Using language models for holistic language variety comparisons

Joshua McNeill

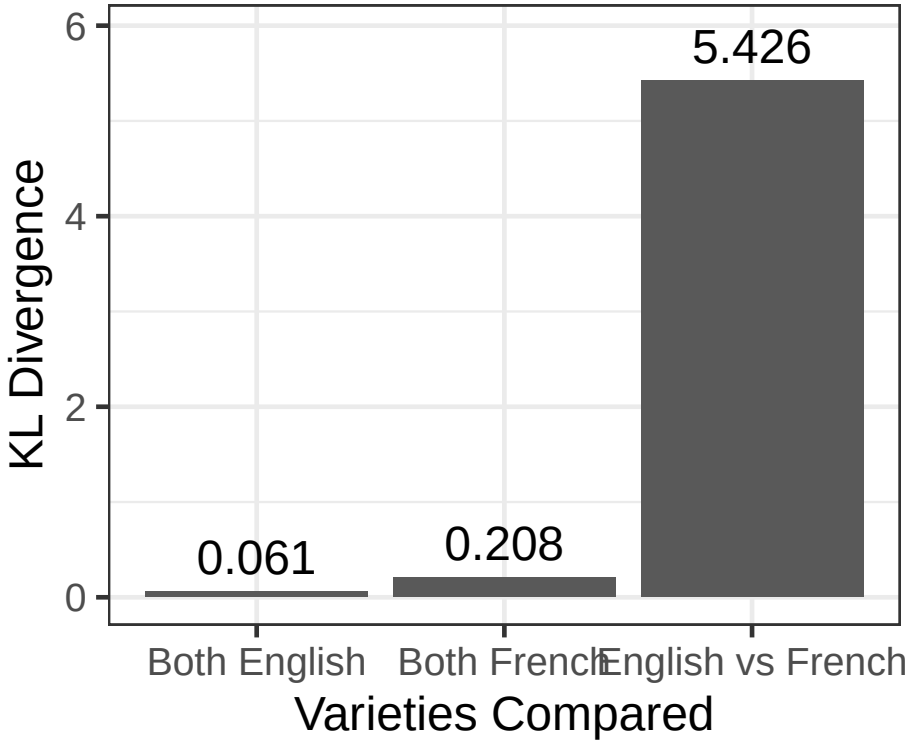
University of Georgia

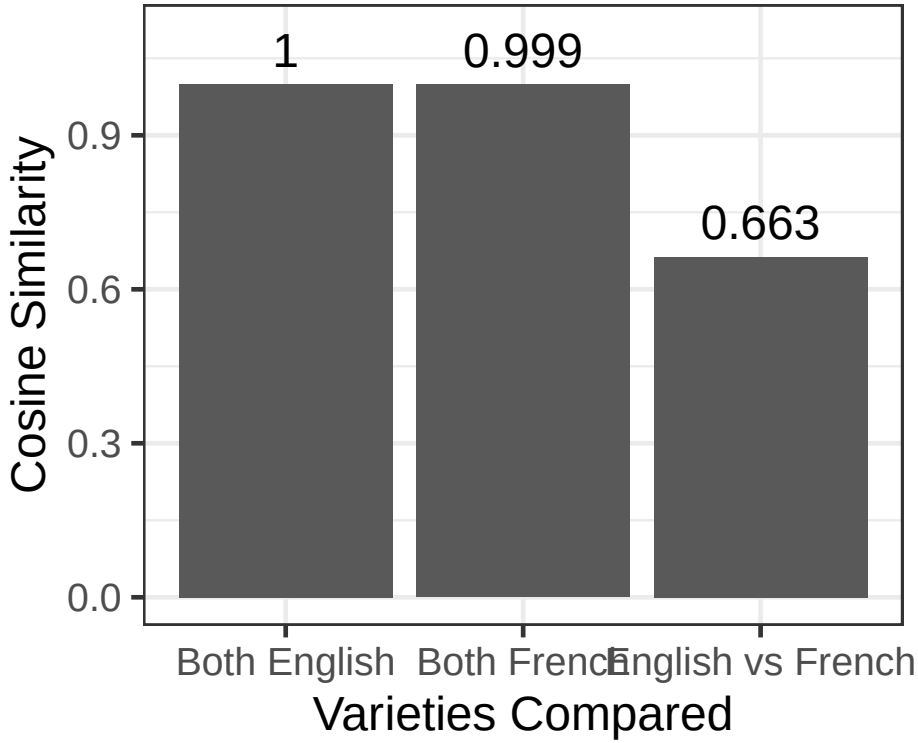
29 July 2021

Data and code available at https://osf.io/9cjpw/.

The study

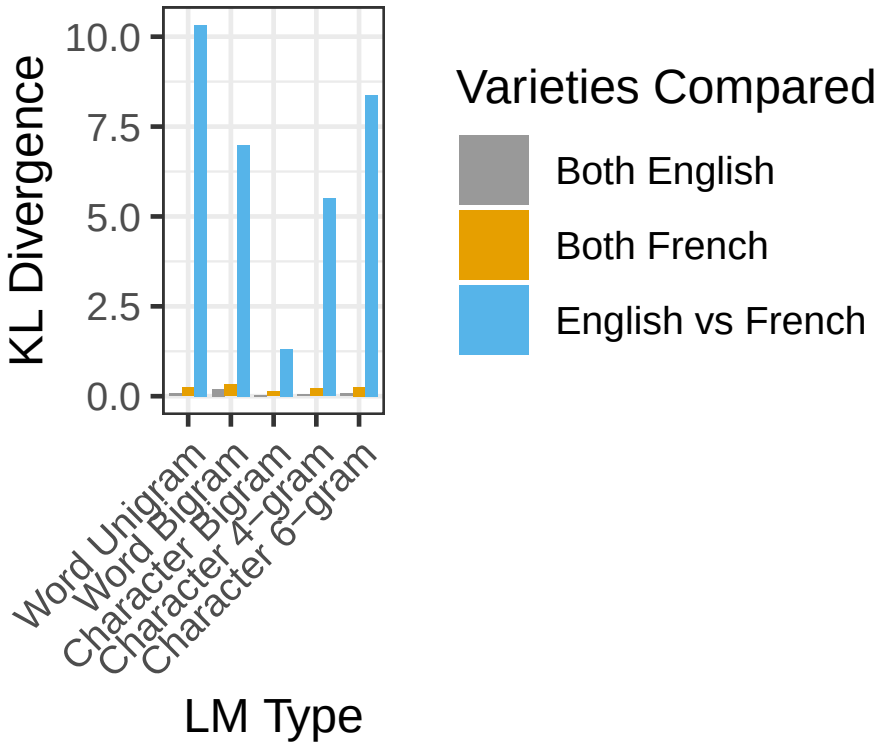
- Goal: Develop a holistic measure of linguistic distance between varieties
- Why: Traditional methods generalize from very small sets of linguistic variables
- Method: Incorporate LMs from NLP to capture all linguistic items



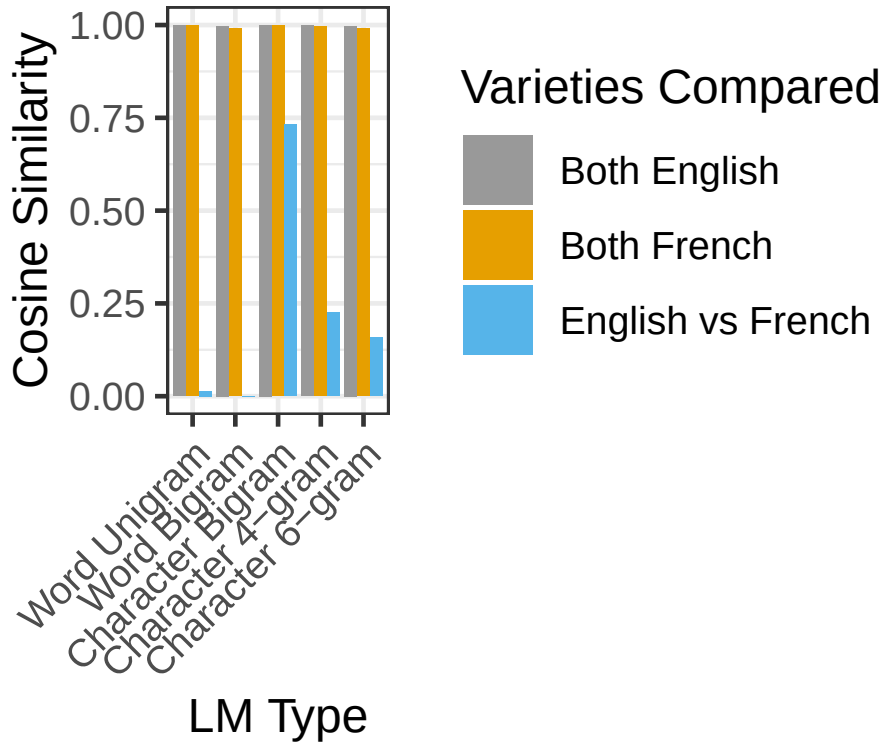


KL divergence values comparing the combination of all LMs, representing all linguistic levels at once, for pairs of language varieties, rounded to the nearest thousandth

Cosine similarity values comparing the combination of all LMs, representing all linguistic levels at once, for pairs of language varieties, rounded to the nearest thousandth



KL divergence values comparing individual LM types, each representing one linguistic level, for pairs of language varieties



Cosine similarity values comparing individual LM types, each representing one linguistic level, for pairs of language varieties