Analyzing the Surprising Variability in Word Embedding Stability Across Languages

Laura Burdick, Jonathan K. Kummerfeld, Rada Mihalcea

University of Michigan *{lburdick, jkummerf, mihalcea}@umich.edu*

Introduction

Does stability vary for different languages? Is stability associated with linguistic properties?

- **Data**
  - **Wikipedia** (40 languages)
  - **Bible** (97 languages)
  - **World Atlas of Language Structures (WALS)**, phonological, lexical, and grammatical properties (>2,000 languages)
- Why word2vec and GloVe?
  - These algorithms continue to be used in many situations, including the computational humanities and low-resource languages!

What is Stability?

Stability = *percent overlap between ten nearest neighbors in an embedding space*

$$\text{stability} = \frac{100}{|\text{words}|} \sum_{\text{words}} \frac{\text{neighbors}_0 \cap \text{neighbors}_1}{10}$$

- **neighbors$_0$** = ten words most similar to the word in embedding space 0
- **neighbors$_1$** = ten words most similar to the word in embedding space 1

*Example: **international** in 2 embedding spaces*

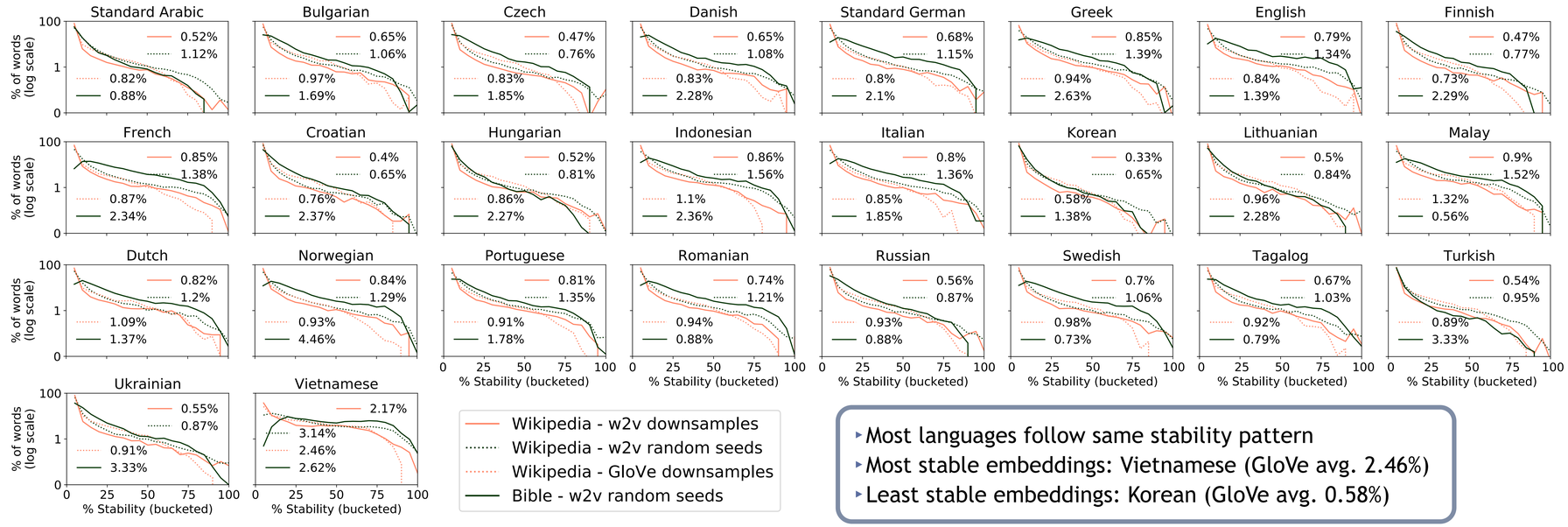
Stability = 40%

Model 1	*Model 2*
metropolitan	ballet
national	**metropolitan**
egyptian	bard
rhode	chicago
society	**national**
debut	state
folk	exhibitions
reinstallation	**society**
chairwoman	whitney
philadelphia	**rhode**

Stability for Wikipedia and the Bible

We compare the stability of embeddings for 26 languages.

- Wikipedia (3 settings): Stability of…
  - GloVe embeddings across 5 downsampled corpora
  - word2vec (w2v) embeddings across 5 downsampled corpora
  - w2v using 5 random seeds on 1 downsampled corpus
- One setting for the Bible: Stability of w2v embeddings using 5 random seeds on 1 downsampled corpus
- 26 languages in both Wikipedia and the Bible
- Each downsampled corpora 100,000 sentences



- Most languages follow same stability pattern
- Most stable embeddings: Vietnamese (GloVe avg. 2.46%)
- Least stable embeddings: Korean (GloVe avg. 0.58%)

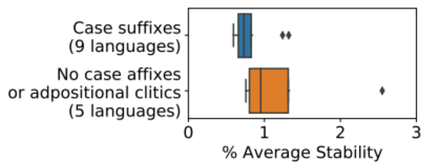
Regression Modeling

We use a regression model to predict stability in a language using linguistic properties.

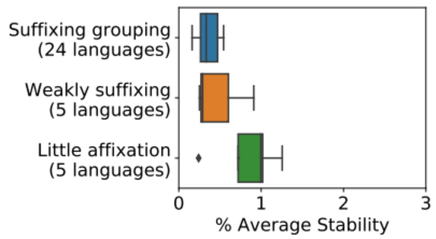
- Ridge regression
- 37 languages
- Input: 97 WALS properties
- Output: Average stability of all the words in a language
- High R^2 score of 0.96 ± 0.00

- More affixing (suffixing and prefixing) associated with lower stability
  - Affixes cause increased word variation
- Languages with no gender system associated with higher stability
  - Languages with gender systems have more word forms

Selected WALS Properties Associated with Affixing

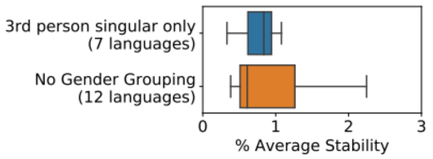


Position of Case Affixes

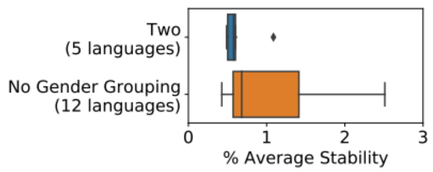


Prefixing v. Suffixing in Inflectional Languages

Selected WALS Properties Associated with Gender



Gender Distinctions in Independent Personal Pronouns



Number of Genders

This material is based in part upon work supported by the National Science Foundation (grant #1815291) and by the John Templeton Foundation (grant #61156).