CSE 517: Project - Final Report
Analyzing the Surprising Variability in Word Embedding Stability Across Languages

Prashant Rangarajan, Neha Kardam
{prashr, nehakrdm}@uw.edu

Reproducibility Summary

Scope of Reproducibility

Two datasets are used for the glove embedding: Wikipedia and the Bible Corpus. In the case of the larger Wikipedia Corpus, the stability was determined using 40 languages in the original paper. We limit ourselves to a subset of these languages, due to datasize and computation time. Each language data set was massive in size; for example, the full English text was 10 gigabytes. To run the code and calculate the stability for the Wikipedia corpus, some parameters were changed for example number of iterations and downsample size. Due to the small size of the Bible corpus, all available languages (~ 100) could be used to determine stability. We evaluated the stability of the dataset for various languages using an easily accessible online version. If we can run it on a GPU server, we can try increasing the number of iterations and bringing it closer to the original mode.

Methodology

A stable embedding measures how changes in the input data or training method impact the output embeddings. We produce data in this research by downsampling a corpus into distinct subsets of smaller corpora and then analyzing stability across these groups. The primary goal here is to see if consistency among down samples produces consistent results that can be compared across languages. We look at the stability of Wikipedia and the Bible in the 26 languages for comparison. Stability is separated into 5 percent buckets, such as 0-5 percent, 5-10 percent, and 95-100 percent. Number of words in each bucket is counted and normalized by the total number of words. Additionally, we examine down sampling without replacement for various down sample sizes.

Results

We were able to to find stability for some languages in the wiki text (downsample without replacement) and for the bible corpus. We were also able to try out a different dataset- the Quran and got similar trends for stability.

What was Easy

The author has provided a code that is simple to understand and interpret. That aided in verification of the code.

What was Difficult

It was challenging to find the original data set from the authors. The large data set for various languages made the computation time lengthy, which hopefully can be resolved if GPU is used. Some libraries used in the code are incompatible with the Python version we were using.

Communication with Original Authors

For Bible corpus, we used a version that is freely available online Christodoulopoulos and Steedman [2015]. However, it is slightly different from the version of the dataset was used in the paper, which we were unable to obtain from them (but gave similar results).

1 Introduction

In this project, we will be analyzing the stability of word embeddings across various languages as described in Burdick et al. [2021].

Word embeddings such as word2vec Mikolov et al. [2013] and GloVe Pennington et al. [2014] are low-dimensional and dense vector representations that are now ubiquitous with the advent of neural network models and larger datasets. They are strong representations that serve as the foundation for a large number of natural language processing systems in both English and other languages. These embeddings can be influenced by many factors, such as the order of the dataset used to train the embedding (curriculum), batching etc. Analyzing these factors is pertinent as they impact the downstream models in which these embeddings are used.

This paper introduces the concept of stability as a property of the embedding space of the word vectors. Stability represents how much changing the input data affects the resultant embeddings. We define stability by examining how closely words in the word vector space model are clustered with each other. As a rule of thumb, stability is defined as the degree to which two embedding spaces are in perfect agreement. For example, for a word W and two embedding spaces A and B, there are ten closest neighbors of W in both A and B. The percentage of overlap between the two lists of closest neighbors represents W's stability. A stability of 100 percent shows complete agreement between the two embedding spaces, whereas a stability of zero percent suggests complete disagreement.

Burdick et al. [2021] investigates the impact of word stability on embedding behavior in a variety of languages. It examines the linguistic traits associated with stability and draws conclusions about their relationships by making use of regression modelling.

2 Scope of Reproducibility

Burdick et al. [2021] make use of 2 datasets - the Wikipedia and the Bible Corpus (described in detail in the Datasets section).

In the case of the (larger) Wikipedia Corpus, stability was calculated for 40 languages. However, for this version of the project report, we tested out only a subset of the languages due to computational restrictions. Each language data set was enormous in size; for example, the full text for English was 10 gigabytes. However, despite using only some languages we were still able to obtain the general trends in the model.

In the case of the Bible corpus, the dataset was not very large, hence all the available languages (~ 100) could be used in order to find the stability. We used a version of the dataset easily available online to evaluate the stability for different languages. For the final proposal, we will try to use the original version of the corpus (if possible) for which we need to contact McCarthy et al. [2020].

To calculate the stability for the Wikipedia corpus some parameters needed to be modified in order to run the code as compared to the original. For example, the GloVe embedding uses 100 iterations to calculate stability; we reduced iteration size to 40, then to 10, in order to run the script successfully on a laptop. If we are able to run it on a GPU, we may try to increase the number of iterations and make it closer to the original model.

2.1 Addressed Claims from the Original Paper

1. The author bucketed the stability of Wikipedia and Bible and got certain trends in the percentage of words i each stability bucket, that we would like to replicate (like most of the words being in the middle range of stability buckets)
2. When comparing GloVe downsamples on Wikipedia, certain languages like Vietnamese has the most stable embeddings whereas Korean has the least stable embeddings. We intend to use the same implementation and see if our results agree with or differ from the authors' claims Burdick et al. [2021].
3. The regression model used has a high R2 score of 96 percent, indicating that the model fits the data well. The study presents significant weights with the greatest magnitude, suggesting that negative weights belong to poor stability and positive weights correspond to high stability. We'd like to perform a similar analysis using other regression models and hyperparameters.
4. Certain features like affixing and gender attributes were found to influence stability - we will investigate if these claims continue to hold.

We will investigate the claims made by the paper listed above and any additional experiments.

3 Methodology

In this project we investigate the relationship between the stability and linguistic traits in various languages.

3.1 Model Descriptions

We use regression modeling to capture correlations between linguistic qualities and a language's average stability, and we get insights into how linguistic variables relate to stability. For example, the languages with greater affixing are less stable. This project employs the word embedding GloVe and word2vec to explore linguistic features and trends. It is important for researchers using the unstable embeddings to account for this in their methodology and error analysis Burdick et al. [2021].

A stable embedding is a measure of how changes in input data or training method affect the resulting embeddings. In some cases, to shift the embeddings, we apply changes such as increasing the size of the context window in order to obtain embeddings that capture semantics better than syntax. However, in other cases, such as changing the algorithm's random seed, we expect there to be little effect on our embeddings.

In this project, we generate data by down sampling a corpus into different subsets of smaller corpora and then assessing stability across these subsets. The option of whether to sample with or without replacement, as well as the size of the sample, is a subtle methodological decision. The main purpose here is to test if consistency among down samples generates consistent findings that can be compared across languages. To begin, we selected 5 sets (down samples) of 500,000 sentences from various languages. This collection contains sentences that were randomly picked and included in all five down samples such that a certain percentage of all the sets had these overlapping sentences. To verify stability, GloVe embedding is employed with 100 iterations, 300 dimensions, a window size of 5, and a minimum word count of 5.

First, stability is divided into buckets of 5 percent width, such as 0-5 percent, 5-10 percent, 95-100 percent. This allows us to observe patterns in stability that a single statistic, such as the overall average, would not reveal. The number of words in each stability bucket is counted and normalized by the total number of words.

Second, we investigate down sampling without replacement for various down sample sizes, such as 50,000, 100,000, and so on. The original authors discovered that down sampling without replacement gives more consistent (and hence equivalent) stability results than down sampling with replacement. Therefore, the original research considers only down sampling without replacement.

We assess stability across the 26 languages covered by both the Wikipedia and the Bible corpora. These findings illustrate three settings for Wikipedia:

1. Stability of GloVe embeddings over five down sampled corpora,
2. Stability of word2vec (w2v) embeddings over five down sampled corpora, and
3. Stability of word2vec embeddings on one down sampled corpus using five random seeds.

We show just the third case for the Bible, since it is too small for down sampling.

3.2 Datasets

In this paper, we use two different datasets, the Wikipedia and the Bible corpus. The Wikipedia dataset Al-Rfou et al. [2013] is larger and is a comparable corpus. It spans 40 languages. It is available in `https://sites.google.com/site/rmyeid/projects/polyglot`. On the other hand, the Bible corpus Christodoulopoulos and Steedman [2015] has data in more languages (97 languages) and is a parallel corpus, available online in `https://christos-c.com/bible/`. In addition, we use the World Atlas of Language Structures (WALS) Dryer and Haspelmath [2013], a database of linguistic features for over 2000 languages that has been curated by a team of 55 authors. It is available online at `https://wals.info/`.

While testing out the code for obtaining the stability, we used four languages for the Wikipedia corpus: English, Arabic, Bulgaria and Hindi. The total text size for English was close to 10GB, Arabic 533MB, Bulgarian 400MB and 300 MB for Hindi. All the four languages have five down samples without a replacement. Each Arabic subset had 58678, Bulgaria 42576 and Hindi subset contains 28897 sentences, whereas each English subset contains 880835 sentences.

In the case of the Bible Corpus, the average size of the corpus was around 3.6MB for each language, with the largest file size being 12MB, and no downsampling was required in order to obtain the stability values for each language.

The additional dataset for this project was obtained from the Quran. 25 Ayah from the Quran were translated into four different languages: English, Hindi, Bulgarian, and Arabic. The text files from the four languages were manually created using Fakhrezi et al. [2021] resource as a reference.

3.3 Hyperparameters

To create the five-word embedding in case of Wikipedia we changed the iteration level from 100 to 50 to see if that makes difference in assigning embedding. We also changed the hyper parameter when running the regression model; originally, it was set to run for 1000 iterations, which was reduced to 10, and the alpha value was reduced to the standard 0.1 from 10. Furthermore, we used different hyper - parameters for regression models for Wikipedia data, such as Lasso regression with an alpha level of 0.7 and Elastic regression with an alpha level of 0.5.

3.4 Implementation

We used the Github code provided by the authors Burdick [2021] to obtain the stability.

For Wikipedia, we first created five GloVe embedding spaces (downsampling without replacement) for the English and Hindi languages, and then ran the author's script `trainGlove_wikipedia.sh`. It took about 1-2 hours to download two languages' full text (we used only two to determine stability though) and write the code to down sample to five subsets, and nearly a day to get the shell script working because the program kept crashing due to the large data size. Therefore, we experimented by reducing the down sample size.

In the case of the Bible corpus, no downsampling was required. However, we did need to use a script to find out the set of languages that had at least 75% of the Bible in the text to be used in future steps.

The next step was to precalculate the five nearest neighbors for each word. We used the scripts `precalculateNearestNeighbors_bible.py` and `precalculateNearestNeighbors_wikipedia.py`. The code makes use of the FAISS library Johnson et al. [2021] in order to compute nearest neighbors and similarity search in an efficient manner. For this step, a significant amount of time was spent in this step figuring out how to get the library to cooperate with the Python code, due to some version incompatibilities.

The final step was to use `stability_bible.py` and `stability_wikipedia.py` to calculate stability for each language in the Bible and Wikipedia.

3.5 Experimental Setup

The experiments were run on the two CPUs. The data from Wikipedia and the Bible were divided among team members to code and analyze, and the results were later compiled. Here's a link Rangarajan [2022] to the code and notebook.

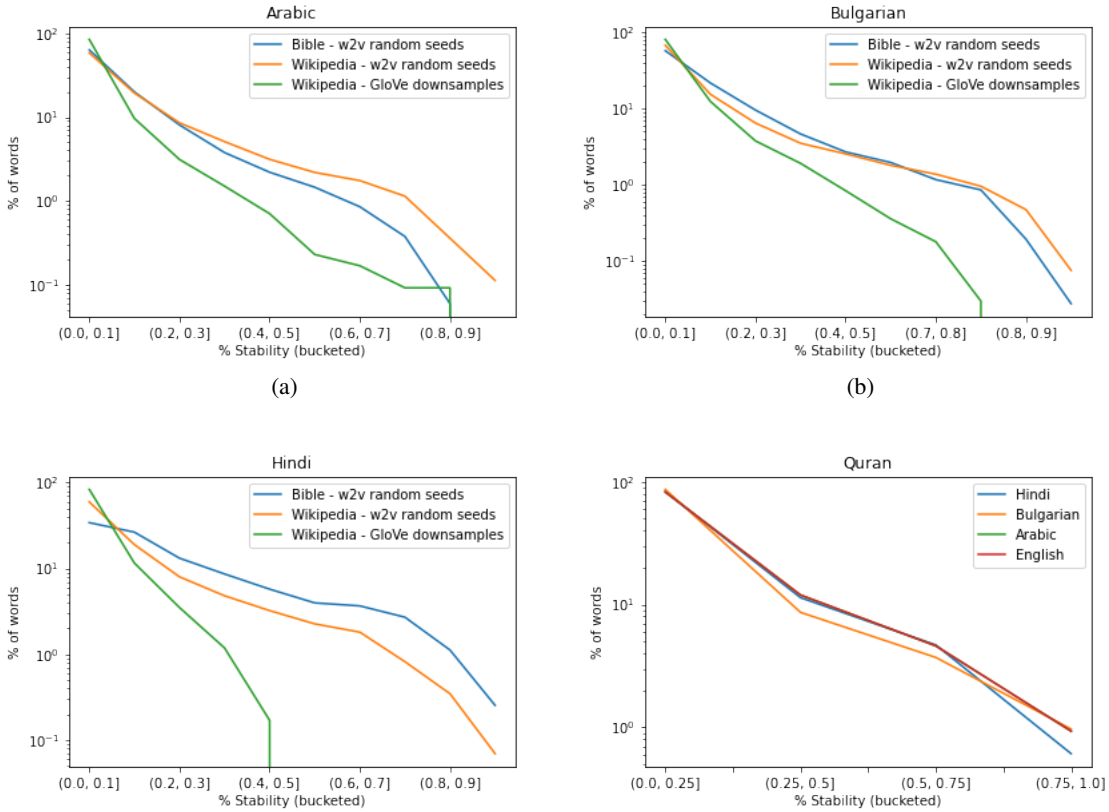
3.6 Computational Requirements

For the four Wikipedia text datasets, it took approximately 140 CPU hours to download text, write down sampling code, experiment with sample size, and get the shell script working. The stability and regression models for Wikipedia using glove embedding and Wikipedia using word2vec with ransom seed took a total of 75 CPU hours to calculate. We attempted to run the Wikipedia dataset using word2vec for all downsamples, but the code failed to create the embedding, so we only proceeded with the other two due to time constraints.

During this analysis we discovered that, even after reducing the vocabulary length, the English language took the longest when compared to other languages in the Wikipedia textdata.

The Bible corpus, on the other hand, contains all 97 languages, and due to the small dataset, it only took 100 CPU hours to complete the stability calculations for all languages and run the regression model.

For Quran, we manually created the text dataset and pre-processing for four languages: English, Hindi, Bulgarian, and Arabic. The stability calculation took a total of 85 CPU hours.



(a) (b) (c) (d)

Figure 1: Bucketed stability for a subset of the languages for Wikipedia and Glove, and stability values for Quran

4 Results

We experiment with two primary aspects of the paper. One methodology is to compare the stability values of various embeddings, for example GloVe versus Word2Vec. Both embedding techniques make use of standard parameters. The 10 closest neighbors of each word are calculated for each embedding. Additionally, we determine the stability of each language's words across all five embedding spaces. Furthermore, the paper's next step is to use regression modeling to determine which linguistic factors have an effect on the model's overall stability.

4.1 Stability in Multiple Languages

Both Wikipedia and the Bible exhibit a high degree of Bucket stability, as illustrated in Figure 1. As a general rule, we observe a large number of somewhat unstable word embeddings followed by a relatively flat distribution of 45 to 75 percent before dropping sharply at high levels of stability. Certain languages have much more stable embeddings than others. In Wikipedia's GloVe downsamples, the most stable embeddings were discovered for Hindi language word2vec with random seed, while the least stable embeddings were discovered for hindi with Wikipedia corpus with Glove downsamples in Figure1 (c). Overall, the trends for the other languages Arabic and Bulgaria looked very similar.

As a result of the findings we see that, any study of a language that makes use of embeddings should train on a large number of embedding spaces in order to generate robust results.

Additionally, altering the data set has a smaller effect on performance than altering the training algorithm. When the data is kept constant but the algorithm is changed, the same patterns emerge. We typically observe large variations in the results when the technique (w2v random seeds) is fixed but the data set is varied. In other words, in order to make meaningful comparisons across languages, the corpus's content must be properly accounted for. Even if the Bible is too small to support down sampling, the Wikipedia experiment results demonstrate that random seed experiments have the same effect as data sample experiments.

Cat	WALS Attribute	Weight
VC	The Perfect: **No Perfect tense**	-0.0493 ± 0.0
M	Locus of Marking in the Clause: **Dependent marking**	-0.0448 ± 0.0
L	Hand and Arm: **Different**	-0.0404 ± 0.0
SC	Alignment of Case Marking of Full Noun Phrases: **Nominative - accusative**	-0.0396 ± 0.0
NC	Position of Pronominal Possessive Affixes: **Possessive suffixes**	-0.0394 ± 0.0
VC	Position of Tense-Aspect Affixes: **Tense-aspect suffixes**	-0.0391 ± 0.0
M	Prefixing vs. Suffixing in Inflectional Morphology: **Strongly suffixing**	-0.0369 ± 0.0
SC	Alignment of Case Marking of Pronouns: **Nominative - accusative**	-0.0353 ± 0.0
M	Inflectional Synthesis of the Verb: **4-5 categories per word**	-0.0353 ± 0.0
VC	The Past Tense: **Present, no remoteness distinctions**	-0.0353 ± 0.0

Table 1: Features with the top 10 most negative weights for Ridge Regression

Cat	WALS Attribute	Weight
NC	The Associative Plural: **No associative plural**	0.0451 ± 0.0
VC	Position of Tense-Aspect Affixes: **No tense-aspect inflection**	0.0435 ± 0.0
M	Prefixing vs. Suffixing in Inflectional Morphology: **Little affixation**	0.0435 ± 0.0
VC	Position of Polar Question Particles: **Final**	0.0426 ± 0.0
NC	Position of Pronominal Possessive Affixes: **No possessive affixes**	0.0399 ± 0.0
SC	Alignment of Case Marking of Full Noun Phrases: **Neutral**	0.0394 ± 0.0
NC	Inclusive/Exclusive Distinction in Verbal Inflection: **No person marking**	0.0361 ± 0.0
SC	Predicative Adjectives: **Verbal encoding**	0.0355 ± 0.0
VC	The Past Tense: **No past tense**	0.0351 ± 0.0
M	Locus of Marking in the Clause: **No marking**	0.0351 ± 0.0

Table 2: Features with the top 10 most positive weights for Ridge Regression

4.2 Regression Modelling

We investigated the linguistic factors that associated with stability using regression modeling. We employed ridge regression, which was originally used in the research paper to determine the average stability of all words in a language given language features. Ridge regression normalizes the model weights in magnitude, resulting in a more interpretable model than non-regularized linear regression. We examined several regularization strengths and conducted cross validation on the following values 0.0001,0.001,0.01,0.1,1,10,100, and 1000. We discovered that the optimal alpha value for the ridge regression model is 10. The R2 value for the ridge regression model is 0.967, indicating that the model matches the data well. The regression model's largest weights are listed in Table 1. demonstrates that three of the top negative features are 'No perfect tense' and that the majority of features are linked to suffixes, whereas the top positive feature in Table 2. is 'No associative plural' and the other feature patterns are quite distinct.

4.3 Additional Results not Present in the Original Paper

The following are two additions that were not included in the original research paper that we experimented:

1. Additional dataset was utilized to determine the stability: We calculated the stability of four languages using the Quran dataset: Hindi, English, Bulgarian, and Arabic. Figure 1. demonstrates the bucket's stability across the four languages. As illustrated in Figure 1 (d), the four languages of the Quran show significant stability. As can be seen, the graph is slightly different from the wikipedia and bible corpus graphs. This is because we used a small dataset and trained the model with word2vec embeddings.

2. Experimented with various types of regression models.: We found the average stability of all languages given their attributes using the Lasso, Elastic Net, and Linear regression models. The best alpha with the least error for Lasso is 0.01 and the R2 score is 0.98816. Elastic net's best alpha value is 0.1 and its R2 score is 0.801. The R2 value for linear regression is 0.638 at alpha 0.001.

5 Discussion

We calculated the stability of the Wikipedia corpus across four languages using glove embedding and word2vec, as well as the bible corpus across 97 languages using word2vec. Additionally, we examined the stability of the Quran's four languages. We discovered that the stability trend for glove and word2vec was very similar across the Wikipedia and bible corpora. However, we discovered that the Hindi language had the least stability for wiki text with glove embedding and the most stability for the same language with a word2vec random seed. In the case of the Quran, the stability trends for all four languages are fairly similar. This could be because we used the exact same data size as the Quran's 25 Ayah.

In the case of regression, we first used the ridge regression model, which produced an excellent R2 score of 0.96 at alpha 10; this indicates that our results are identical to those of the authors. Additionally, we tested three additional regression models, Lasso, Elastic, and linear, and found that Lasso regression achieved the highest R2 score of 0.988, outperforming all other regression models.

5.1 What was Easy

The author has provided a code that is simple to understand and interpret for some of the sections only.

5.2 What was Difficult

Some of the Python notebooks contained bugs, such as an incorrect loop function or an incorrect file path assignment, which required considerable time to resolve. The pre-processing time for the Wikipedia data was huge we were able to test four languages English, Hindi, Arabic and Bulgarian.

5.3 Recommendations for Reproducibility

Our recommendation to the author is that they include explicit instructions on how to use the Wikipedia corpus with word2vec using all five downsampling. This point is only stated briefly in the paper and does not appear at all in the code.

Communication with Original Authors

While gathering information on the dataset for the paper, we discovered that the wiki text was accessible via Polyglot Al-Rfou et al. [2013] in the form of zip files for 40 different languages. We downloaded the first file from the list when writing the project proposal, assuming that the remaining files would be available as well. However, this was not the case; we were required to contact the author to obtain access to additional language files. We had to wait a while for the author to respond, but we eventually got permission to use two language files which we used in project report V1. Furthermore, the author informed us that we could use a newly released cleaned data set for 40 language wiki text on Pytorch. However, the tensor flow dataset used differs from the author's data set and necessitates pre-processing for each language, which took a long time. As a result, we asked the author once more for access to the raw data and were granted permission for two more languages. We emailed the author again, requesting access to the other language dataset as well, but did not receive a response.

As for the Bible corpus, as described earlier, we used a version of the Bible corpus available online Christodoulopoulos and Steedman [2015] which gave reasonable results. However, in the paper, a slightly different version of the dataset was used, so we will attempt to to obtain this alternate version of the dataset from the authors.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-3520`.

Laura Burdick. Multilingual stability. `https://github.com/laura-burdick/multilingual-stability`, 2021.

Laura Burdick, Jonathan K. Kummerfeld, and Rada Mihalcea. Analyzing the surprising variability in word embedding stability across languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL `https://arxiv.org/abs/2004.14876`.

Christos Christodoulopoulos and Mark Steedman. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49:375 – 395, 2015.

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/`.

Muhamad Fahmi Fakhrezi, Moch Arif Bijaksana, and Arief Fatchul Huda. Implementation of automatic text summarization with textrank method in the development of al-qur'an vocabulary encyclopedia. *Procedia Computer Science*, 179:391–398, 2021.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.352`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Prash Rangarajan. Multilingualstability. `https://github.com/PrashRangarajan/MultiLingualStability.git`, 2022.