Analyzing the Surprising Variability in Word Embedding Stability Across Languages

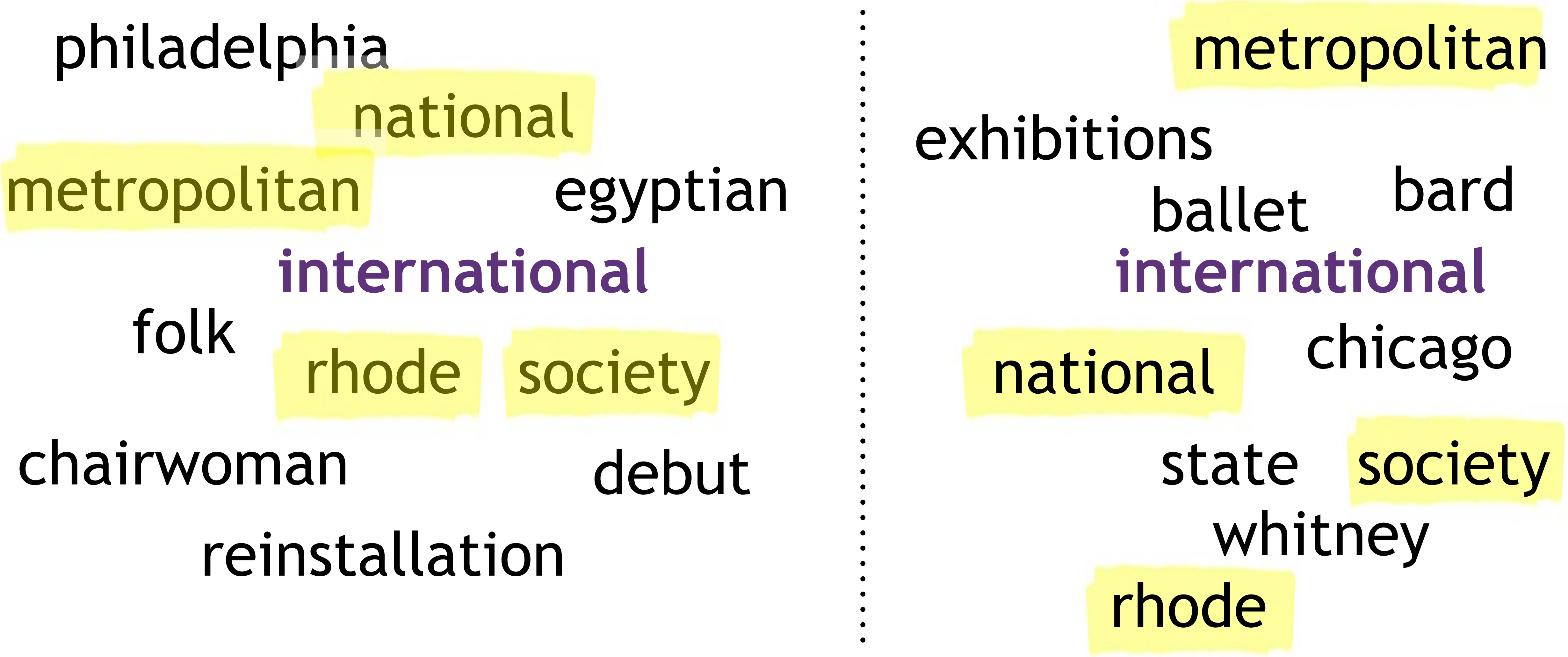
Laura Burdick, Jonathan K. Kummerfeld, Rada Mihalcea

University of Michigan

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

What is Stability?

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



philadelphia

national

metropolitan

egyptian

international

folk

rhode

society

chairwoman

debut

reinstallation

metropolitan

exhibitions

bard

ballet

international

chicago

national

state

society

whitney

rhode

Stability = 40%

This Work

Does stability vary for different languages?

Is stability associated with linguistic properties?

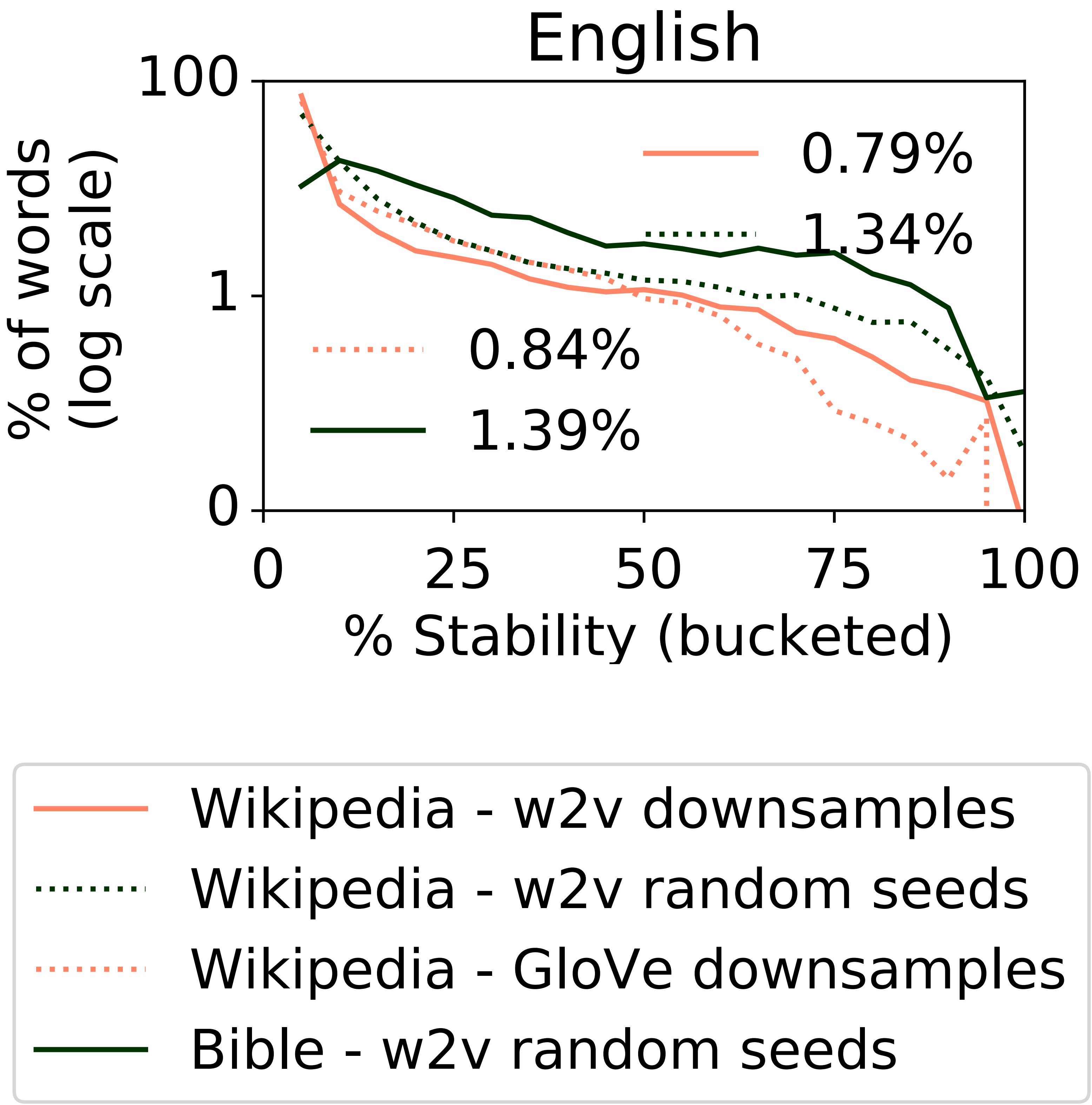
- **Data**
  - **Wikipedia** (40 languages)
  - **Bible** (97 languages)
  - **World Atlas of Language Structures (WALS)**, phonological, lexical, and grammatical properties (>2,000 languages)

Stability for Wikipedia and the Bible

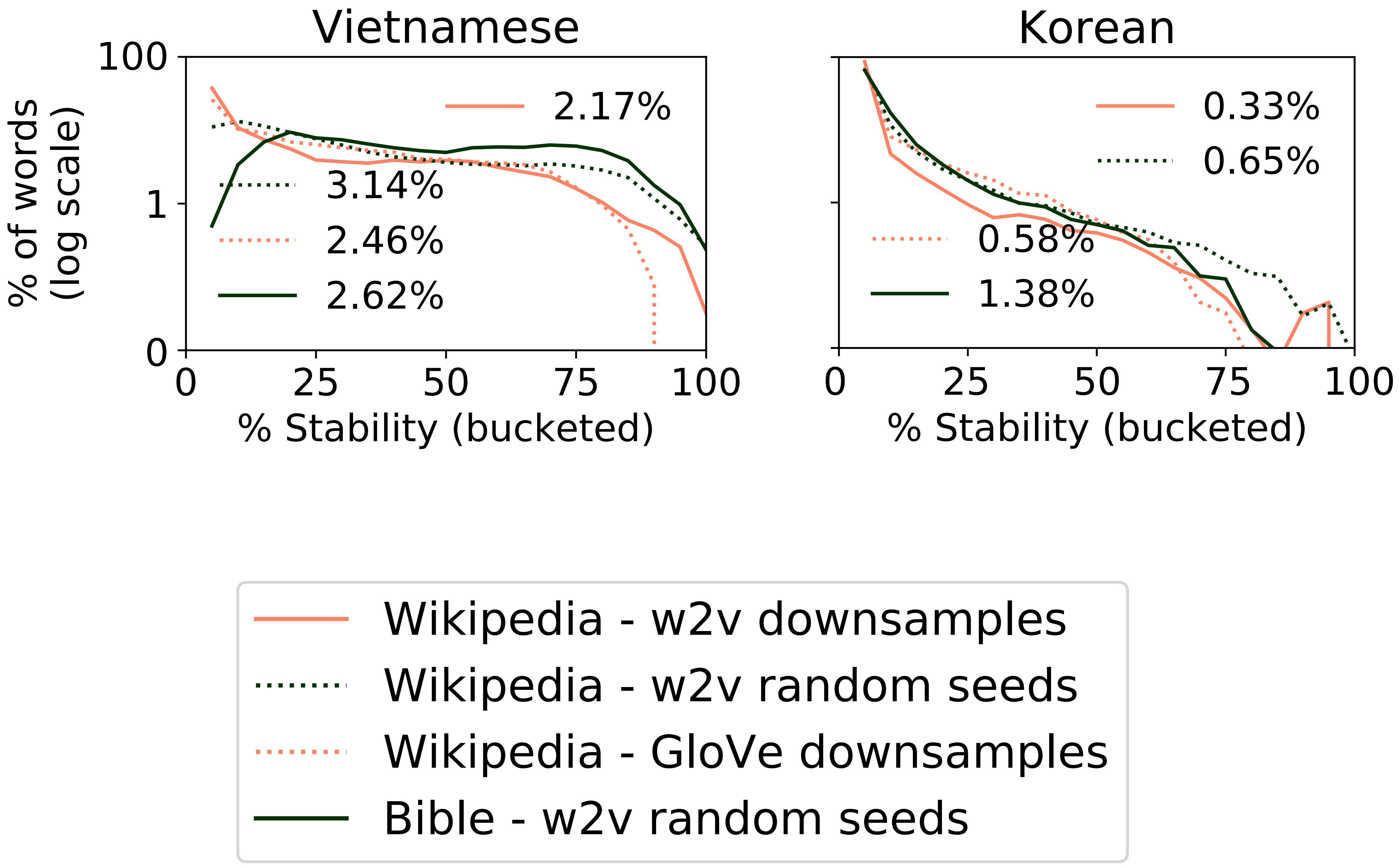
We compare the stability of embeddings for 26 languages.

- **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
- Each downsampled corpora 100,000 sentences

Stability for Wikipedia and the Bible



Stability for Wikipedia and the Bible



Regression Modeling

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

- Ridge regression

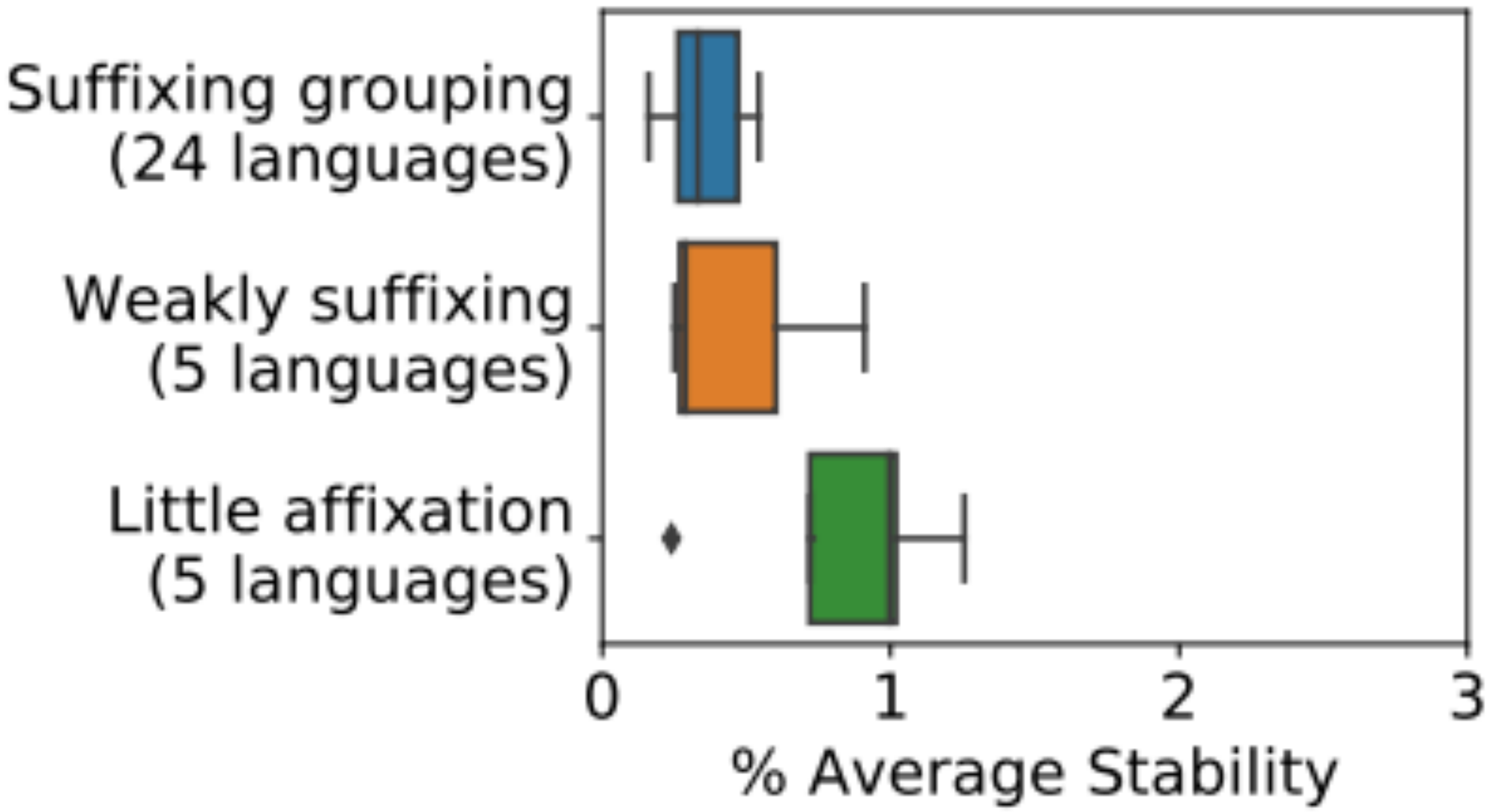
- 37 languages

- Input: 97 WALS properties
- Output: Average stability of all the words in a language

- High R^2 score of 0.96 ± 0.00

Regression Modeling

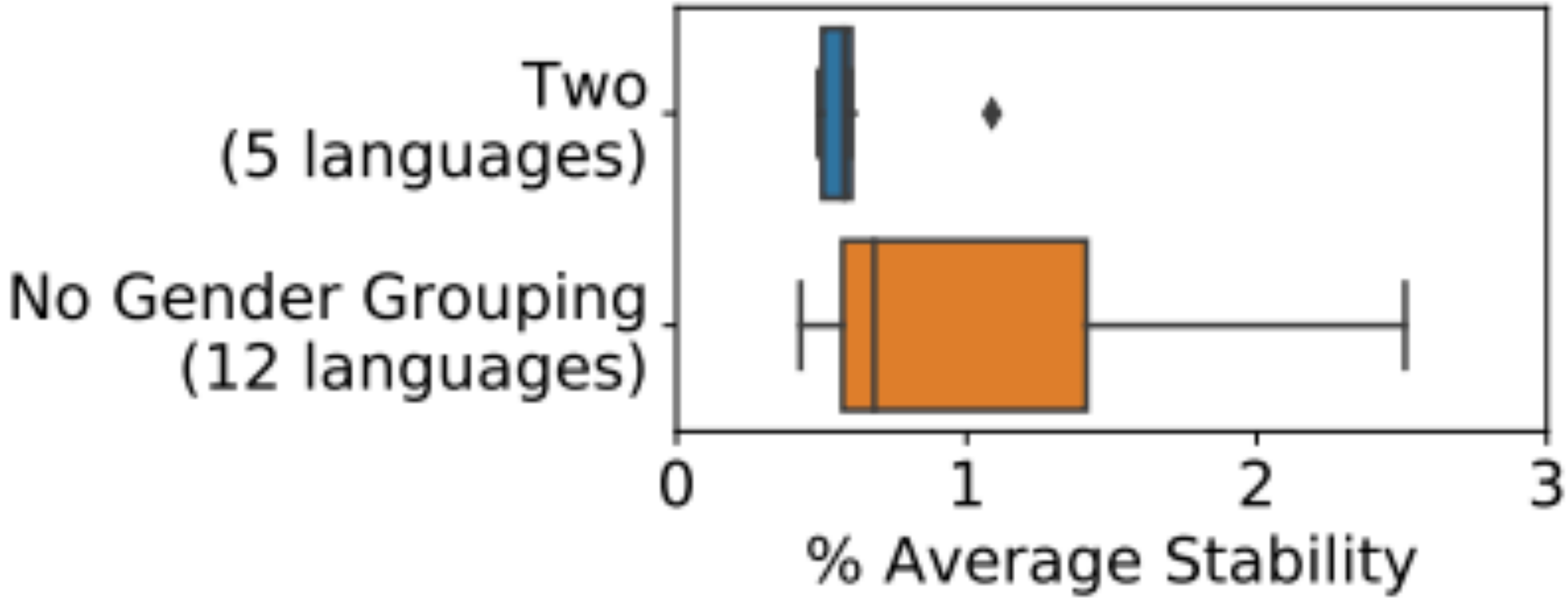
- More affixing (suffixing and prefixing) associated with lower stability
  - Affixes cause increased word variation



Prefixing v. Suffixing in Inflectional Languages

Regression Modeling

- Languages with no gender system associated with higher stability
  - Languages with gender systems have more word forms



Number of Genders

Final Thoughts

- *Languages with more affixing tend to have less stable embeddings*
- *Languages with no gender systems tend to have more stable embeddings*
- *Future embedding design needs to take into account gendered words and morphologically rich words with affixes*

Download our code:

`http://lit.eecs.umich.edu/downloads.html`