Analyzing the Surprising Variability in Word Embedding Stability Across Languages

Laura Burdick, Jonathan K. Kummerfeld and **Rada Mihalcea**
Computer Science & Engineering
University of Michigan, Ann Arbor
{lburdick,jkummerf,mihalcea}@umich.edu

Abstract

Word embeddings are powerful representations that form the foundation of many natural language processing architectures, both in English and in other languages. To gain further insight into word embeddings, we explore their stability (e.g., overlap between the nearest neighbors of a word in different embedding spaces) in diverse languages. We discuss linguistic properties that are related to stability, drawing out insights about correlations with affixing, language gender systems, and other features. This has implications for embedding use, particularly in research that uses them to study language trends.

1 Introduction

Word embeddings have become an established part of natural language processing (NLP) (Collobert et al., 2011; Wang et al., 2020a). Stability, defined as the overlap between the nearest neighbors of a word in different embedding spaces, was introduced to measure variations in local embedding neighborhoods across changes in data, algorithms, and word properties (Antoniak and Mimno, 2018; Wendlandt et al., 2018). These studies found that many common English embedding spaces are surprisingly unstable, which has implications for work that uses embeddings as features in downstream tasks, and work that uses embeddings to study specific properties of language.

However, research to date on word embedding stability has been exclusively done on English and so is not representative of all languages. In this work, we explore the stability of word embeddings in a wide range of languages. Better understanding the differences caused by diverse languages will provide a foundation for building embeddings and NLP tools in all languages.[1]

[1] Code is available at https://lit.eecs.umich.edu/downloads.html.

In English and other very high resource languages, it has become common practice to use contextualized word embeddings, such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019). These algorithms require huge amounts of computational resources and data. For example, it takes 2.5 days to train XLNet with 512 TPU v3 chips. In addition to requiring heavy computational resources, most contextualized embedding algorithms need large amounts of data. BERT uses 3.3 billion words of training data. In contrast to these large corpora, many datasets from low-resource languages are fairly small (Maxwell and Hughes, 2006). To support scenarios where using huge amounts of data and computational resources is not feasible, it is important to continue developing our understanding of context-independent word embeddings, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). These algorithms continue to be used in a wide variety of situations, including the computational humanities (Abdulrahim, 2019; Hellrich et al., 2019) and languages where only small corpora are available (Joshi et al., 2019).

In this work, we consider how stability varies for different languages, and how linguistic properties are related to stability—a previously understudied relationship. Using regression modeling, we capture relationships between linguistic properties and average stability of a language, and we draw out insights about how linguistic features relate to stability. For instance, we find that embeddings in languages with more affixing tend to be less stable. Our findings provide crucial context for research that uses word embeddings to study language properties and trends (e.g., Heyman and Heyman, 2019; Abdulrahim, 2019), which often rely on raw embeddings created by GloVe or word2vec. If these embeddings are unstable, then research using them needs to take this into account in terms of methodologies and error analysis.

2 Related Work

Word embeddings are low-dimensional vectors used to represent words, normally in downstream tasks, such as word sense disambiguation (Scarlini et al., 2020) and text summarization (Moradi et al., 2020). They have been shown to capture both syntactic and semantic properties of words, making them useful in a wide range of NLP tasks (Wang et al., 2020b). In this work, we explore word embeddings that generate one embedding per word, regardless of the word's context. We consider two widely used algorithms: word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014).

Our work analyzes embeddings in multiple languages, which is important because embeddings are commonly used across many languages. In particular, there has been interest in embeddings for low-resource languages (Chimalamarri et al., 2020; Stringham and Izbicki, 2020).

In this work, we use stability to measure the quality of word embeddings. Similar to the work we present here on stability, other research looks at how nearest neighbors vary as properties of the embedding spaces change. Pierrejean and Tanguy (2018) found that the lowest frequency and the highest frequency words have the highest variation among nearest neighbors. Additional research has explored how semantic and syntactic properties of words change with different embedding algorithm and parameter choices (Artetxe et al., 2018; Yaghoobzadeh and Schütze, 2016). Unlike our work, previous studies only considered English.

Finally, while our work is *not* a form of embedding evaluation, it is related to the topic (Chiu et al., 2016; Rogers et al., 2018; Qiu et al., 2018). There has been extensive work on evaluating word embeddings, seen in the recent RepEval workshops (Rogers et al., 2019), and going back to work comparing them with counting based methods (Baroni et al., 2014). Our findings indicate that work on embedding evaluation should take into consideration stability, using multiple training runs to confirm results. Similarly, stability should be considered when studying the impact of embeddings on downstream tasks. Leszczynski et al. (2020) specifically looked at the downstream instability of word embeddings, and found that there is a stability-memory tradeoff, and higher stability can be achieved by increasing the embedding dimension.

3 Data

In order to explore the stability of word embeddings in different languages, we work with two datasets, Wikipedia and the Bible. While Wikipedia has more data, the Bible covers more languages. Wikipedia is a comparable corpus, whereas the Bible is a parallel corpus.

Wikipedia Corpus. We use pre-processed Wikipedia dumps in 40 languages taken from Al-Rfou' et al. (2013).[2] The size of these Wikipedia corpora varies from 329,136 sentences (Tagalog) to 75,241,648 sentences (English), with an average of 9,292,394 sentences. For all of our experiments, we downsample each corpus to work with comparably sized data (details in Section 4.2).

Bible Corpus. We consider 97 languages from the pre-processed Bible corpus (McCarthy et al., 2020):[3] all languages for which at least 75% of the Bible ($\geq 23,326$ verses) is present.[4] This excludes many languages for which there is only a partial Bible, e.g., just the New Testament, which would be insufficient for training word vectors. We consider two sets of languages with the Bible corpus: languages that overlap with the set of Wikipedia languages (26 languages), and all languages in the Bible corpus (97 languages).

WALS. To gain linguistic properties of these languages, we use the World Atlas of Language Structures (WALS),[5] a database of phonological, lexical, and grammatical properties for over 2,000 languages (Dryer and Haspelmath, 2013). This expert-curated resource contains 192 language features. For example, WALS records subject, object, and verb word order for various languages.

4 Calculating Stability in Many Languages

The first part of our work is a comparison of stability across languages. Before presenting our measurements, we define stability and analyze some important methodological decisions.

[2] Available online at https://sites.google.com/site/rmyeid/projects/polyglot.

[3] Available by contacting McCarthy et al. (2020).

[4] To work with a maximum number of languages, we only consider the complete Protestant Bible (i.e., all of the verses that appear in the English King James Version of the Bible).

[5] Available online at https://wals.info.

Model 1: **indie**, **punk**, **progressive**, **pop**, *roll*, **band**, blues, brass, class, **alternative**
Model 2: **punk**, **indie**, **alternative**, **progressive**, **band**, sedimentary, bands, psychedelic, *climbing*, **pop**
Model 3: **punk**, **pop**, **indie**, **alternative**, **band**, *roll*, **progressive**, folk, *climbing*, metal

Table 1: Ten nearest neighbors for the word **rock** in three GloVe models trained on different subsets of Large English Wikipedia. Words in all lists are in bold; words in only two lists are italicized. Models 1 and 2 have 6 words (60%) in common, models 1 and 3 have 7, and models 2 and 3 have 7. Therefore, this word has a stability of 66.7%, the average word overlap between the three models.

4.1 Defining Stability

Stability is defined as the percent overlap between nearest neighbors in an embedding space. To calculate stability, given a word W and two embedding spaces A and B, take the ten nearest neighbors (measured using cosine similarity) of W in both A and B. The stability of W is the percent overlap between these two lists of nearest neighbors.[6] 100% stability indicates perfect agreement between the two embedding spaces, while 0% stability indicates complete disagreement. Table 1 shows a simple example. This definition of stability can be generalized to more than two embedding spaces by considering the average overlap between pairs of embedding spaces. Let X and Y be two sets of embedding spaces. Then, for every pair of embedding spaces (x, y), where $x \in X$ and $y \in Y$, take the ten nearest neighbors of W in both x and y and calculate percent overlap. Let the stability be the average percent overlap over every pair of embedding spaces (x, y).

Previous work has explored stability for English word embeddings. For instance, it was found that the presence of certain documents in the training corpus affects stability (Antoniak and Mimno, 2018), and that training and evaluating embeddings on separate domains is less stable than training and evaluating on the same domain (Wendlandt et al., 2018). In this work, we expand this analysis to a more diverse set of languages.

[6]While alternative definitions of stability are possible, e.g., considering a vector of similarities with a large set of words, we chose to use a prior definition of stability that has been rigorously studied. Similarly, sets of nearest neighbors smaller and larger than ten have been tried previously, with comparable results (Wendlandt et al., 2018).

4.1.1 The Effect of Downsampling on Stability

Stability measures how changes to the input data or training algorithm affect the resulting embeddings. Sometimes we make changes with the goal of shifting the embeddings, such as increasing the context window size to try to get embeddings that capture semantics more than syntax. In other cases, we would hope a change would not substantially change embeddings, such as changing the random seed for the algorithm. For our experiments, we consider a previously unstudied source of instability: different data samples from the same distribution. This is a case where we hope embeddings remain stable, given a sufficiently large sample.

We generate data samples by downsampling a corpus to create multiple smaller corpora; we then measure stability across these downsamples. The choice of sampling with or without replacement, and the size of the sample are subtle methodological choices. In this section, we consider whether stability across downsamples produces consistent results that we can compare across languages.

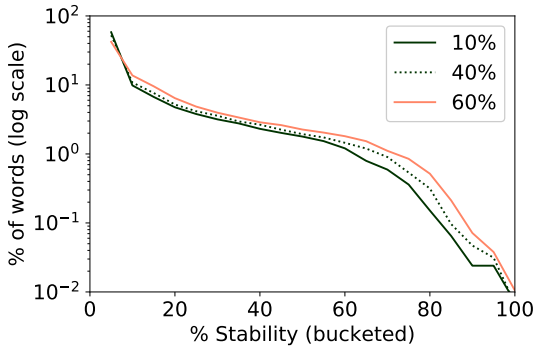
First, we consider downsampling with replacement, shown in Figure 1a. We use data drawn from an English Wikipedia corpus of 5,269,686 sentences (denoted "Large English Wikipedia").[7] We randomly sample five sets of 500,000 sentences multiple times, controlling the amount of overlap between downsamples (from 10% to 60% shared across all five samples). For a specific overlap amount X%, X% of 500,000 sentences is randomly sampled and included in all of the five downsamples. The remaining (100-X)% sentences are randomly sampled for each downsample.

Stability is calculated using GloVe embeddings and the words that occur in every downsample for every overlap percentage. In Figure 1a, we group stability into buckets of size 5% (i.e., 0-5%, 5-10%, etc). This allows us to see patterns in stability that are not visible from a single statistic, such as the overall average. We see that while stability trends are similar for different overlap amounts, stability is consistently higher as the overlap amount increases. This means that if we use downsampling with replacement, we cannot reliably compare stability across multiple corpora of varying sizes (e.g., Wikipedia and the much smaller Bible corpus). The overlap amount would change de-

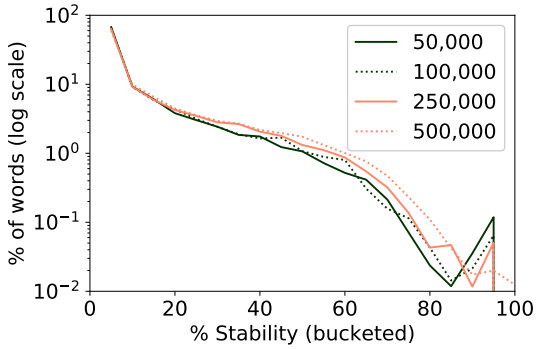
[7]This data was used in Tsvetkov et al. (2016) and is available by contacting the authors of that paper.

Experiment	Machine	Timing
Training one w2v embedding on one Wikipedia corpus (Section 4)	Machine 1	13 sec.
Training one GloVe embedding on one Wikipedia corpus (Section 4)	Machine 1	12 min.
Calculating stability on one Wikipedia corpus (Section 4)	Machine 1	17 sec.
Training one w2v embedding on one Bible corpus (Section 4)	Machine 1	5 sec.
Calculating stability on one Bible corpus (Section 4)	Machine 1	12 sec.
Training regression model (Section 5)	Machine 2	< 7 sec.
Leave-one-out cross-validation (Section 6)	Machine 2	< 4 sec.

Table 2: Runtimes for different experimental portions of this work. Machine 1 is four Intel(R) Xeon(R) CPU E5-1603 v3 @ 2.80 GHz processors. Machine 2 is a 2.9GHz Dual-Core Intel Core i5.



(a) Sampling *with* replacement, varying percentage overlap between samples.



(b) Sampling *without* replacement, varying sample size.

Figure 1: Measuring the impact of data sampling parameters on stability measurements. Results when sampling *with* replacement consistently increase as overlap increases (a). This poses a problem, as results may reflect corpus size rather than intrinsic stability. Results when sampling *without* replacement do show a consistent pattern, even when the sample is only 50,000 sentences, a tenth of the largest sample size (b).

pending on the size of the corpus, changing our stability measurement.

Instead of downsampling with replacement, we consider downsampling without replacement, shown in Figure 1b for different downsample sizes. We see that varying the size of the downsample does not have a large effect on the patterns of stability. Particularly when looking at lower stability, the trends are remarkably consistent, even when the downsample size varies from 50,000 sentences to 500,000 sentences. The pattern grows less consistent when looking at higher stability, especially with smaller downsample sizes.

This comparison (Figures 1a and 1b) shows that downsampling without replacement produces more consistent (and thus comparable) stability results than downsampling with replacement. Thus, we only consider downsampling without replacement.

4.2 Stability for Wikipedia and the Bible

Our first study, shown in Figure 2, considers stability across the 26 languages included in both Wikipedia and the Bible. These results show three settings for Wikipedia: (1) Stability of GloVe embeddings across five downsampled corpora, (2) Stability of word2vec (w2v) embeddings across five downsampled corpora, and (3) Stability of word2vec embeddings using five random seeds on one downsampled corpus. For the Bible, we only show the third case, since it is too small for downsampling.

Each downsampled corpus is 100,000 sentences, and words that occur with a frequency less than five are ignored. Previous work (Pierrejean and Tanguy, 2018) has indicated that words that appear this infrequently will be very unstable. We use standard parameters for both embedding algorithms.[8] For each embedding, we calculate the ten nearest neighbors of every word using FAISS[9] (Johnson et al., 2019). Finally, for each language, we calculate the stability for every word in that language across all five embedding spaces. Experimental runtimes are listed in Table 2.

Figure 2 shows bucketed stability for both Wikipedia and the Bible. Most languages have the same overall trend: a large number of relatively unstable word embeddings, then a fairly flat distri-

[8]For GloVe (Pennington et al., 2014), we use 100 iterations, 300 dimensions, a window size of 5, and a minimum word count of 5; these parameters led to good performance in Wendlandt et al. (2018). For word2vec (Mikolov et al., 2013), we use 300 dimensions, a window size of 5, and a minimum word count of 5.

[9]We use exact, not approximate, search.

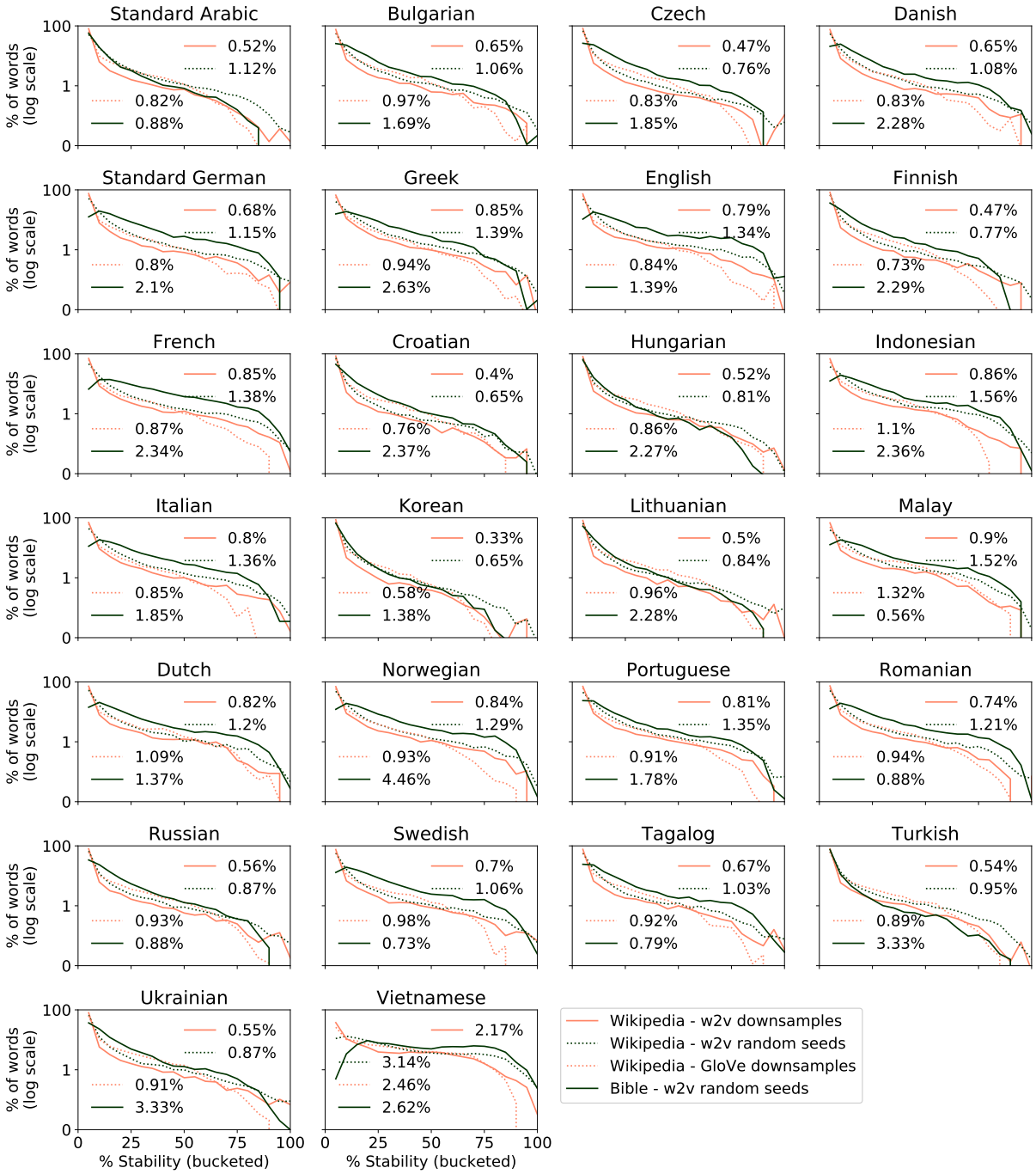
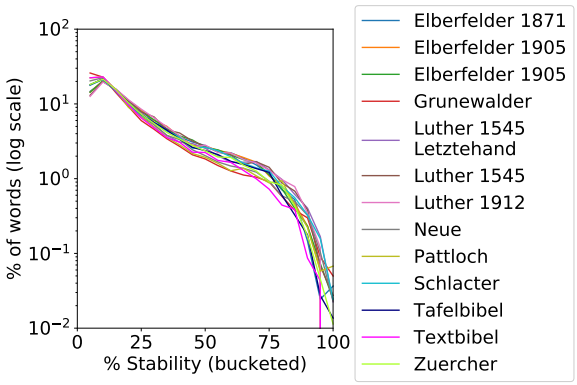


Figure 2: Percentage of words that occur in each stability bucket for four different methods, three on Wikipedia and one on the Bible. The 26 languages in common are shown here. The average stability for each method is shown on the individual graphs.

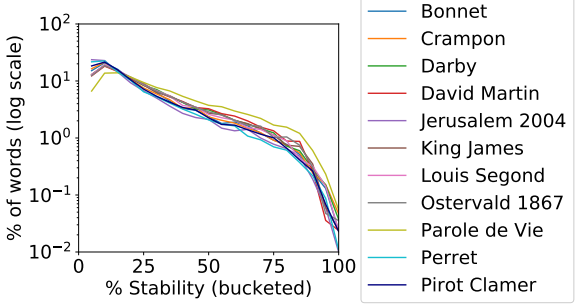
bution between 25% and 75%, and a sharp drop at high stability. This indicates that the conclusions from prior work on English apply to other languages as well. In particular, it means that any work that uses embeddings to study a language should train multiple embedding spaces to ensure robust findings.

Some languages have substantially more stable embeddings than others. Comparing GloVe downsamples on Wikipedia, Vietnamese has the most stable embeddings (avg. 2.46%), while Korean has the least stable embeddings (avg. 0.58%). The plot for Vietnamese has a different trend than many of the other plots in Figure 2. Vietnamese is the only Austro-Asiatic language in our dataset, so there could be multiple distinctives that are related to it exhibiting different patterns than the other languages.

Finally, varying the training algorithm has a smaller impact than changing the dataset. Keeping



(a) German



(b) French

Figure 3: Percentage of words that occur in each stability bucket for different Bible translations.

the dataset fixed (Wikipedia) and varying the algorithm, we see similar trends. Keeping the algorithm fixed (w2v random seeds) and varying the dataset, we often see substantial shifts. This means that in order to compare languages we need to carefully control for the content of the corpus (which the Bible data allows us to do). While the Bible is too small to support downsampling, these results on Wikipedia suggest that experiments varying the random seed lead to similar variations to experiments varying the data sample.

To confirm this finding, we consider two languages with multiple Bible translations: German and French. We average stability across five word2vec embeddings using five random seeds on one downsampled corpus. The downsampled corpus is 100,000 sentences, randomly sampled. Figure 3 shows the stability patterns for each. The results are very consistent, indicating that variations in translator behavior do not impact stability the way shifting from one corpus to another does. The largest shift is for the French Parole de Vie translation (top line in yellow in Figure 3b), which intentionally uses simpler, everyday language. For further experiments on languages with multiple Bible translation, we choose the Bible translation with the highest average stability.

It is difficult to infer more from these figures alone. In the next section, we use regression modeling to identify patterns in the results. Based on the observations above, we use results from GloVe across five downsampled corpora for Wikipedia, and results across five random seeds for the Bible.

5 Regression Modeling

We now explore linguistic factors that correlate with stability. To draw conclusions about specific linguistic features, we use a ridge regression model (Hoerl and Kennard, 1970)[10] to predict the average stability of all words in a language given features reflecting language properties. Regression models have previously been used to measure the impact of individual features (Singh et al., 2016). Ridge regression regularizes the magnitude of the model weights, producing a more interpretable model than non-regularized linear regression. We experiment with different regularization strengths and use the best-performing value ($\alpha = 10$).[11] We choose to use a linear model here because of its interpretability. While more complicated models might yield additional insight, we show that there are interesting connections to be drawn from a linear model.

5.1 Model Input and Output

Our model takes linguistic features of a language as input and predicts stability as output. Since WALS properties are categorical, we turn each property into a set of binary features. If a particular language does not have a known value for a given property, then all of these features are marked zero.

In order to draw out important correlations between linguistic features and stability, we filter the languages and WALS properties that we consider. We only include languages that have at least 25% of all WALS properties. Then, we only consider WALS properties that cover at least 25% of the filtered languages. We remove all WALS properties that do not have at least two features that each include at least five languages. Note that because all of our input features are binary, all weights are easily comparable. After this filtering, we end up

[10] Run using the Python package `sklearn.linear.model.Ridge` (Pedregosa et al., 2011) with default parameters except $\alpha = 10$.

[11] We run leave-one-language-out cross-validation, described in Section 5.2, using the α values of 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, and 1000, choosing the α value with the lowest average absolute error.

with 37 languages,[12] and 97 WALS properties.

We also group highly correlated WALS features. We create the groupings by combining features with a Pearson correlation greater than 0.8. A feature is included in a particular grouping if it correlates highly with any of the features already in the group. Each grouped feature is marked as one if *any* of the included features are marked as one.

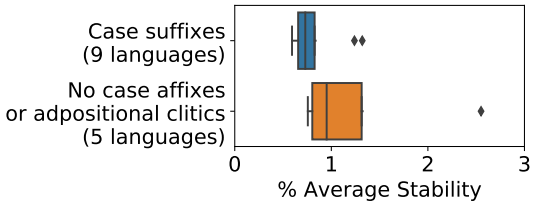
For each model, we bootstrap over the input features $1,000$ times, allowing us to calculate standard error for the R^2 score and the model weights. Calculating significance for each feature allows us to discard highly variable weights and focus on features that consistently contribute to the regression model, giving us more confidence in the results.

The output of our model is the average stability of a language, which is calculated by averaging together the stability of all of the words in a language. If a language is present in both corpora, we average the stabilities from the two corpora.

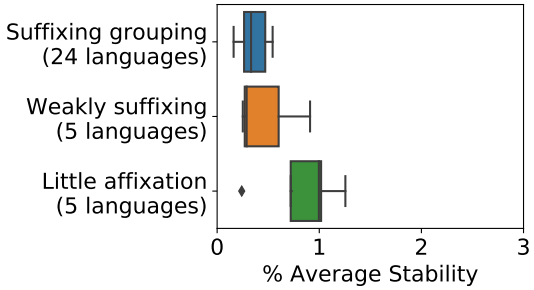
5.2 Evaluation

We evaluate our model in two ways. First, we measure goodness of fit using the coefficient of determination R^2.[13] This measures how much variance in the dependent variable y (average stability) is captured by the independent variables x (WALS properties). A model that always predicts the expected value of y, regardless of the input features, will have an R^2 score of 0. The highest possible R^2 score is 1, and R^2 can be negative. Second, in addition to the R^2 score, we run leave-one-out cross-validation across all languages, and report absolute error on the left-out language. We compare this to a baseline of choosing the average stability over all training languages.

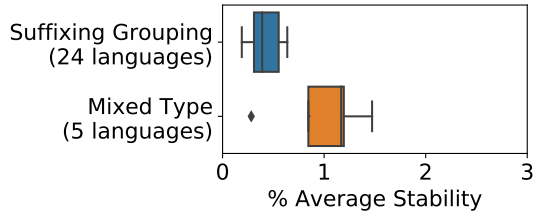
We use the individual feature weights to measure how much a particular feature contributes to the overall model. When reporting weights, we train the model using all 37 languages. Because we are primarily using regression modeling to learn associations between certain features and stability, no test data are necessary. The emphasis is on the model itself and the feature weights it learns, not on the model's performance on a task.



(a) Position of Case Affixes



(b) Prefixing vs. Suffixing in Inflectional Morphology



(c) Position of Tense-Aspect Affixes

Figure 4: Affixing properties compared using box-and-whisker plots.

6 Results and Discussion

Our regression model has a high R^2 score of 0.96 ± 0.00, indicating that the model fits the data well. Significant weights with the highest magnitude are shown in Table 3. Running leave-one-out cross-validation across all languages, we get an average absolute error of 0.62 ± 0.53.[14] For comparison, using the average stability gives an average absolute error of 0.86 ± 0.55. (A two sample t-test comparison gives a p-value of 0.060.)

Table 4 breaks down the regression results by broad WALS category, listing both the number of binary features per category, as well as the average magnitude of weights for features in that category. The two most important groups of features are Nominal Categories and Verbal Categories. Both of these categories have a large number of features and a high average magnitude. While the Lexicon category has a high average magnitude, it contains

[12] Bengali, Bulgarian, Cherokee, Comanche, English, Estonian, Finnish, Haitian, Haitian Creole, Hebrew, Hindi, Hmong Njua, Hungarian, Indonesian, Italian, Japanese, Korean, Latin, Latvian, Linda, Lithuanian, Ma'di, Mam, Mandarin, Maybrat, Norwegian, Persian, Pohnpeian, Polish, Portuguese, Russian, Somali, Spanish, Swedish, Thai, Turkish, Ukrainian, Vietnamese

[13] Measured using the Python package `sklearn.linear_model.Ridge.score`.

[14] Cross-validation has an average R^2 score of 0.92 on the training data.

Cat.	WALS Attribute	Weight
VC, M	*Suffixing Grouping:* ·Prefixing vs. Suffixing in Inflectional Morphology: **Strongly Suffixing**; ·Position of Tense-Aspect Affixes: **Tense-aspect suffixes**	-0.14 ± 0.0
L	Hand and Arm: **Different**	-0.11 ± 0.0
CS	Relativization on Obliques: **Gap**	-0.10 ± 0.0
VC	Overlap between Situational & Epistemic Modal Marking: **Overlap for both possibility & necessity**	-0.09 ± 0.0
NC	Ordinal Numerals: **First, second, three-th**	-0.08 ± 0.0
NC	Comitatives and Instrumentals: **Differentiation**	-0.08 ± 0.0
P	Rhythm Types: **Trochaic**	-0.08 ± 0.0
WO	Order of Adjective and Noun: **Adjective-Noun**	-0.07 ± 0.0
WO	Order of Adposition and Noun Phrase: **Postpositions**	-0.07 ± 0.0
NC	*No Gender Grouping:* · Systems of Gender Assignment: **No gender**; · Sex-based and Non-sex-based Gender Systems: **No gender**; · Gender Distinctions in Independent Personal Pronouns: **No gender distinctions**; · Number of Genders: **None**	0.05 ± 0.0
P	Voicing and Gaps in Plosive Systems: **Other**	0.06 ± 0.0
M	Prefixing vs. Suffixing in Inflectional Morphology: **Little affixation**	0.06 ± 0.0
CS	'Want' Complement Subjects: **Subject is expressed overtly**	0.06 ± 0.0
VC	The Morphological Imperative: **No second-person imperatives**	0.06 ± 0.0
CS	Purpose Clauses: **Balanced**	0.06 ± 0.0
WO	*Prepositions Grouping:* ·Order of Adposition and Noun Phrase: **Prepositions**; ·Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase: **VO and Prepositions**	0.06 ± 0.0
WO	Order of Demonstrative and Noun: **Noun-Demonstrative**	0.07 ± 0.0
NC	Position of Case Affixes: **No case affixes or adpositional clitics**	0.11 ± 0.0

Table 3: Weights with the highest magnitude in the regression model. Negative weights correspond with low stability, and positive weights correspond with high stability.

WALS Category	Num. Features	Avg. Magnitude
Simple Clauses (SC)	30	0.019
Nominal Syntax (NS)	2	0.021
Other (O)	2	0.023
Complex Sentences (CS)	11	0.028
Morphology (M)	18	0.031
Word Order (WO)	32	0.031
Phonology (P)	21	0.032
Nominal Categories (NC)	40	0.036
Verbal Categories (VC)	27	0.036
Lexicon (L)	6	0.039

Table 4: Number of binary features and average magnitude of weights in the regression model for different WALS categories. Grouped features are included in each category that they cover.

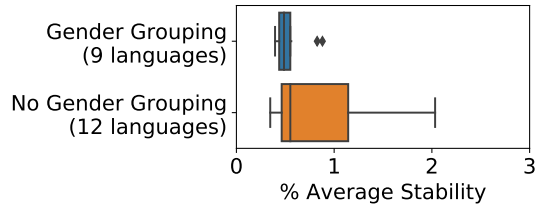
very few features. To further explore these results, we highlight a few WALS property in more detail.

Suffixes and prefixes. Table 3 shows that three of the top features are related to affixes (suffixes and prefixes). Specifically, three main properties deal with affixes: Position of Case Affixes (Dryer, 2013a), Prefixing vs. Suffixing in Inflectional Morphology (Dryer, 2013c), and Position of Tense-Aspect Affixes (Dryer, 2013b). Distributions of these features in the 37 languages used for the regression model are shown in Figure 4 (categories with fewer than five languages are not shown).

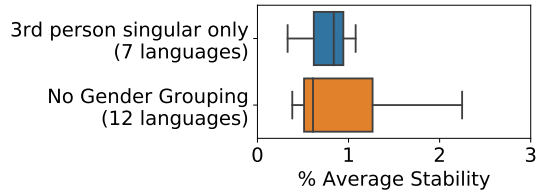
For all three of these properties, more affixing is associated with lower stability. When considering word embeddings, this result makes intuitive sense. Affixes cause there to be many different word variations (e.g., *walk*, *walked*, *walking*, *walker*), which may not be handled consistently by the embedding algorithm, leading to lower average stability.

Gendered Languages. Table 3 also highlights a grouping of WALS properties related to whether a language is gendered or not. Four WALS properties are relevant to this: Systems of Gender Assignment (Corbett, 2013c), Sex-based and Non-sex-based Gender Systems (Corbett, 2013b), Gender Distinctions in Independent Personal Pronouns (Siewierska, 2013), and Number of Genders (Corbett, 2013a). In general, a language is considered to have a gender system if different parts-of-speech are required to agree in gender (as opposed to simply having gendered nouns). Distributions of these features are shown in Figure 5.

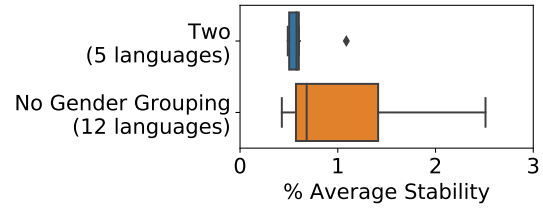
For all of these properties, languages with no gender system tend to have higher average stability. Again, this result makes sense in the context of



(a) Systems of Gender Assignment; Sex-based and Non-sex-based Gender Systems (*Gender Grouping:* No gender; Sex-based)



(b) Gender Distinctions in Independent Personal Pronouns



(c) Number of Genders

Figure 5: Gender properties compared using box-and-whisker plots. Note, the 12 languages with "No Gender Grouping" are not the same across the three plots.

word embeddings. Languages with gender systems will have more word forms (e.g., both male and female word forms), which may not be handled consistently by the embedding algorithm.

7 Conclusion

In this paper, we considered how stability varies across different languages. This work is important because algorithms such as GloVe and word2vec continue to be effective methods in a wide variety of scenarios (Arora et al., 2020), particularly the computational humanities and languages where large corpora are not available. We studied the relationship between linguistic properties and stability, something that has been previously understudied. We drew out several aspects of this relationship, including that languages with more affixing tend to have less stable embeddings, and languages with no gender systems tend to have more stable embeddings. These insights can be used in future work to inform the design of embeddings in many languages. For example, this work suggests that future embedding space designs need to take into account gendered words and morphologically rich words with affixes.

8 Acknowledgements

This material is based in part upon work supported by the National Science Foundation (grant #1815291) and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or John Templeton Foundation.

References

Abdul Z Abdulrahim. 2019. Ideological drifts in the US constitution: Detecting areas of contention with models of semantic change. In *NeurIPS Joint Workshop on AI for Social Good*, Vancouver, Canada.

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. 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.

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 282–291, Brussels, Belgium. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Santwana Chimalamarri, Dinkar Sitaram, and Ashritha Jain. 2020. Morphological segmentation to improve crosslingual word embeddings for low resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–15.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Greville G. Corbett. 2013a. Number of genders. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Greville G. Corbett. 2013b. Sex-based and non-sex-based gender systems. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Greville G. Corbett. 2013c. Systems of gender assignment. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer. 2013a. Position of case affixes. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013b. Position of tense-aspect affixes. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013c. Prefixing vs. suffixing in inflectional morphology. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2019. Modeling word emotion in historical language: Quantity beats supposed stability in seed word selection. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11, Minneapolis, USA. Association for Computational Linguistics.

Tom Heyman and Geert Heyman. 2019. Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology*, pages 2084–2109.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

Ishani Joshi, Purvi Koringa, and Suman Mitra. 2019. Word embeddings in low resource Gujarati language. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 5, pages 110–115. IEEE.

Megan Leszczynski, Avner May, Jian Zhang, Sen Wu, Christopher R. Aberger, and Christopher Ré. 2020. Understanding the downstream instability of word embeddings. In *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org.

Mike Maxwell and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 29–37, Sydney, Australia. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. 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. European Language Resources Association.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada.

Milad Moradi, Maedeh Dashti, and Matthias Samwald. 2020. Summarization of biomedical articles using domain-specific word embeddings and graph ranking. *Journal of Biomedical Informatics*, 107:103452.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher,

Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bénédicte Pierrejean and Ludovic Tanguy. 2018. Towards qualitative word embeddings evaluation: Measuring neighbors variation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 209–221. Springer.

Anna Rogers, Aleksandr Drozd, Anna Rumshisky, and Yoav Goldberg, editors. 2019. *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Minneapolis, USA.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Anna Siewierska. 2013. Gender distinctions in independent personal pronouns. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Nathan Stringham and Mike Izbicki. 2020. Evaluating word embeddings on low-resource languages. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 176–186, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 130–139, Berlin, Germany. Association for Computational Linguistics.

Shirui Wang, Wenan Zhou, and Chao Jiang. 2020a. A survey of word embeddings based on deep learning. *Computing*, 102(3):717–740.

Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. 2020b. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, pages 1–20.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Intrinsic subspace evaluation of word embedding representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 236–246, Berlin, Germany. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Thirty-third Conference on Neural Information Processing Systems*, pages 5754–5764, Vancouver, Canada.