Using language models for holistic language variety comparisons*

Joshua McNeill

July 9, 2021

1 Introduction

One of the goals of scientific research is to discover generalizable facts, and work on language variation is no different. Variationists are often interested in describing language varieties and distinguishing between language varieties. This task is typically accomplished by analyzing several salient linguistic variables in fine detail and either generalizing to whole varieties from these several variables. The focus on deeply analyzing a small number of variables has advanced the understanding of language variation in more ways than can be mentioned here. While this paradigm shift from the broad analyses of many variables done dialectology, out of which language variation developed, has been valuable, modern technological advances have made it possible to once again consider analyzing broad swaths of linguistic variables at once. The objective for this study is thus to develop a method for holistically quantifying the distance between language varieties in a way that can account for practically all potential linguistic variables.

Some work on language variation has to already returned to its dialectological roots at least in spirit, and so a look at what has been done recently will help situate the method being proposed here. As such, section 1.1 will discuss some advantages and disadvantages of traditional methods of analyzing language variation to show what can be gained from a more holistic approach, and section 1.2 will review the limits of how holistic such traditional studies tend to get. Section 1.3 will explain a bit of what is done in the field of natural language processing (NLP), and section 1.4 will review how such technological advances have been applied to variationist studies, leading to the inclusion of many more variables than was previously practical.

1.1 Generalizing from saliency

In traditional variationist work, generalizations about whole varieties are sometimes drawn from a relatively small set of linguistic variables, perhaps as few as five. The assumption underlying that makes this possible is that the variables analyzed are salient and so have social meaning and so are what really distinguish varieties. This connection between saliency and social meaning has indeed been acknowledged in variationist literature (Podesva, 2011, p. 235) and has at least been implicit in variationist theory since its foundations.

Labov (1972) himself introduced the idea of saliency into language variation in proposing the concepts of linguistic variables being either indicators, markers, or stereotypes, the latter two carrying saliency. It was later claimed that most linguistic variables are markers (Bell, 1984), carrying saliency but not being explicitly commented on by speakers when describing one variety versus another. The importance of saliency for classifying types of linguistic variables was also part of the concept of orders of indexicality (Silverstein, 2003), where second and third order indexicals would be somewhat equivalent to markers and stereotypes, respectively, in terms of saliency.

Saliency has not only been a consistent aspect of variationist theorizing, but it has proved useful for analyses. One example is in analyzing how varieties comes to be social constructs, recognized as existing in the minds of speakers, which has been referred to in language variation as enregisterment (Agha, 2003). In essence, the process of enregisterment is based on features increasing in saliency. This was initially applied to Received Pronunciation (Agha, 2003) but has also been applied to the development of Pittsburghese based mostly on an analysis of (aw) monophthongization, though some other features are discussed also such as the term *yinz* 'you (pl.)' (Johnstone et al., 2006). Additionally, the saliency has been used to distinguish varieties spoken by different groups within a single geographic area in that residents will explicitly describe linguistic features that they associate with each group. This was the case in a study of Beijing where two types of Beijingers and one type of non-Beijinger were associated with one linguistic variable each (Zhang, 2005).[1]

The sort of explicit commentary noted in Zhang (2005) is one way to establish whether a linguistic variable is salient or not. To quantify such commentary requires developing indices where perhaps certain types of commentary each count in the index or commentary in general is counts as just one item among other types in the index. In fact, saliency has sometimes even been quantified on purely linguistic grounds (Podesva, 2011, p. 237).[2] Perhaps a more common way of identifying saliency has come through experimentation, however. Perceptual studies that invovle experimentation fall under paradigms such as eye-tracking (e.g., D'Onofrio, 2015) or matched-guise experiments (e.g., Campbell-Kibler, 2009; Delforge, 2012). Findings in such studies include showing that preconceptions of Valley Girls and Californians

---

[1]Technically, one group, the "smooth operators", was associated with two variables, but each variable involved [ɹ] (Zhang, 2005, pp. 441-444).

[2]Podesva (2011) does of course acknowledge social saliency, as well.

can impact perceptions of the TRAP vowel (D'Onofrio, 2015), that the alveolar variant of (ing) is judged differently depending on the listener (Campbell-Kibler, 2009), and that vowel devoicing in Andean Spanish is less noticeable by listeners when they do not want to notice it, as those who denied having the feature still had it at similar rates to those from elsewhere in whom the feature was regularly perceived (Delforge, 2012).

The point is that saliency does indeed exist, can be measured, and carries import for analyses, particularly if one is interested in social meaning. To the extent that varieties are social constructs, it is also a rather safe assumption to say that salient variables are what distinguish varieties as that which is not socially meaningful could not be a defining feature of a social construct. However, when looking at varieties more as concrete systems that are coherent, consistent, and objectively different from each other, comparing salienct variables alone may miss even a large number of distinctions that a more holistic comparison could capture. Furthermore, being able to compare between varieties both holistically and by looking at only salient variables opens the possibility of finding contrasts that may be enlightening, such as the possibility that two varieties different in many more ways than what is noticeable to speakers, yet the focus in variationist work has mainly been on salient variables alone.

1.2   Looking at clusters of linguistic variables

Language variation studies have of course examined clusters of variables and systems of interrelated variables. The focus in these cases is still linguistic variables that are thought to be salient, though the number of variables analyzed has at times approached a relatively larger scale. This is particularly true when the analysis focuses on "features", which are effectively treated as linguistic variables that have two realizations: present or absent.

Studies of African-American English (AAE)[3], perhaps due simply to there being so many, include both analyses of small clusters of variables but also larger sets of features. In some cases, there is an apparent relationship between the variables, such as in Rickford and McNair-Knox (1994) where the analysis included the zero copula, habitual invariant *be*, the plural *-s* morpheme, the 3rd person present *-s* verbal inflection, and the possessive *-s* morpheme. The first two of these both implicate the verb *to be* whereas the last three all involve homophonous suffixes. Such relationships are not always present, however, as in an examination of Martin Luther King, Jr's language that included the zero copula, (ing), (t/d) deletion, (r), word-final consonant cluster reduction, and prosody (Wolfram et al., 2016). If a structural relationship exists between these linguistic variables, it is not evident at first glance.

When sets of features from AAE are studied, the number of variables grows substantially. In the case of Van Hofwegen and Wolfram (2010), 32 features were tallied in individuals' speech as a way to measure how heavily AAE their

---

[3]I am including what researchers may refer to as African-American Vernacular English (AAVE) under this term as the distinction is not particularly important for the present study.

speech was. The measure itself, along with the chosen features, was developed in research on speech pathology and education and where it is referred to as the dialect density measure (DDM). While the proposal of an appellation such as DDM suggests something of a novel approach, similar measures have in fact been used elsewhere, as well. In a study looking into the use of AAE features in gay British men on Twitter, 25 AAE features were analyzed according to their frequencies (Ilbury, 2020). Also involving Twitter, over 30 non-standard spellings thought to be associated with AAE were analyzed in order to draw AAE dialect maps in the US (Jones, 2015). While these feature-conceptualized studies include far more than the typical five to seven clustered variables, they still fall well short of including all possible variables in these varieties.

If studies of AAE sometimes include clusters of linguistic variables that appear structurally related, it is not always clear that this structural relationship is what drove the choice of variables. In many other cases, there is little question that the researchers are interested in a cluster of variables because of their structural relationship. This is particularly evident in examinations of vowel systems or large portions of vowel systems, where there is often though not always a chain shift involved. This has been done since the foundational variationist work in New York City (Labov, 2006) and has continued in studies of other areas such as Belfast (L. Milroy, 1987), King of Prussia (Payne, 1980), Syndey (Horvath & Sankoff, 1987), the Detroit area (Eckert, 2000), and Philadelphia (Eckert & Labov, 2017). Similarly, though without implicating a chain shift, the majority of the pronoun system of Louisiana French as spoken in Houma has been analyzed as a whole (Rottet, 1995) as well as chunks of the verbal morphology system and and copula system in Cajun English (Dubois & Horvath, 2003). Finally, perhaps one of the most explicit treatments of a system of covariation was Guy's (2013) analysis of Brazilian Portuguese as spoken in Rio de Janeiro which had the explicit goal of examining the existence or non-existence of distinctive sociolects by determining if four linguistic variables covaried. While these sort of studies go further towards covering all the variables of particular linguistic systems, there is once again much left to be described if one wishes to compare the distance between varieties in their entireties. Furthermore, not all subsystems of varieties are contained in such a relatively small number of variables. For instance, it is not possible the lexical system of a variety with five, ten, or even one hundred variables.

1.3 Advances in NLP

A significant part of what has made it possible to analyze much larger numbers of linguistic variables at once is the continued development of NLP technology. Indeed, not only has development in NLP become more and more active, NLP has also made its way into other fields. This includes sociolinguistics, where the resulting interdisciplinary field has become known as computational sociolinguistics (Nguyen et al., 2016). While a thorough discussion of NLP is beyond the scope of the present study, a brief overview of the basics of what can be done and how is relevant to make the holistic method of comparing vari-

4

eties proposed here more understandable. NLP has certainly found its way into language variation, but not every variationist is a computational sociolinguist.

The building of large corpora is not purely an activity related to NLP, but they are central to the development of NLP tools and sizes have exploded over the last decade. Sociolinguistic interviews themselves yield corpora, of course, and the first corpus to be over one million words of written English, the Brown corpus, was collected already in the 1960s (Francis & Kucera, 1964). Efforts to digitize written language have continued through the present day with corpora such as the British National Corpus (BNC) consisting over just under 100 million words (Burnard, 2007), the Corpus of Contemporary American English (COCA) consisting of over 400 million words at its start (Davies, 2010), and Google Books consisting of over 40 million books (Lee, 2019).

In addition to digitized written language, data mining techniques have made collecting large contemporary corpora from social media platforms so prevalent that there is more of a question about the ethics behind such activities than about how to accomplish the task (D'Arcy & Young, 2012). Corpora from Twitter, for instance, have reached sizes such as 62 million tweets (Hong et al., 2011, p. 519) or 114 million tweets (Eisenstein, 2013, p. 13) or even 924 million tweets (Huang et al., 2016, p. 244), each tweet containing up to 140 characters[4]. Depending on the scope of the sort of data the researcher wishes to collect, such corpora can be constructed in relatively little time, as well, with even the largest example given taking only a year to finish. Of course, the extent to which such large corpora present opportunities for linguistic research is limited by the tools available to deal with so much data. NLP provides a number of tools that help expedite the sort of analyses that variationists might carry out, including part of speech tagging, sentence parsing, and document classification.

Corpora have been tagged for parts-of-speech (POS) since at least the development of the Brown corpus, which was done by hand. A common NLP task then is to perform POS tagging automatically on new corpora. This typically involves training a tagger on documents that were first hand-coded for POS, such as the aforementioned Brown corpus. The general idea behind training is to construct a dictionary made up of each word that has been seen in the hand-coded corpus with its most likely POS tag. This can be as simple as associating the most frequent POS tag for each word to that word but can also get much more complex, perhaps by taking into account the POS of the surrounding words, as well. For example, training might lead to listing *run* as most likely a verb, but if taking into account context, it might also list *run* preceded by a determiner as a noun. There are many possibilities and approaches, but the key point is that hand-coded data is always used as input for the training algorithm.

Related to POS tagging is the task of parsing sentences, which is also a classic task in NLP. This is done in much the same way as POS tagging where some training data is used to train a parser that can then automatically parse

---

[4]The character limit for tweets was subsequently raised to 280 characters (Rosen, 2017).

new sentences, though naturally the training data must be annotated for no just POS tags but also syntactic structures. Fortunately, like the hand-coded POS tags in the Brown corpus, there are today a number of corpora hand-coded with syntactic structures. These are referred to as treebanks, perhaps the most well-known example being the Penn Treebank for English (Prasad et al., 2019). Both automatic sentences parsers and POS taggers are useful for variationists looking at syntax but also potentially any other variationist whose linguistic variables may be syntactically conditioned. For instance, one may want a POS tagger if one is studying the (ing) variable and suspects it to varying according to lexical category (Houston, 1985).

A common NLP task that is particularly relevant to the current study is that of document classification, where a "document" is understood as any amount of text that acts as a unit, be it a tweet or a book. A classic example of this task is that of identifying the topic of a document. The goal in these cases might be to spot spam in e-mail (Dada et al., 2019), to classify news articles, to mine opinions in text, and so on (Minaee et al., 2021). The general approach to this problem is to take a collection of documents that have already been classified by hand and once again use them as input for a training algorithm in order to train a classifier that can be used on new documents. The training algorithm may draw conclusions from the input data such as the use of the word "senator" over a certain number of times increases the likelihood that a new article is about politics. The result is not entirely different from the dictionaries generated by POS tagging.

The classification task related to the current study is that of language variety identification, which as it sounds, refers to classifying the language variety in which a document is written. There is a related task called language identification, but the former involves varieties that are quite similar and the latter varieties that are quite different. Language variety identification is not as old of a task in NLP as some of the others, but work has been done on Arabic varieties during the MADAR workshop (Bouamor et al., 2019), and Balkan languages, Indonesian versus Malaysian, Portuguese varieties, Spanish varieties, French varieties, and Persian versus Dari during the VarDial workshop (Zampieri et al., 2017). A particularly effective approach to the tasks involved in the Var-Dial workshop made use of n-gram language models (LMs) (Duvenhage, 2019). An n-gram LM is generated from training documents in that the training algorithm makes a list of all the n-grams found in the training documents along with their frequencies. An n-gram itself is some n number of consecutive symbols, which are typically consecutive characters or consecutive words. For instance, in sentence 1 below, the word bigrams are ⟨this is⟩, ⟨is a⟩, and ⟨a sentence⟩, whereas the character 4-grams are ⟨this⟩, ⟨his ⟩, ⟨is i⟩, ⟨s is⟩, ⟨ is ⟩, and so on. An n-gram LM then would be a list of all the n-grams of whichever type the researcher decides to use from all the documents that are known to be in a particular language variety. The classifier would then make a decision on the language variety of new texts based on their relationship to the n-gram LMs that were constructed. In the case of Duvenhage (2019), he achieved very good results by using a combination of word unigrams and bigrams and character

bigrams, 4-grams, and 6-grams which influenced the use of these LM types in the present study, as will be discussed further in section 2.2 of the Methods section.

1. ⟨this is a sentence⟩

These tools help variationists relegate much work that used to be done by hand to a machine, allowing the scale projects to increase substantially. Admittedly, it is still prudent to go through samples of data that has been automatically coded for quality assurance, but doing so is much quicker than coding by hand. Additionally, perhaps the biggest yet simplest technological advancement that has aided variationists, particularly when considering how many variables can be analyzed at once, has just been counting. A few lines of basic code can return counts for tokens of linguistic variables from millions to even billions of words of text, something that would have taken an enormous amount of time in the early days of the field.

1.4   The impact of NLP on language variation research

As computational sociolinguistics has developed in recent years, bringing NLP technology into language variation research, studies have begun to analyze many more variables at once to diverse ends. Often these studies mine data from Twitter which presents certain advantages such as broad demographic coverage (Wojcik & Hughes, 2019, p. 5), good API access, and fewer ethical grey areas than other platforms as tweets are public by default, accessible even to those without Twitter accounts themselves, which leads users to being less likely to post confidential information (Ehrlich & Shami, 2010, p. 46). However, the methods employed in large-scale analyses such as those described below are not limited to social media data.

Research examining very many linguistic variables at once is often focused on regional variation, though objectives as been as diverse as looking at language change, style, and gender. Huang et al. (2016) studied 211 sets of lexical variables in order to examine regional variation throughout the US as expressed on Twitter. Similarly, though not including quite as many variables, Grieve et al. (2018) looked at 54 lexical items on Twitter that they judged to be innovations with the goal of mapping their geographic diffusion over time. With a very different interest, Pavalanathan and Eisenstein (2015) put Bell's (1984) audience design framework for style shifting to the test by analyzing 200 lexical variables on Twitter, treating each variable as a binary where the lexical item is either present or absent, which is not a typical approach to linguistic variables in variationist research where the norm is to analyze manifest tokens of variables, though Pavalanathan and Eisenstein's approach is not at all atypical for computational sociolinguistics.

The impact of gender on variation has also been analyzed on Twitter. Bamman et al. (2014) employed 10,000 lexical items to do so. One of their goals was to see which lexical items were associated with specific genders, though

they also used these items to train a classifier to identify the gender of Twitter users in the first place as this information is not part of users' profiles nor is it always clear how users identify from cursory glances. Their approach of training a classifier involved generating a LM, which is at the heart of the method proposed in the present study to compare language varieties holistically.

In some ways, these studies harken back to the early days of dialectology where the goal was to produce atlases that demonstrated the boundaries between regional varieties such as the Atlas linguistique de la France for French (Gilliéron & Edmont, 1902) or the Sprach-Atlas for German (Wenker, 2020). Technologies that are ever-present today but unimaginable in the 19th and early 20th centuries allow new insights to be gained from studies with similar objectives to those early ones that relied on questionnaires alone. Indeed, technological advances have always served as catalysts for paradigm changes in methods in science in general but even in sociolinguistics. As J. Milroy and Milroy (2012) once remarked, the development and availability of tape recorders is what made sociolinguistic research based on real speech a reality starting in the 1960s (p. 52).

1.5 Generalizing from n-gram language models

The aforementioned studies and took advantage of advances in NLP to analyze many more variables at once than was ever done or even possible before, making them much more holistic examinations of varieties than studies based in traditional methods. However, it is possible to go further. Importantly, all of these large-scale studies have looked at lexical variation alone despite the large number of linguistic variables used, leaving out the other linguistic levels. In the case of variation in written language, a truly holistic analysis would aim to capture morphological, syntactic, and even orthographic variation in addition to lexical variation.

An approach that would allow for comparing language varieties on all linguistic levels would involve using character n-gram and word n-gram LMs. n-gram LMs have been used successfully in NLP for the task of language identification and language variety identification, as discussed in section 1.3. Achieving high accuracy in distinguishing betweening two closely related language varieties suggests that similar methods can be used to quantify the distance between two already-identified varieties, as well. As LMs are simply probability distributions over some set of symbols, existing statistical metrics meant to quantify the distance between two distributions can be used. My research question is thus the following:

RQ: Are n-gram LMs useful for quantifying the linguistic distance between varieties in a more holistic way than was previously possible?

2   Methods

There are three important factors to consider when using LMs for comparing language varieties: which corpus/corpora to use, which LM type(s) to use, and which metric to use to quantify the distance between the LMs. These three factors will be considered in sections 2.1, 2.2, and 2.3 below, respectively.

2.1   Data

The primary interest of this study is to develop baseline results for the method, meaning that the method will be applied to language varieties that are *a priori* distinct enough that they are socially considered different languages as well as language varieties that are *a priori* identical for all intents and purposes. A natural choice for a corpus given this interest is the Europarl corpus (Koehn, 2009).

The Europarl corpus is a parallel corpus that includes speech from proceedings of the European Union dating back to 1996. A parallel corpus aligns text in one language with its translation in one or more other languages. In the case of the Europarl corpus, speakers are recorded in their own language and then translated into the other official languages of the European Union, which at the outset of the recording of these proceedings meant ten other languages.

The portions of the corpus used to construct LMs for *a priori* different languages were relatively simple to choose, but the portions used for identical varieties requires a couple assumptions. For different languages, the portions of the corpus considered to be English were compared to the portions considered to be French, meaning LMs for the English texts were compared to LMs for French texts. These two languages were chosen primarily because I am very comfortable speaking both myself, allowing for investigation of any anomalies if they may arise that would not be possible for languages that I do not know. The decision was not related to how different English and French are from each other compared to any other language pairs. Naturally, these languages have a relationship and will thus have some similarities in their models, such as the presence of interlingual homographs such as English ⟨coin⟩ 'coin' and French ⟨coin⟩ 'corner', but that which is important is that they are distinct enough overall that quantifying their distance would show very great distance. The greatest possible distance in the case of written language comparisons would be between languages that use different scripts entirely, but a baseline need not show the maximum and minimum distances but rather very large and very small distances that help situate comparisons between similar but different language varieties.

The portions of the corpus used to construct LMs for the *a priori* identical varieties came from English and French again. However, in this case, the English texts were split in half with LMs constructed for the first half to be compared to LMs constructed for the second half. The same was done for the French texts. The idea is that the first half of the texts for any given language should yield LMs that are fairly identical to those of the second half of texts from the same

language. There are of course different speakers represented in the corpus, and it is quite possible that those who spoke early in the corpus are not the same as those who spoke later and that these two groups of speakers use different language varieties. This is assumed to not be a great issue due to the corpus representing formal proceedings. All speakers are national representatives of their respective countries communicating in an exceptionally formal forum and so are assumed to converge on the same language norms in these texts.

2.2 Language models

For each portion of the Europarl corpus that was chosen for comparison to another corpus, an LM or LMs were required. I followed Duvenhage (2019) due to his success with applying his language identification system to the Discriminating between Similar Languages (DSL) portion of the 2017 VarDial workshop. Duvenhage achieved an overall accuracy of 98.70% when classifying texts as being written in language varieties that are similar, whereas the winner of the DSL task achieved an accuracy of 92.74% (p. 4). Part of what led to such high accuracy was training a combination of n-gram LMs that included word unigram, word bigram, character bigram, character 4-gram, and character 6-gram LMs. As such, the same LM types were used in the present study under the assumption that good sensitivity in language variety discrimination would translate into more accurate measures of distance between language varieties when comparing them.

These five LMs arguably capture four linguistic levels. Perhaps the least obvious linguistic level captured is syntax by way of the word bigram LM. For instance, one can imagine the following word bigrams making up part of an English LM:

- ⟨the record⟩

- ⟨the law⟩

- ⟨the consequence⟩

- ⟨the state⟩

While the LM would not include information about the lexical category for each word, it is still codifying the idea that a word like *the* is likely followed by a word that happens to be a noun or in other words NP → Det N. What is missed in such an LM, however, is dependencies that are even moderately long distance. If sentence 2 below was found in the training corpus, a word bigram LM would have no way of codifying the idea that the presence of ⟨Tony and Kim⟩ necessitate the use of ⟨their⟩ as opposed to possibilities such as ⟨his⟩ or ⟨its⟩. While it would be beneficial to have the LM deal with such long distance dependencies, it cannot be done with large word n-grams – an 8-gram in this case – because such LMs yield very many types with one token each.

2. ⟨Tony and Kim developed a plan for their stock exchanges.⟩

10

The other captured linguistic levels are perhaps more obvious. The word unigram LM captures the lexicon of the variety, and all of the character n-grams capture both the morphology and the orthography of the variety. For instance, a character 4-gram that may occur many times in an English corpus but not a French corpus is ⟨ing ⟩. The LM does not annotate this as a morpheme, but it does effectivey codify that the three characters ⟨ing⟩ often come before a space and at the end of a word, suggesting a suffix. Likewise, a character bigram such as ⟨té⟩ occurring many times in a French corpus is codifying the idea that acute accents occur in the orthography of this French variety, whereas the same character bigram would likely never be found in an English corpus.

One limitation that should be noted about capturing a variety's morphology in a LM is that drawing comparisons between the morphological systems of two different varieties afterward is dependent on how the sound system is being represented in writing. An extreme example would be comparing a character n-gram LM constructed from an English corpus with one constructed from a Mandarin corpus where it would be literally impossible to find that both varieties used a particular morpheme as the writing systems are in no way related. However, this is as it should be as it is also very unlikely that a particular morpheme is used in both English and Mandarin. A less extreme and more probable example would be in comparing closely related language varieties that may be written in different scripts regardless, such as Romanian possibly being written in Latin script in some communities and Cyrillic in others. A particular morpheme may indeed be exactly the same between two such varieties but the representations are completely different. For spoken language (i.e., in the sense of written transcripts of speech), a solution may be to use IPA transcriptions when constructing character n-gram LMs, but for written language, IPA transcriptions may not be appropriate. This limitation is not at issue with the current analysis, though, as both English and French use Latin script.

2.3 Analysis

As has been mentioned already, an n-gram LM is effectively a distribution over some set of symbols. Table 1 shows what this might look like for two different word unigram LMs, though it only presents a very small sample. The entirety of each LM would include considerably more n-grams. Indeed, the word unigram LM constructed in this study from the entire English corpus contains 620,866 unigrams, where the most frequent, ⟨the⟩, occurs 3,577,328 times. The difference between distributions such as these or their similarities can be quantified, which for the purposes here would represent the distance between two varieties. Two metrics were used in this study: KL divergence and cosine similarity.

KL divergence specifically quantifies the difference between two probability distributions. As such, the frequencies in LMs such as those in Table 1 must be converted into probabilities. As LMs are typically trained on very large corpora – the English portion of the Europarl corpus used here is made up of 52,562,008 words – the empirical probability is an appropriate estimate of the

Table 1: Hypothetical example of the frequencies of a small portion of two different LMs

Unigram	a	about	is	laws	phrase	words	this
English 1	41	7	28	–	2	1	12
English 2	49	10	25	2	1	–	17

true probability for each n-gram. The relative frequency of each n-gram was thus used in this study. The resulting values yielded by KL divergence range from zero and up where zero represents no difference and large values can be understood to represent great differences.[5] This range makes KL divergence fairly easy to interpret, though it perhaps makes comparing differences between different pairs of language varieties difficult as one is comparing values that could range from zero to infinity.

Cosine similarity, on the other hand, yields values ranging from -1 to 1, making it simpler to compare the cosine similarities obtained from comparing different pairs of language varieties. In this case, a value of 1 means the distributions are identical, and a value of -1 means the distributions are not similar in any way. However, interpreting these values on their own is not as straight forward as with KL divergence. Technically, each distribution is being treated as a very high dimensional vector space, which is not easily conceptualized.

Another advantage of cosine similarity is that, unlike KL divergence, counts of zero do not cause any problems. Zero counts are quite frequent in LMs, and particularly frequent when aligning two LMs from very different language varieties. For instance, the word unigram ⟨ai⟩, the 1st person singular habitual inflection of the verb meaning 'to' in French, is very frequent in an LM trained from a French corpus – it occurs $4{,}617$ times in the French LM used here – but naturally is unlikely to ever occur in an LM trained from an English corpus, as is the case here. To calculate KL divergence between distributions that contain zero counts, data smoothing must be used.

A very simply form of smoothing, and the form used here, is Laplace smoothing, which involves adding an imagined pseudocount of one to every n-gram in the distribution. In some cases, the actual counts may be zero simply because the sample corpus was not large enough to obtain any observations of the n-gram in question. These are ideal cases for Laplace smoothing. In other cases, such as having a zero count for ⟨ai⟩ in an English LM, Laplace smoothing is more dubious as the count is expected to remain zero no matter how large the corpus sample is. The alternative to Laplace smoothing then would be to remove all n-grams with zero counts from both distributions. However, this was not done here as these are exactly the n-grams that may distinguish these language varieties from each other the most.

Both KL divergence and cosine similarity were calculated for pairs of like

---

[5]Technically, KL divergence is how much information is lost when one probability distribution is used as an estimate for another.

n-gram LMs. For instance, character 4-gram LMs were only compared to other character 4-gram LMs and never to other LMs. Doing so would have amounted to comparing different linguistic levels between varieties and would have likely yielded very high differences even between similar varieties.

3 Results

Figure 1: 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
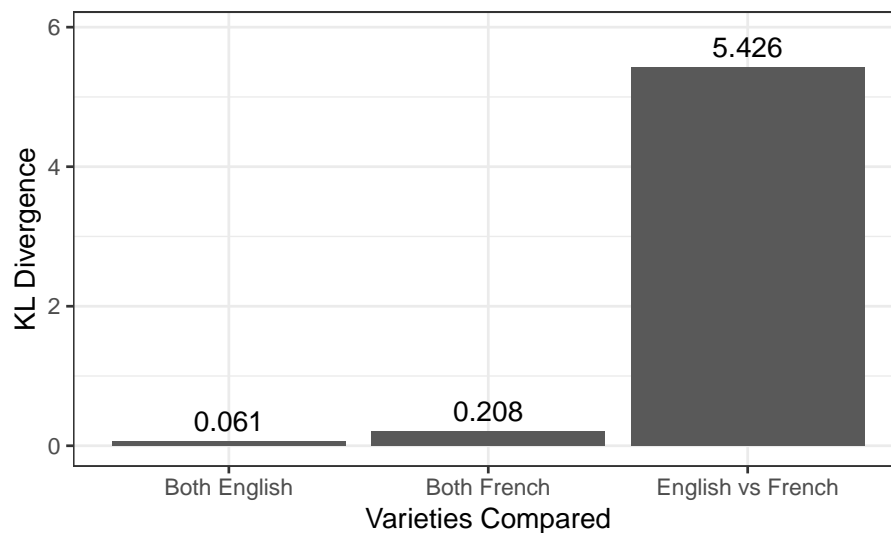
4 Discussion

5 Conclusion

References

Agha, A. (2003). The social life of cultural value. *Language & Communication, 23*(3–4), 231–273. https://doi.org/10.1016/S0271-5309(03)00012-0

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160. https://doi.org/10.1111/josl.12080

Bell, A. (1984). Language style as audience design. *Language in Society, 13*(2), 145–204. https://doi.org/10.1017/s004740450001037x

Figure 2: 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
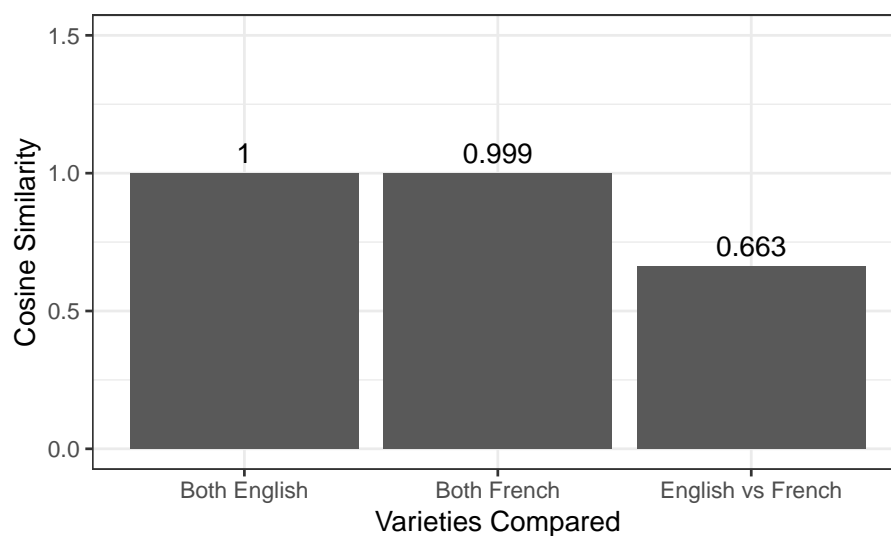

Figure 3: KL divergence values comparing individual LM types, each representing one linguistic level, for pairs of language varieties
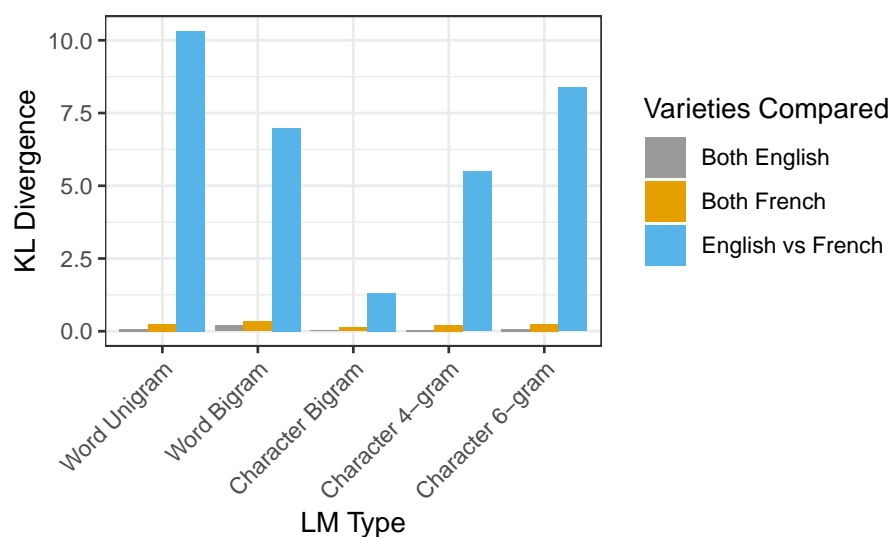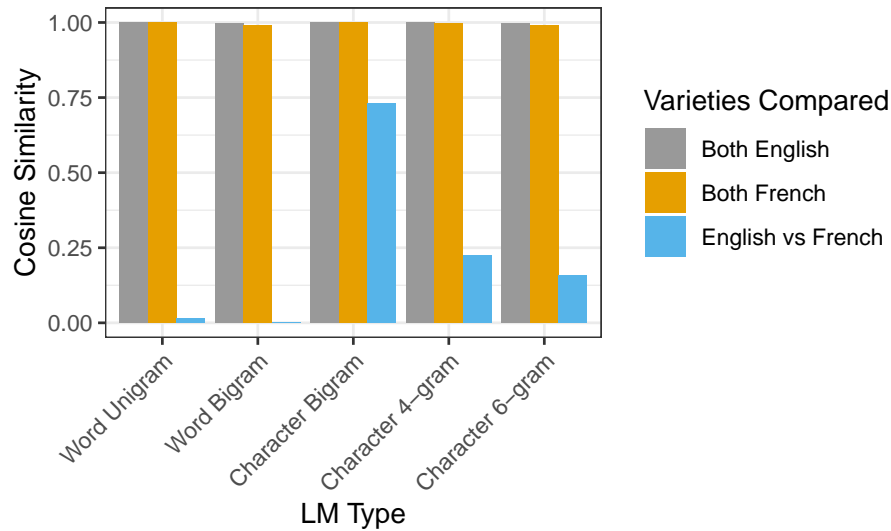
Figure 4: Cosine similarity values comparing individual LM types, each representing one linguistic level, for pairs of language varieties

Bouamor, H., Hassan, S., & Habash, N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 199–207.

Burnard, L. (2007). Reference Guide for the British National Corpus (XML Edition). Retrieved January 17, 2019, from http://www.natcorp.ox.ac.uk/docs/URG/

Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change*, *21*(1), 135–156. https://doi.org/10.1017/S0954394509000052

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon*, *5*(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

D'Arcy, A., & Young, T. M. (2012). Ethics and social media: Implications for sociolinguistics in the networked public1. *Journal of Sociolinguistics*, *16*(4), 532–546. https://doi.org/10.1111/j.1467-9841.2012.00543.x

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Delforge, A. M. (2012). 'Nobody wants to sound like a provinciano': The recession of unstressed vowel devoicing in the Spanish of Cusco, Perú. *Journal of Sociolinguistics*, *16*(3), 311–335. https://doi.org/10.1111/j.1467-9841.2012.00538.x

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley Girls and California vowels. *Journal of Sociolinguistics*, *19*(2), 241–256. https://doi.org/10.1111/josl.12115

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]*. Retrieved December 9, 2020, from http://arxiv.org/abs/1911.07555

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Blackwell Publishers, Inc.

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

Ehrlich, K., & Shami, N. (2010). Microblogging Inside and Outside the Workplace. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 42–49.

Eisenstein, J. (2013). Phonological Factors in Social Media Writing. *Proceedings of the Workshop on Language in Social Media*, 11–19.

Francis, W. N., & Kucera, H. (1964). Brown corpus.

Gilliéron, J., & Edmont, E. (1902). *Atlas linguistique de la France. Notice, servant a l'intelligence des cartes*. Paris H. Champion. Retrieved June 19, 2021, from http://archive.org/details/atlaslinguistnot00gilluoft

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

Guy, G. R. (2013). The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics*, *52*, 63–71. https://doi.org/10.1016/j.pragma.2012.12.019

Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters in Twitter: A Large Scale Study. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 518–521.

Horvath, B., & Sankoff, D. (1987). Delimiting the Sydney speech community. *Language in Society*, *16*(2), 179–204. https://doi.org/10.1017/s0047404500012252

Houston, A. C. (1985). *Continuity and Change in English Morphology: The Variable (ing)* (PhD). University of Pennsylvania. Philadelphia, PA.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, *59*, 244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Ilbury, C. (2020). "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, *24*(2), 245–264. https://doi.org/10.1111/josl.12366

Johnstone, B., Andrus, J., & Danielson, A. E. (2006). Mobility, Indexicality, and the Enregisterment of "Pittsburghese". *Journal of English Linguistics*, *34*(2), 77–104. https://doi.org/10.1177/0075424206290692

Jones, T. (2015). Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech*, *90*(4), 403–440. https://doi.org/10.1215/00031283-3442117

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

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Labov, W. (2006). *The Social Stratification of English in New York City* (2nd) [origdate = 1966]. Cambridge University Press.

Lee, H. (2019). 15 years of Google Books. Retrieved June 30, 2021, from https://blog.google/products/search/15-years-google-books/

Milroy, J., & Milroy, L. (2012). *Authority in Language: Investigating Standard English* (4th) [original-date: 1985]. Routledge.

Milroy, L. (1987). *Language and Social Networks* (2nd) [origdate = 1980]. Blackwell.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning–based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, *54*(3), 62:1–62:40. https://doi.org/10.1145/3439726

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational Sociolinguistics: A Survey [Publisher: MIT Press]. *Computational Linguistics*, *42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258

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

Payne, A. (1980). Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. In W. Labov (Ed.), *Locating Language in Time and Space*. Academic Press.

Podesva, R. J. (2011). Salience and the Social Meaning of Declarative Contours: Three Case Studies of Gay Professionals. *Journal of English Linguistics*, *39*(3), 233–264. https://doi.org/10.1177/0075424211405161

Prasad, R., Webber, B., Lee, A., & Joshi, A. (2019). Penn Discourse Treebank Version 3.0. *LDC2019T05*. https://doi.org/10.35111/QEBF-GK47

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.

Rosen, A. (2017). Tweeting Made Easier. Retrieved June 30, 2021, from https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier

Rottet, K. J. (1995). *Language shift and language death in the Cajun French-speaking communities of Terrebonne and Lafourche parishes, Louisiana* (PhD). University of Indiana. Bloomington, IN.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & Communication, 23*(3–4)*,* 193–229. https://doi.org/10.1016/S0271-5309(03)00013-2

Van Hofwegen, J., & Wolfram, W. (2010). Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics, 14*(4), 427–455. https://doi.org/10.1111/j.1467-9841.2010.00452.x

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* [Google-Books-ID: rMrtDwAAQBAJ]. Walter de Gruyter GmbH & Co KG.

Wojcik, S., & Hughes, A. (2019). *Sizing Up Twitter Users* (tech. rep.). Pew Research Center. Washington, DC.

Wolfram, W., Myrick, C., Forrest, J., & Fox, M. J. (2016). The Significance of Linguistic Variation in the Speeches of Rev. Dr. Martin Luther King Jr. *American Speech, 91*(3), 269–300. https://doi.org/10.1215/00031283-3701015

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., & Aepli, N. (2017). Findings of the VarDial Evaluation Campaign 2017. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*. https://doi.org/10.18653/v1/w17-1201

Zhang, Q. (2005). A Chinese yuppie in Beijing: Phonological Variation and the Construction of a New Professional Identity. *Language in Society, 34*(3), 431–466. https://doi.org/10.1017/s0047404505050153