Using language models for holistic language variety comparisons

Joshua McNeill

University of Georgia

9 October 2021

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

1 Distinguishing between varieties

2 Language models (LMs)

3 Methods

4 Results

5 Discussion

6 References

1 Distinguishing between varieties

2 Language models (LMs)

3 Methods

4 Results

5 Discussion

6 References

Past studies

1. **Atlases in dialectology** used many linguistic features, often lexical (Gilliéron & Edmont, 1902; Wenker, 1881/2020)
2. **Modern studies** have focused on a handful of socially salient variables (Dubois & Horvath, 2003; Eckert & Labov, 2017; Rickford & McNair-Knox, 1994)
3. **Recent online dialectology** returns to including many lexical items (Eisenstein, 2014; Grieve et al., 2018; Pavalanathan & Eisenstein, 2015)

Research question

Given these limits:

1. Dialectology > Many but only **lexical** features
2. Sociolinguistics > **Few** but diverse socially salient variables

Can language models be used to capture many diverse linguistic features to more holistically quantify the difference between language varieties?

1 Distinguishing between varieties

2 Language models (LMs)

3 Methods

4 Results

5 Discussion

6 References

Basic idea: n-gram LMs

Europarl (Koehn, 2009)

```
Madam President, can you tell me why this Parliament does
not adhere to the health and safety legislation that it
actually passes? Why has no air quality test been done on
this particular building since we were elected?
```

Word Bigram	Tokens
Madam President	1
President can	1
can you	1
you tell	1
tell me	1
me why	1
…	…
$bigram_n$	$tokens_n$

Character 4-gram	Tokens
mada	1
adam	1
dam_	1
am_p	1
m_pr	1
_pre	1
…	…
$4gram_n$	$tokens_n$

1 Distinguishing between varieties

2 Language models (LMs)

3 Methods

4 Results

5 Discussion

6 References

Data

The Europarl corpus (Koehn, 2009)

1. Transcriptions of speeches from proceedings of the European Union dating back to 1996
2. Parallel corpus (i.e., speeches translated into several languages)
3. 50,263,238 words in English
4. 52,562,008 words in French
5. Well controlled for speech style and topic

Excerpt

```
Still on the subject of Wednesday' s sitting, I have
another proposal regarding the oral question on capital
tax. The PPE-DE Group is requesting that this item be taken
off the agenda.
```

LMs used

Training data:

1. English transcriptions vs French transcriptions
2. 1st half of the English transcriptions vs 2nd half of the English transcriptions
3. 1st half of the French transcriptions vs 2nd half of the French transcriptions

The following LMs were constructed from each training set following Duvenhage (2019):

4. Word n-gram LMs
   $\rightarrow$ unigram and bigram
5. Character n-gram LMs
   $\rightarrow$ bigram, 4-gram, and 6-gram

Analysis

To quantify the difference between varieties:

1. **KL divergence**
   - Range from 0 (no similarity) to ∞ (very similar)
2. Cosine similarity
   - Range from -1 (opposite) to 0 (perpendicular) to 1 (identical)

1 Distinguishing between varieties

2 Language models (LMs)

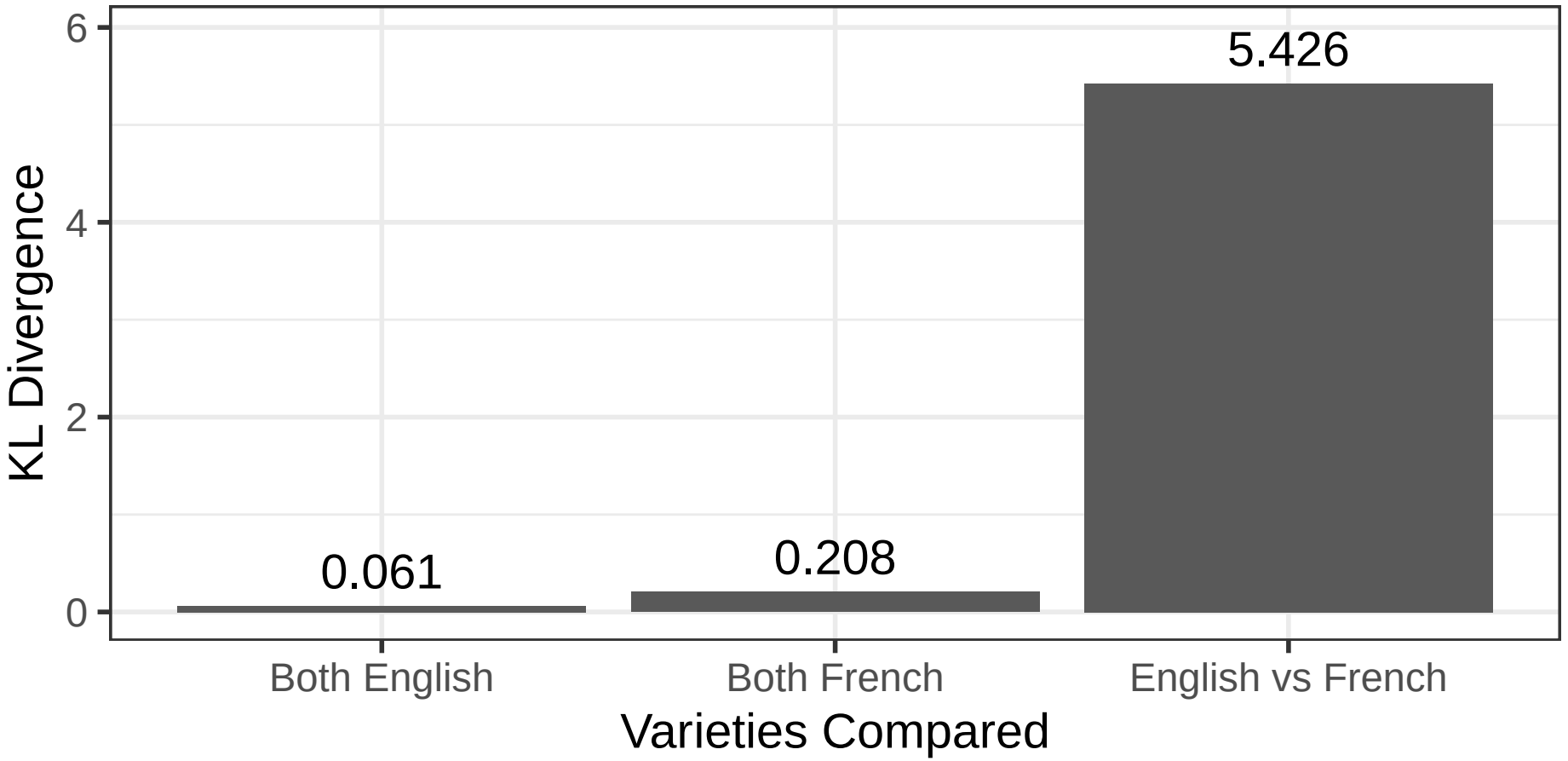
3 Methods

4 Results

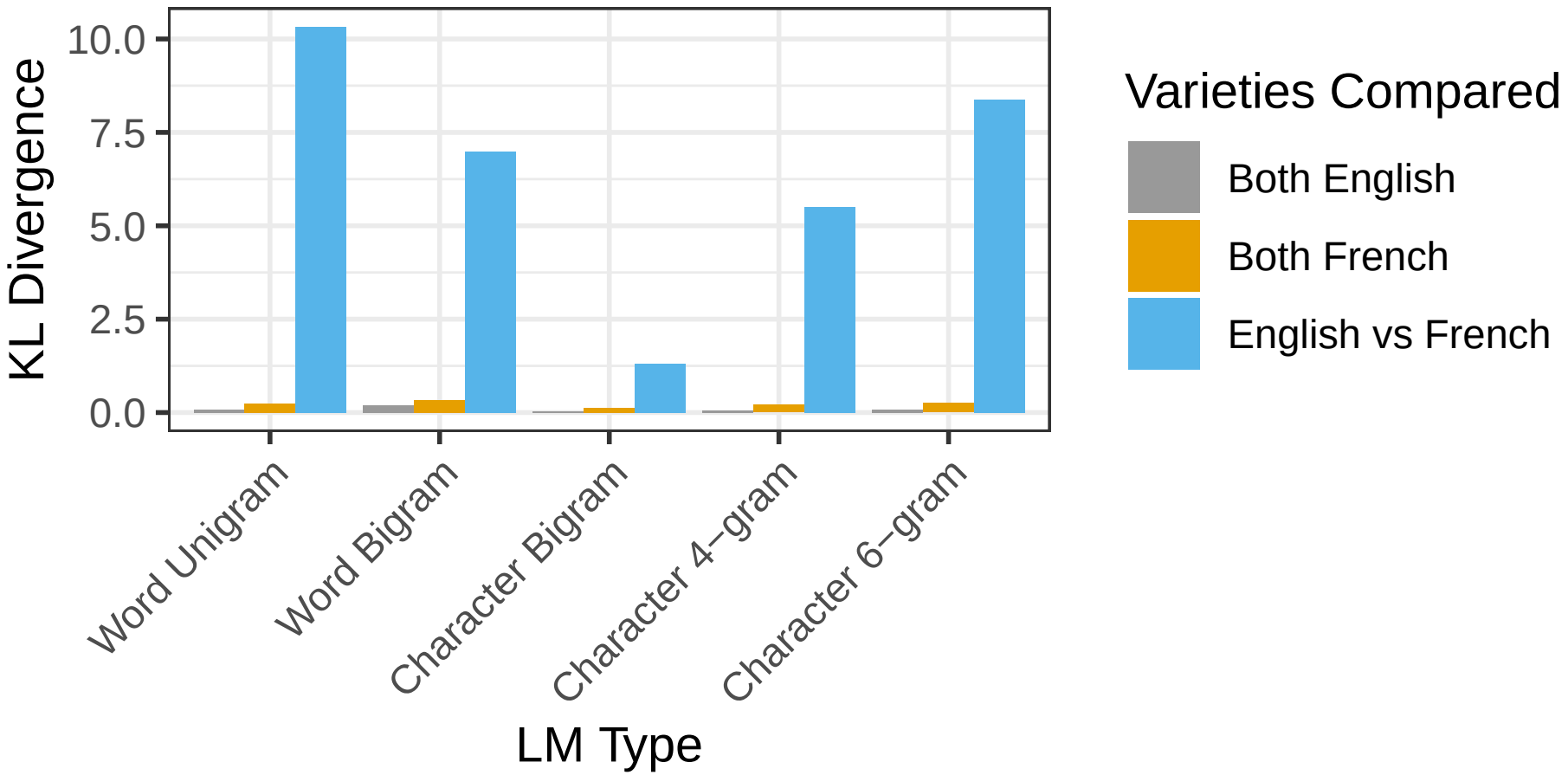
5 Discussion

6 References

Collated LMs



By LM type


KL Divergence
10.0
7.5
5.0
2.5
0.0
Word Unigram
Word Bigram
Character Bigram
Character 4−gram
Character 6−gram
LM Type
Varieties Compared
Both English
Both French
English vs French

Character bigrams

10 Most Frequent	
English	French
e _	e _
_ t	s _
th	_ d
s _	es
_ a	_ l
he	on
t _	t _
n _	en
in	de
on	nt

Some matches are spurious

1. *on* is a preposition in English but a subject pronoun in French
2. Words ending in ⟨t⟩, ⟨s⟩, or ⟨e⟩ are common in both languages, but only ⟨s⟩ is likely to have a shared meaning (i.e., plurality)

Character bigrams arguably do not carry enough linguistically distinguishing information

Capturing the lexicon: Word unigrams

5 Most Frequent	
English	French
the	de
of	la
to	et
and	le
in	Ã
Eng Half 1	Eng Half 2
the	the
of	of
to	to
and	and
in	in

Matches are as expected

1. No matches between English and French
2. Everything matches between English and English

Function words are given greater weight

Capturing syntax: Word bigrams

5 Most Frequent	
English	French
of the	de la
in the	Ã la
to the	et de
the European	que nous
on the	Monsieur le
Eng Half 1	Eng Half 2
of the	of the
in the	in the
to the	to the
the European	the European
on the	on the

This is lexical rather than syntactic

3. *of the* = *de la* : P Det
4. *in the/to the* = *à la* : P Det

Training on a PoS tagged corpus would overcome this

5. Can be tagged manually or automatically

Capturing morphology: Character 4-grams

5 Most Frequent	
English	French
_ the	_ de _
the _	_ la _
_ of _	tion
_ to _	ent _
and _	ion _
Eng Half 1	Eng Half 2
_ the	_ the
the _	the _
_ of _	_ of _
_ to _	_ to _
and _	and _

Function words are captured most frequently but also derivational morphemes in French

1. *ion*, *tion*, and *ent*
2. Only the form of the morphemes is contrasted

Capturing morphology: Character 6-grams

5 Most Frequent	
English	French
_ that _	ement _
_ of th	ation _
n the _	_ de la
of the	de la _
f the _	_ dans _
Eng Half 1	Eng Half 2
_ that _	_ that _
_ of th	_ of th
n the _	n the _
of the	of the
f the _	f the _

Again, function words are the most frequent items captures but also some derivational morphemes in French

1. *ment* and *tion*
2. For both function words and morphemes, additional context is captured
3. Again, only forms are contrasted

Capturing morphology: Character 6-grams

5 Most Frequent	
English	French
_ that _	ement _
_ of th	ation _
n the _	_ de la
of the	de la _
f the _	_ dans _
Eng Half 1	Eng Half 2
_ that _	_ that _
_ of th	_ of th
n the _	n the _
of the	of the
f the _	f the _

This really captures orthographic differences and, to some extent, phonetic differences

2. Training character n-gram LMs on a phonetic transcription would better capture phonetic differences

1 Distinguishing between varieties

2 Language models (LMs)

3 Methods

4 Results

5 Discussion

6 References

Limitations

1. Special care needs to be given to the LMs used
2. Special care needs to be given to the corpora used
   - This method likely won't work well with small corpora either
3. This method gives no great value to particularly salient features

What's next

1. Explore using PoS tagging to better capture syntax in the LMs
2. Look into the impact of using corpora of different sizes
3. Apply the method to data that has already been analyzed with traditional variationist methods

1 Distinguishing between varieties

2 Language models (LMs)

3 Methods

4 Results

5 Discussion

6 References

Dubois, S., & Horvath, B. M. (2003). Verbal Morphology in Cajun Vernacular English: A Comparison with Other Varieties of Southern English. *Journal of English Linguistics*, *31*(1), 34–59. https://doi.org/10.1177/0075424202250296

Duvenhage, B. (2019). Short Text Language Identification for Under Resourced Languages [arXiv: 1911.07555]. *arXiv:1911.07555 [cs]*.

Eckert, P., & Labov, W. (2017). Phonetics, phonology and social meaning. *Journal of Sociolinguistics*, *21*(4), 467–496. https://doi.org/10.1111/josl.12244

Eisenstein, J. (2014). Identifying regional dialects in online social media. *Preprint*, 1–15. https://doi.org/10.1002/9781118827628.ch21

Gilliéron, J., & Edmont, E. (1902). *Atlas linguistique de la France. Notice, servant a l'intelligence des cartes*. Paris H. Champion.

Grieve, J., Nini, A., & Guo, D. (2018). Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics*, *46*(4), 293–319. https://doi.org/10.1177/0075424218793191

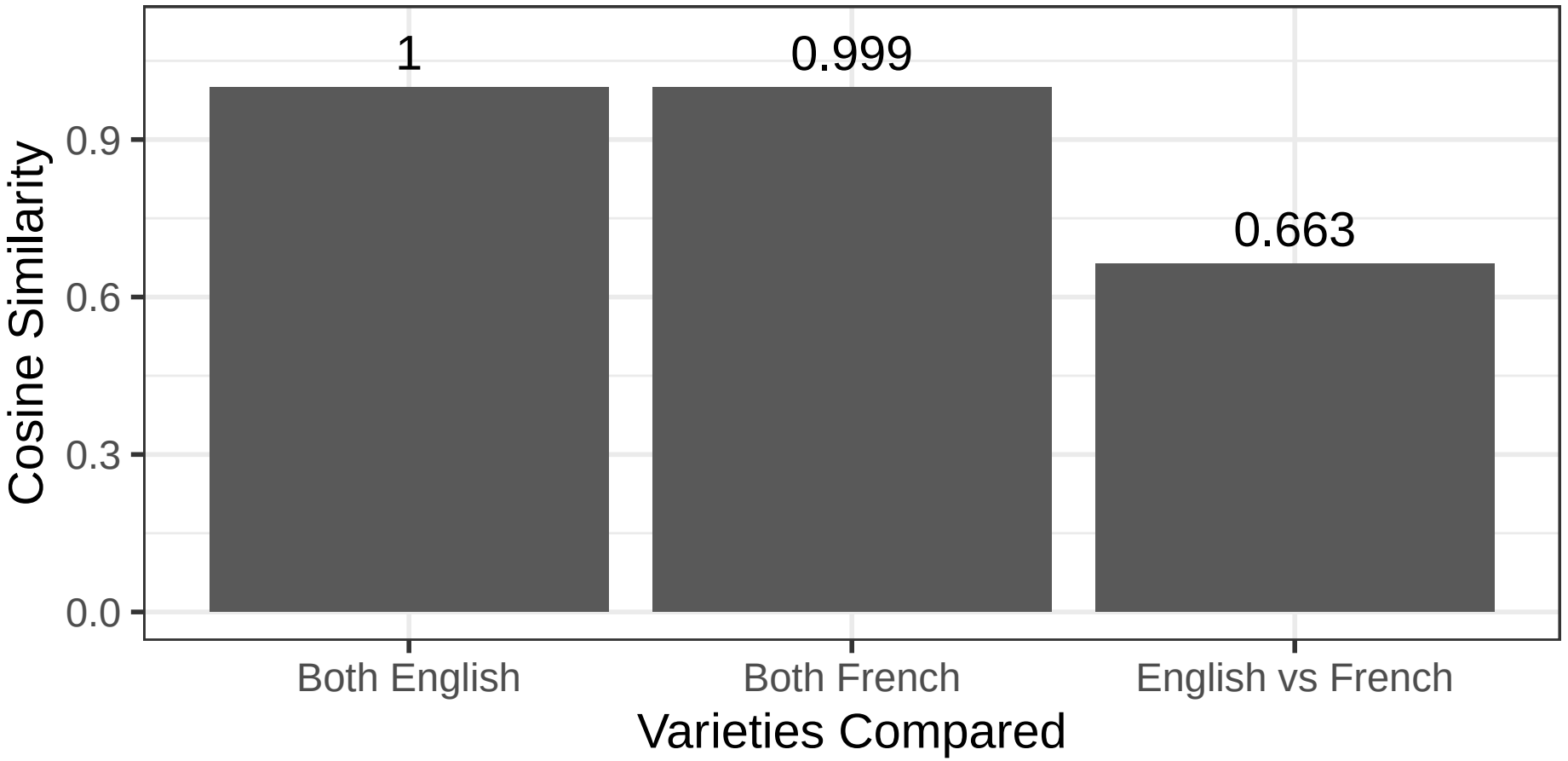
Koehn, P. (2009). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of MT Summit X*, 8.

Pavalanathan, U., & Eisenstein, J. (2015). Audience-Modulated Variation in Online Social Media. *American Speech*, *90*(2), 187–213. https://doi.org/10.1215/00031283-3130324

Rickford, J., & McNair-Knox, F. (1994). Addressee- and Topic-Influenced Style Shift: A Quantitative Sociolinguistic Study. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 235–276). Oxford University Press.

Wenker, G. (2020). *Sprach-Atlas von Nord- und Mitteldeutschland auf Grund von systematisch mit Hülfe der Volksschullehrer gesammeltem Material aus circa 30.000 Orten bearbeitet, entworfen und gezeichnet: Text. Einleitung.* Walter de Gruyter GmbH & Co KG. (Original work published 1881)

Bonus slides



Bonus Slides