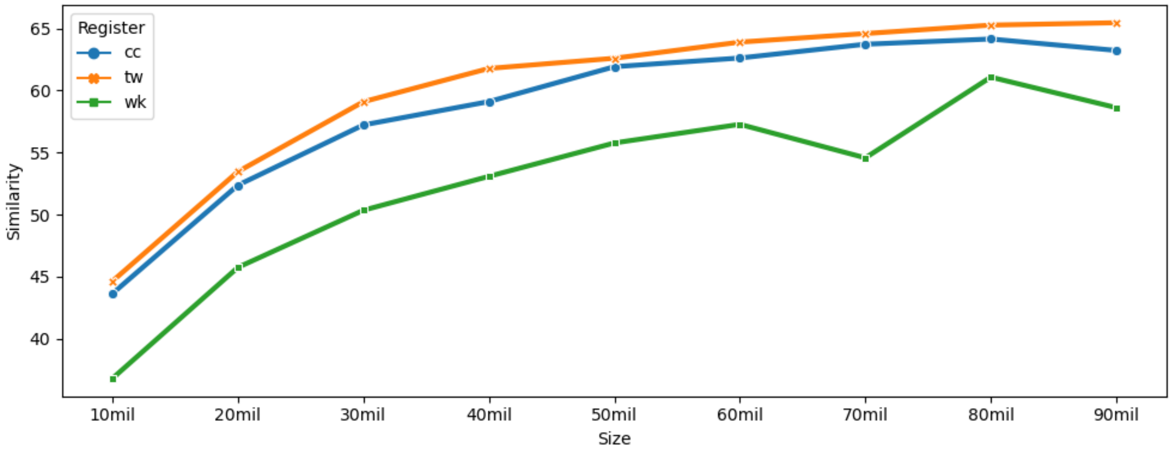

Figure 2: Embedding Stability Within Registers by Increased Training Data, Arabic

of embeddings. Thus, higher overlap scores indicate higher agreement. The overall embedding similarity between two conditions (such as tweets vs Wikipedia at 50 million words) is represented using the average overlap across these top 1,000 words.

The embeddings themselves are character-based, trained using the skip-gram negative sampling method with 50 negative examples per observation and trained for 20 epochs. Implemented using the fastText framework (Mikolov et al., 2018), these popular embeddings have the advantage of being character-based, which we expect to reduce differences that are caused by morphology or writing system (c.f., Table 1).

We also need to measure the similarity between the training corpora themselves, before training takes place (Kilgarriff, 1997; Fothergill et al., 2016). Recent work on corpus similarity measures has shown that a frequency-based approach with 5k bag-of-words features and Spearman's *rho* performs well across many languages (Dunn, 2021). A frequency-based approach to corpus similarity uses a vector of n-gram frequencies, usually restricted to the most common n-grams. Spearman's *rho* has been shown to be highly accurate in comparing these frequency vectors, with more similar corpora having a higher correlation coefficient. Unlike other measures (Kilgarriff, 2001), Spearman's *rho* is not dependent on corpus size. We use the same background corpora as above to select the top n-gram features for each language.

While the embedding similarity measure has been evaluated before (Wendlandt et al., 2018), we provide an evaluation of the corpus similarity measure here. The underlying task for validation is to predict whether two samples come from the same register: for example, given two samples, do both represent tweets? A high accuracy means that the corpus similarity measures always distinguish between data sources. To convert this continuous measure into a categorical prediction, we set a threshold; the more often this threshold leads to correct predictions, the more accurate the measure is. The threshold calculation takes the lowest average similarity for same-register pairs (for example, CC-CC) and the highest average similarity for cross-register pairs (for example, CC-TW). The threshold is set halfway between these minimum and maximum values (Ali, 2011; Leban et al., 2016). We conduct the evaluation using a five-fold cross-validation design, with the threshold calculated on the training data for each fold. Each *sample* in this design is a 20k word sub-set of a corpus.

Language	Family	Features	Acc.
Arabic	Afro-Asiatic	C4	99%
German	IE:Germanic	C4	98%
Greek	IE:Hellenic	W1	97%
Farsi	IE:Indo-Iranian	W1	96%
Finnish	Uralic	C4	94%
French	IE:Romance	W1	100%
Indonesian	Austronesian	C4	99%
Italian	IE:Romance	W1	94%
Japanese	Isolate	C2	88%
Korean	Isolate	C4	99%
Dutch	IE:Germanic	W1	100%
Polish	IE:Balto-Slavic	W1	99%
Portuguese	IE:Romance	C4	98%
Russian	IE:Balto-Slavic	C4	100%
Spanish	IE:Romance	C4	99%
Swedish	IE:Germanic	C4	96%
Turkish	Turkic	C4	100%

Table 2: Accuracy of Corpus Similarity Measures with Feature Type in a Register Identification Task

The accuracy of this register-prediction task is shown in Table 2, along with the best type of feature (word or character n-gram size). While there is some variation in performance, with Japanese being particularly low, the generally high performance provides a validation of the corpus similarity measure. This is important because it means that these measures are indeed able to distinguish between corpora representing the three registers used in these experiments (CC, TW, WK). In other words, these results show that it is possible to measure differences between training corpora before we train embeddings.

While the corpus similarity measure performs well, the scores are not directly comparable across languages because each language has a different central tendency. For comparison across different sources (i.e., the web vs tweets), we retain the raw similarity value and restrict ourselves to within-language comparisons. For comparison within sources (i.e., the web vs the web), we use the z-score to standardize the measure across all registers, in order to make better downstream predictions (c.f., Section 8). Finally, the similarity between large corpora are estimated by sampling 200 unique pairs of sub-corpora, each containing 20k words. We then find the mean similarity across all samples to represent the relationship between the larger corpora. A Python package for reproducing these corpus similarity measures is available here.

6. Experiment 1: Register

The first experiment asks whether more similar corpora produce more similar embeddings. In other words, is there a relationship between the input (a corpus) and the output (word embeddings)? A different way of asking this same question is whether register variation, a long-studied linguistic phenomenon, has a predictable impact on embeddings. We create three cross-register comparisons for each language: CC-TW, TW-WK, and CC-WK. For each comparison, we compute both the corpus similarity (upstream) and the embedding similarity (downstream). We can visualize embedding similarity in this context as in Figure 1, which shows similarity for each pair of register-specific embeddings (y-axis) over increasing amounts of training data (x-axis) for Arabic. Each line here represents the similarity between two different sets of embeddings. We see, for instance, that all register-specific embeddings become more similar as the amount of training data increases. At the same time, the tweet-based and Wikipedia-based embeddings (in yellow) are much less similar than the others. This indicates that register variation does, in fact, have an impact on these sets of embeddings.

The main question here concerns the similarity between each pair of register-specific embeddings at different data sizes (i.e., at 100 million words). We see across languages that there is a clear set of relationships between embeddings trained on different corpora: register has a significant impact on embeddings, as we expect from previous work on register variation. The full set of figures for each language is provided in the supplementary material, available here. The question, however, is whether we can predict these relationships between embeddings using corpus similarity measures.

We take a closer look in Table 3 with Finnish. For each pair of registers, in the first column, we see both the similarity between embeddings and the similarity between the training corpora. For both measures, higher values indicate higher similarity; the scales, however, are quite different. What we see here is that the same pair of registers (CC-TW) produce the most similar embeddings (24.80% overlap) and also have the highest corpus similarity score (0.72). In fact, there is a strong correlation of 0.999 between these two sets of values. This means that, for Finnish, we can predict which embeddings will be more similar even before we train them.

Register Pair	**Embedding Sim.**	**Corpus Sim.**
CC-TW	24.8	0.72
TW-WK	14.3	0.59
CC-WK	19.6	0.66

Table 3: Relationship between Embedding Similarity and Corpus Similarity for Finnish

Language	**Family**	**10 mil**	**100 mil**
Arabic	Afro-Asiatic	0.727	0.909
German	IE:Germanic	0.977	0.993
Greek	IE:Hellenic	0.817	0.914
Farsi	IE:Indo-Iranian	0.591	0.729
Finnish	Uralic	0.988	1.000
French	IE:Romance	0.947	0.924
Indonesian	Austronesian	0.988	0.981
Italian	IE:Romance	0.986	0.928
Japanese	Isolate	0.994	0.971
Korean	Isolate	0.541	0.972
Dutch	IE:Germanic	1.000	0.981
Polish	IE:Balto-Slavic	0.947	0.898
Portuguese	IE:Romance	0.788	0.863
Russian	IE:Balto-Slavic	0.836	1.000
Spanish	IE:Romance	0.838	0.966
Swedish	IE:Germanic	0.999	0.995
Turkish	Turkic	0.678	0.906
Average	**All**	**0.861**	**0.937**

Table 4: Relationship between Embedding Similarity and Corpus Similarity

We take a cross-linguistic view of this relationship between input and output in Table 4. The quantity we are interested in is the relationship between the two measures, corpus similarity and embedding similarity: how well could we predict embedding similarity downstream given corpus similarity upstream? This table shows the relationship both at 10 million words and at 100 million words. Note that the corpus similarity measure remains quite stable across corpus size (c.f., Section 9) while embeddings become more similar given more training data (c.f., Figure 1). We see that the relationship becomes stronger as the embeddings have more training data, from an average correlation of 0.86 to 0.93. This is because the embeddings themselves become more stable with increased training data (c.f., Section 7).

Analyzing Table 4, we can use the properties of each language to understand what causes specific outliers. For example, the relationship for Korean is quite low at 10 million words but rather high at 100 million words.

We might think this is caused by the logographic writing system, but that same pattern is not shown in Japanese. The lowest relationship is shown by Farsi, with a maximum correlation of 0.729. However, we know that this is not caused by the writing system (shared with Arabic) or by the type of morphology (shared by several other languages). This would rather seem to be a specific property of Farsi corpora, rather than a property based on either Farsi's writing system or morphology.

This section has shown two important properties of the impact of register variation on embeddings: First, across all languages there is a significant difference between register-specific embeddings. This means that it is more accurate to formulate Arabic-Wikipedia embeddings rather than universal Arabic embeddings: the downstream embeddings remain register-specific, at least with this amount of training data. In other words, register variation has a consistent impact downstream on trained embeddings. Second, there is a strong relationship between the training corpora themselves and the similarity between embeddings trained on those corpora. In other words, more similar corpora produce more similar sets of embeddings. This is an important finding because it suggests that we should be able to predict the conditions under which embeddings will be both stable and reliable.

7. Experiment 2: Size

The second experiment quantifies the amount of change that occurs within register-specific embeddings as the amount of training data is increased. The central question is whether it is possible to predict the growth curve of embedding stability, the rate at which embeddings become more similar when trained from different subsets of the same corpus. For example, Figure 2 shows embedding similarity by size within each register for Arabic. Here each line represents a single register, with the comparison made between embeddings trained using different amounts of data. For example, the green line represents Wikipedia. With less data, the agreement between the 10 million and 20 million word conditions (on the left) is below 40%; but with more data the agreement between the 90 million and 100 million word conditions (on the right) is closer to 60%.

We refer to this as *stability* because the two corpora overlap to a large degree. This gives us a baseline for stability within each language: cross-register similarity, for example, should never exceed within-register similarity in this setting. As before, we expect that more training data leads to more stable embeddings; this trend is found across all languages. The more meaningful question, however, is whether we can predict this increasing rate of stability using corpus similarity.

In other words, we might expect that more homogeneous corpora, data sets that are more self-similar, will produce more stable embeddings because they contain less internal variation. We test this hypothesis in two ways: First, we take the amount of increase in embedding stability for each condition. For example, Arabic web corpora have an increase of 20.49% in embedding stability between the 10-20 million and 90-100 million word comparisons. But the Arabic Wikipedia corpora have a higher increase of 24.26%. To test whether we can predict the amount of increased stability, we look at the correlation between (i) the increased stability of within-register embeddings and (ii) the homogeneity of the training corpus. *Homogeneity* here is the same corpus similarity measure calculated across 200 unique chunks from a single corpus. However, there is no consistent relationship. A second approach looks at the relationship between the slope of increased stability across all conditions and corpus homogeneity. Again, there is no consistent relationship across languages.

Language	Feature	CC	TW	WK
Arabic	C4	63.22	65.45	58.61
German	C4	64.28	67.88	59.71
Greek	W1	62.99	68.21	57.59
Farsi	W1	62.53	68.87	57.51
Finnish	C4	63.34	67.94	57.92
French	W1	58.99	65.63	51.71
Indonesian	C4	67.77	57.90	60.79
Italian	W1	65.16	62.66	62.65
Japanese	C2	54.81	59.42	50.29
Korean	C4	59.26	62.74	55.61
Dutch	W1	64.56	67.17	60.43
Polish	W1	66.50	69.78	52.93
Portuguese	C4	64.80	60.10	57.58
Russian	C4	57.31	67.96	60.18
Spanish	C4	67.39	70.59	60.43
Swedish	C4	65.02	65.64	52.89
Turkish	C4	61.40	66.14	54.62
Average		**62.90**	**65.53**	**57.14**

Table 5: Embedding Stability by Language and Register, 90 million and 100 million word comparison

This experiment therefore reaches a negative result: it is not possible to use corpus homogeneity to predict embedding stability in this context. Table 5 shows within-register similarity at the 90-100 million word comparison condition. On the one hand, for each condition there is higher similarity in this same-register comparison than we saw in the previous cross-register comparison, as we would expect. However, it is not possible to predict the rate or the degree of increased embedding stability given corpus homogeneity.

8. Experiment 3: Reliability

The third experiment asks whether we can predict the reliability of embeddings in very low-resource settings. First, we train embeddings using only 1 million words from each language in each register: 1 million words of Arabic tweets, for example. We repeat this process multiple times and calculate similarity between ten unique pairs of same-register embeddings in each language. For many low-resource languages we have only 1 million

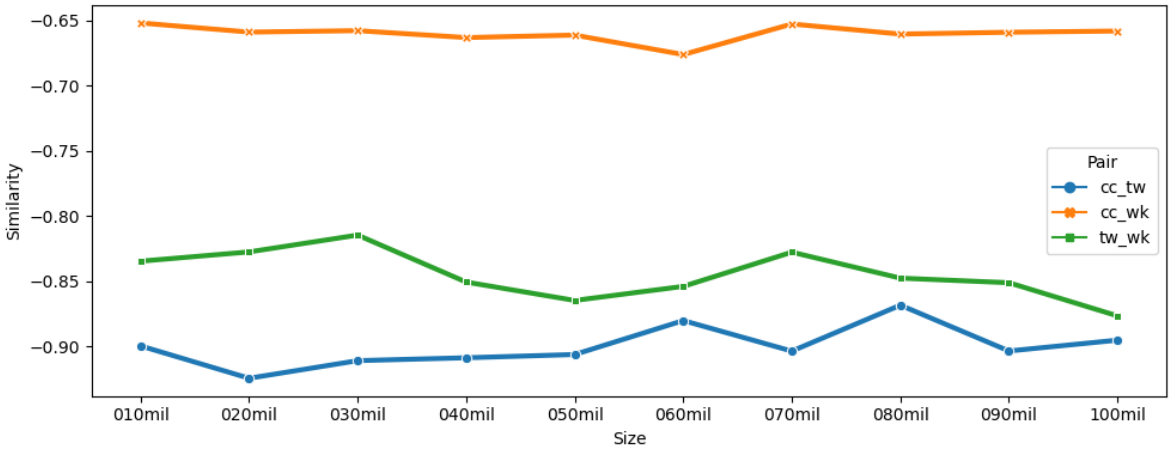

Figure 3: Corpus Similarity Between Registers by Increased Training Data, Arabic

Register	Reliability	Homogeneity
CC	21.63	-0.28
TW	26.42	-0.19
WK	16.88	-0.36

Table 6: Relationship between Embedding Reliability and Corpus Homogeneity for Italian

words available, so that embeddings must be trained on whatever data there is. This experiment simulates many different samples from high-resource languages in order to capture the distribution of embedding similarity values in a low-resource setting: how different would register-specific embeddings have been if we had instead observed some other sub-set of the data?

As we have seen, larger training sets lead to more stable embeddings. This means, for example, that representations based on a very small amount of data are expected to be less stable. Our question here is about degrees of instability: can we predict which low-resource embeddings will be more reliable across different samples? The basic idea is that, if such predictions are possible, we can estimate reliability in truly low-resource settings for which multiple training sets are not available.

We use Italian as an example in Table 6, showing the reliability of embeddings in a low-resource setting together with the homogeneity of the training corpora. *Reliability* here is the average agreement across 10 random pairs of embeddings, each representing a unique sub-set of the same corpus. The overall relationship in this case is quite strong, a correlation of 0.999. For example, the most homogeneous corpus (TW), has the highest reliability across random pairs of embeddings. Note that the homogeneity scores have been standardized using the z-score across all register combinations; thus, the mean is 0 with values above 0 indicating high similarity and values below 0 indicating lower similarity. The central question is whether corpus homogeneity can be used to predict the reliability of low-resource embeddings.

Language	Family	Correlation
Arabic	Afro-Asiatic	0.567
German	IE:Germanic	0.897
Greek	IE:Hellenic	0.827
Farsi	IE:Indo-Iranian	0.993
Finnish	Uralic	0.847
French	IE:Romance	0.979
Indonesian	Austronesian	0.976
Italian	IE:Romance	0.999
Japanese	Isolate	0.999
Korean	Isolate	1.000
Dutch	IE:Germanic	0.976
Polish	IE:Balto-Slavic	0.890
Portuguese	IE:Romance	0.910
Russian	IE:Balto-Slavic	0.786
Spanish	IE:Romance	0.999
Swedish	IE:Germanic	0.962
Turkish	Turkic	0.972
Average	**All**	**0.916**

Table 7: Relationship between Embedding Reliability and Corpus Homogeneity

The relationship between homogeneity and embedding reliability in low-resource settings is shown in Table 7. This relationship is rather strong on average, 0.916, allowing us to predict the conditions under which embeddings in a low-resource setting will be more reliable. This finding is important because, if corpus similarity measures remain robust across different data sizes, we could predict the degree of confidence we should have for embeddings trained in truly low-resource contexts.

While this relationship is strong in general, there are some clear outliers: for example, in Arabic the correlation is only 0.567. This means that there is only a weak relationship between corpus homogeneity and the reliability of low-resource embeddings. Russian is another language with a relatively weak relationship, in this case 0.786. While Arabic is the only Afro-Asiatic language in this data set, the other Balto-Slavic language (Polish) has a much stronger relationship than Russian. One potential factor is that Arabic morphology, a unique root-and-pattern system, is not shared by any other language in this study. This raises a question for future work about whether related languages like Hebrew would evidence this same characteristic.

The experiment in this section has shown that most languages have a strong relationship between the homogeneity of the training corpus and the reliability of very low-resource embeddings (trained on only 1 million words). The implications of this relationship are important because it suggests that we can predict the reliability of low-resource embeddings.

9. Experiment 4: Corpus Similarity

The final experiment asks how much change we see in corpus similarity measures given the size of the data set. We have calculated corpus similarity by dividing a corpus into many chunks of 20k words and calculating pairwise similarity between these chunks. That means we are estimating the overall similarity by sampling a number of individual observations (200 as implemented here). Here we measure the stability of corpus similarity measures over corpus size, as shown for Arabic in Figure 3. In this figure, the y-axis represents corpus similarity and the x-axis represents the size of the corpus. Overall, the measures are quite stable.

Language	**Ftr.**	CC-TW	CC-WK	TW-WK
Arabic	C4	0.056	0.024	0.062
German	C4	0.040	0.051	0.036
Greek	W1	0.033	0.037	0.038
Farsi	W1	0.041	0.023	0.061
Finnish	C4	0.014	0.031	0.066
French	W1	0.051	0.045	0.068
Indonesian	C4	0.069	0.045	0.093
Italian	W1	0.031	0.035	0.033
Japanese	C2	0.025	0.028	0.038
Korean	C4	0.046	0.063	0.058
Dutch	W1	0.023	0.038	0.042
Polish	W1	0.022	0.069	0.059
Portuguese	C4	0.071	0.030	0.056
Russian	C4	0.067	0.069	0.071
Spanish	C4	0.019	0.040	0.055
Swedish	C4	0.014	0.054	0.040
Turkish	C4	0.045	0.028	0.104
Average	**All**	**0.039**	**0.041**	**0.057**

Table 8: $Max - Min$ of Similarity Across Sizes

We explore the stability of corpus similarity measures in Table 8. This table takes the average corpus similarity value across each amount of data (10 million words, 20 million words, and so on) and then finds the range between the maximum similarity and the minimum similarity for each condition: for example, the similarity between Arabic tweets and Arabic Wikipedia articles at different sample sizes. This table shows that there is only a small variation across sample size in each condition. We further test this using a one-sample t-test to see if each condition actually constitutes a single population. In all cases, there is no significant difference among the population of corpus similarity values by size, a confirmation of the visual trend in Figure 3. Thus, corpus similarity measures are robust regardless of the size of the corpus.

10. Conclusions

This paper has simulated low-resource settings in a cross-lingual and cross-register context in order to measure the similarity, stability, and reliability of embeddings. The basic idea has been to examine the ability of corpus similarity measures, applied to training data, to predict downstream differences in embeddings. The important background is that corpus similarity measures remain stable across corpus size, so that they can be applied in truly low-resource settings.

The first findings, in Sections 6 and 7, showed that (i) register-specific embeddings are significantly different and (ii) that embeddings become more stable within registers as the amount of training data increases. Both of these findings are expected. The important new contribution is the fact that the degree of difference in register-specific embeddings can in fact be predicted by differences in the training corpora themselves. Because low-resource languages have a reduced inventory of register-specific corpora, it is not possible to directly measure the impact of register on embeddings in such languages. Corpus similarity measures thus allow us to indirectly measure the impact of register in truly low-resource settings.

The second important finding, in Section 8, is that the stability of low-resource embeddings can be predicted given corpus homogeneity measures. In a truly low-resource setting, we would never be able to measure embedding reliability because of data limitations. We can, however, measure corpus homogeneity even with limited corpus sizes (c.f., Section 9). The combination of these two findings, then, means that it is possible to predict which low-resource embeddings are more reliable and which are less reliable. This constitutes a significant advance in validating low-resource language resources, providing a measure of confidence for embeddings trained from small corpora.

The caveat of the experiments in this paper, however, is that we have focused on *simulated* low-resource settings rather than *actual* low-resource settings. This is a necessary choice given the need to undertake a large number of comparisons within each language. Further, the 17 languages used here represent a range of language families and types of morphology, but we know that truly low-resource languages often belong to families that are not represented here.

A Python package is made available for working with these corpus similarity measures and the full experimental results are available in the supplementary material.

11. Bibliographical References

Adelmann, B., Menzel, W., and Zinsmeister, H. (2021). The Impact of Word Embeddings on Neural Dependency Parsing. In *Proceedings of the 17th Conference on Natural Language Processing*, pages 1–13, Düsseldorf, Germany. KONVENS 2021 Organizers.

Agić, Ž. and Vulić, I. (2019). JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, jul. Association for Computational Linguistics.

Ali, A. (2011). *Textual similarity*. Bachelors thesis, Technical University of Denmark.

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

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press, Cambridge, UK.

Biber, D., Egbert, J., and Keller, D. (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.

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

Burdick, L., Kummerfeld, J. K., and Mihalcea, R. (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, Online and Punta Cana, Dominican Republic, nov. Association for Computational Linguistics.

Christodoulopoulos, C. and Steedman, M. (2015). A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Dunn, J. and Adams, B. (2020). Geographically-Balanced Gigaword Corpora for 50 Language Varieties. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2528–2536, Marseille, France, may. European Language Resources Association.

Dunn, J. (2020). Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54:999–1018.

Dunn, J. (2021). Representations of Language Varieties Are Reliable Given Corpus Similarity Measures. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–38. Association for Computational Linguistics.

Fothergill, R., Cook, P., and Baldwin, T. (2016). Evaluating a Topic Modelling Approach to Measuring Corpus Similarity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 273–279, Portorož, Slovenia, may. European Language Resources Association (ELRA).

Hellrich, J., Kampe, B., and Hahn, U. (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.

Kilgarriff, A. (1997). Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. In *Proceedings of the Fifth Workshop on Very Large Corpora*, pages 231–245. Association for Computational Linguistics.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK. University of Liverpool.

Leban, G., Fortuna, B., and Grobelnik, M. (2016). Using News Articles for Real-time Cross-Lingual Event Detection and Filtering. In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval*, volume 1568, pages 33–38. CEUR Workshop Proceedings.

Lu, J., Henchion, M., and Mac Namee, B. (2020). Diverging Divergences: Examining Variants of Jensen Shannon Divergence for Corpus Comparison Tasks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6740–6744, Marseille, France, may. European Language Resources Association.

Miaschi, A. and Dell'Orletta, F. (2020). Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online, jul. Association for Computational Linguistics.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 52–55. European Language Resources Association.

Nanayakkara, P. and Ranathunga, S. (2018). Clustering Sinhala News Articles Using Corpus-Based Similarity Measures. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 437–442.

Piperski, A. (2017). Sum of Minimum Frequencies as a Measure of Corpus Similarity. In *Proceedings of the Corpus Linguistics Conference*, Birmingham, UK.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the International Conference on Language Resources and Evaluation*, page 2214–2218. European Language Resources Association.

Torres-Moreno, J.-M., Sierra, G., and Peinl, P. (2014). A German Corpus for Text Similarity Detection Tasks.

International Journal of Computational Linguistics and Applications, 5:9–24.

Wendlandt, L., Kummerfeld, J. K., and Mihalcea, R. (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, New Orleans, Louisiana, jun. Association for Computational Linguistics.

12. Language Resource References

Dunn, J.; Li, H.; & Sastre, D. (2022). *Corpus Similarity: A Python package*. https://github.com/jonathandunn/corpus_similarity.