Journal of French Language Studies (2022), **32**, 165–196
doi:10.1017/S0959269521000223

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Phonological variation on Twitter: Evidence from letter repetition in three French dialects[1]

Jeffrey Lamontagne[1],* and Gretchen McCulloch[2]

[1]Indiana University Bloomington and [2]Lingthusiasm
*Corresponding author. Email: jlamonta@iu.edu

(Received 15 February 2021; revised 21 September 2021; accepted 29 September 2021; first published online 09 February 2022)

Abstract

Writing on social media often departs from prescriptive norms through the use of non-standard words, spellings and punctuation. Amongst these traits is the repetition of letters (e.g. <ouiiiii> for oui 'yes'). In this study, we draw upon a corpus of over 65 million tweets from three dialects of French (Laurentian, Metropolitan and Midi) to test phonological motivations for the choice of repeated letter in a word with repetition. Using mixed-effects multinomial regression, we compare dialectal differences in whether repetition targets final consonants (silent or pronounced), word-final orthographic <e> corresponding to phonological schwa, and prosodically accented penults. We demonstrate that repetition covertly signals phonological properties. We conclude that prosody mediates morphological and phonological effects and that grapheme-to-phoneme correspondences vary between regions, thereby producing phonological patterns that writers likely did not intend to convey at the time of writing. We also propose that orthographic repetition on Twitter has two prosodic sources: the default pitch accent in French (shifted or not) and focus.

1. INTRODUCTION

Twitter constitutes a large corpus with spontaneous or semi-spontaneous writing, which allows for both formal and informal communication. The informality of the data source enables tweet authors to write using non-standard orthography or flouting other prescriptive conventions, thereby conveying group identities or aspects of spoken language (Wikström 2017). Studies on phonological variation

[1]We thank the audiences of *New Ways of Analyzing Variation* 46, the 2018 meeting of the *Canadian Linguistics Association* and *Les français d'ici* 2018 for comments on previous versions of this analysis, which is available in French through the proceedings of the last of these conferences (cited here as McCulloch & Lamontagne 2020). We further thank three anonymous reviewers for their comments as well as Amanda Dalola and Heather Goad. Any errors are our own. This work was supported in part by the *Social Sciences and Humanities Research Council of Canada* [fellowship number 766-2013-0862].

on (written) social media are relatively sparse, but work has shown that writers can consciously convey spoken variants through non-standard spelling (e.g. Tatman 2016) and that such spellings may show phonological conditioning similar to those found in speech (Einstein 2013). To our knowledge, existing work only involves cases where the variant in pronunciation is directly communicated through the spelling – and presumably consciously, for studies not examining errors. As such, there has not yet been a test of Tatman's (2016) proposal that phonological features represented on Twitter must be salient to the tweet author.

In the current study, we examine a novel case in that the phonological traits we are inferring from the non-standard spelling are not strictly those we believe the tweet author was highlighting through their choice of spelling. When authors repeat letters – for example, writing <merciiiiii> for *merci* 'thanks' – they are communicating that the word should be read with emphasis or heightened emotional affect that increases its duration (Schnoebelen 2012). In principle, the author could repeat any or all letters in the word or could consistently repeat letters in a specific position (for example, always repeating the last letter of the word or always repeating the rightmost vowel). This is not the case, however; as we will demonstrate, there are patterns with respect to the letters that undergo repetition, and these patterns can inform us about the author's phonology. Comparing three varieties (Laurentian French, Metropolitan French and Midi French) that have the same standardized spelling system allows us to focus on differences in pronunciation with less concern that prescriptive spelling norms are responsible for the differences in spelling, as would typically be the case when comparing distinct languages. We will show that phonological properties can be signalled indirectly through letter repetition choices, which we interpret as a covert side effect of the author's intent to communicate prosody. This suggests a revision to Tatman's (2016) proposal; non-salient features must predict or influence a salient feature (like prosody) in order to be represented on Twitter.

2. BACKGROUND

This study draws on work both on Twitter and on French. In section 2.1, we provide context into previous work examining sociolinguistic variation through non-standard spelling on Twitter and highlight how the style of writing on Twitter lends itself to such sociolinguistic use. In section 2.2, we offer an overview of phonological or phonetic studies using Twitter as their data source for French, which is the backdrop for this study. In section 2.3, we briefly describe the dialects under examination and their key phonological traits. Finally, we introduce three test cases (sections 2.4-2.6) that will allow us to show how writers act upon their linguistic intuitions in non-standard writing in ways they cannot when following standardized writing conventions. The test cases were selected to examine instances where the dialects are phonologically similar to establish a baseline (test case 1; section 2.4), one where dialects are clearly phonologically distinct to show potential differences between dialects (test case

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

2; section 2.5), and one where the differences between dialects are not known due to the patterning only being confirmed in one dialect (test case 3; section 2.6).

2.1. Sociolinguistic variation and Twitter

Twitter presents a medium for communication that is highly variable with respect to style (i.e. formality or register) due to its large user base with distinct aims (e.g. communication between friends, marketing or advertising for businesses, professional communication). This produces what Biber (1995) would describe as an inherently variable domain of communication because it is predisposed to a greater range of internal variation in style because of the competing or distinct uses of the medium. It can therefore reflect some level of conflict between rigid standardized writing norms and the desire for informal or playful communication, as has been observed on pre-existing computer-mediated platforms (e.g. Kotzur 2015; Thurlow 2003; Werry 1996; Wikström 2017). For many users, such communication thus comes closer than standard writing to embodying the "language of immediacy" in Koch and Oesterreicher's (2012) dichotomy between the language of immediacy (more dialog-like and reflecting private communication between interlocutors who are familiar with each other) and the language of distance (more complex or elaborate communication between non-intimate individuals in public contexts).

As Kotzur (2015) highlights, the terms long used for informal and non-standard writing suggest a speech-like quality – for instance, "netspeak", "textspeak", "typed talk". Previous work on Twitter reveals that authors may consciously use this type of writing to index sociolinguistic or political affiliations. We present two examples: the first illustrates the practice of gender identity, while the second demonstrates that individuals can consciously use informal communication to mark their affiliation to a speech group. Bamman, Eisenstein and Schnoebelen (2014) shows that individuals whose written communication does not correspond to an automated classification of gender (which incorporates factors like topic of communication) write for a social network that contains a significantly larger proportion of individuals of the opposite gender. Tatman (2015, 2016) demonstrates that individuals may use non-standard spelling to communication regional pronunciation variants like non-rhoticity (e.g. <beah> for "beer") and interdental stopping (e.g. <duh> for "the").

In the current study, we aim to discern regional pronunciation differences in contexts where the tweet author did not explicitly communicate the phonological or phonetic variant, and therefore may not have been intending to communicate this variation. The following section provides an overview of work that examines French pronunciation variants on Twitter.

2.2. Studying French pronunciation through Twitter

As previously noted, tweet authors can use non-standard spelling on Twitter to communicate variants in pronunciation. To our knowledge, work on French is thus far limited to two studies, and they find limited conditioning reflecting effects of spoken language. The first study was conducted by Dalola (2017), who

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

demonstrated that tweet authors use variant spellings on Twitter to convey the devoicing high vowels often undergo phrase-finally in speech, for example pronouncing *merci* /mɛrsi/ 'thanks' as [mɛrsi̥] and conveying this realization through spellings like <mercih>, <mercish>, <mercich> or <merciche>. Her findings show that gender differences and frequency effects found in speech are also present when analysing non-standard spelling variants on Twitter.

The second study was conducted by Law (2017), which targeted the ongoing backing and raising of /ɑ̃/, causing it to merge with or shift towards /ɔ̃/. This is reflected in non-standard spelling by replacing digraphs indicating /ɑ̃/ (e.g. <an> and <en>) with ones conveying /ɔ̃/ (<on>, <om>), both being used in the French lexicon. For example, *genre* 'like (discourse particle)' can be spelled as <geonre> to signal the vowel quality. Law observed that the patterning of this phonological variable is different in speech compared to non-standard writing, however; there was a frequency effect in the Twitter data not present in the spoken phenomenon, and phonological predictors found for speech (stress, adjacent labial segments) were not found in the Twitter data. He proposed that this was in part an effect of the medium: the variant spelling often has to be avoided in order to retain word recognizability.

The current study builds upon a previous pilot study on letter repetition in English (Lamontagne & McCulloch 2017), which tested whether phonetic tendencies for lengthening sounds in spoken English would be reflected in the likelihood of repeating letters associated with those sounds. Letter repetition in online media has been noted since at least 1996, when Werry proposed it compensated for the lack of prosody in writing. Schnoebelen (2012) later observed its presence on Twitter, referring to the phenomenon as 'affective lengthening'. While our 2017 study on English letter repetition found phonetic tendencies, the size of the effects paled in comparison to the overall preference to repeat the word-final letter. A large aspect of this, we suspect, results from the orthographic system of English: while repeating consonants would be possible in principle, repeating vowels is more restricted because the number of vowels is a component of the grapheme-to-phoneme correspondence (e.g. *bet* vs. *beet*).

In this study, we expand on the English letter repetition work by turning to French. Our goal is to evaluate the ability to test aspects of pronunciation in cases where the tweet author may not have intended to communicate those differences (i.e. communicating different information may have resulted in unintentionally communicating phonetic or phonological information).[2] This contrasts with existing work, in which the spelling itself conveys the target phonological information directly. If successful, the range of potential (socio) linguistic studies that can be conducted using large public datasets like Twitter is expanded considerably.

[2] See section 4.2 for a context in which authors unambiguously intended to communicate the target variant in pronunciation, which we consequently examine separately.

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

2.3. Varieties of French

We analyse three varieties of French as part of the current study: Laurentian, Metropolitan and Midi. First, **Laurentian French** is one of the two historical varieties of French in Canada, getting its name from the St. Lawrence River, where it originates (Côté 2012). Frequently called Canadian French, Quebecois or Quebec French, the variety extends from the province of Quebec westward. Second, **Metropolitan French** here refers to the variety spoken in Northern France, predominantly associated with Paris and often called Standard French (Berit Hansen 2012). Finally, **Midi French** is spoken in southern France, with strong historical influences from Occitan (Coquillon & Turcsan 2012). Midi French is particularly known for its high rate of schwa retention: for example, *mode* 'mode' is /mɔdə/ and is generally realized as two syllables with prominence consequently being assigned to the penult, whereas in other varieties the final schwa has typically been lost. As we will show, a comparison of these three varieties offers phonological differences that can be used as test cases to determine whether phonological factors influence the tweet author's choice of repeated letter.

2.4. Final consonants in French

In this study, we examine three test cases, comparing the three dialects with respect to each case. Our first test case is **final consonants**: we examine whether authors of each variety show the same patterns for repeating consonants based on whether those consonants are silent (e.g. the <t> in *salut* 'hello', pronounced [saly]) or not (e.g. the <t> in *zut* 'drat', pronounced [zyt]). We expect that the repetition patterns in these cases should be consistent across varieties, since there is relatively little variation across dialects based on whether final orthographic consonants are mapped onto consonants in the pronunciation. We predict that orthographic consonants mapping onto pronounced consonants will be repeated more often than those that do not map onto pronounced consonants, meaning that <saluttttttt> (silent <t>) is expected to be less common than <zutttttttt> (pronounced <t>).

Additionally, we will examine one group of phonologically active consonants in particular: lengthening consonants (e.g. Walker 1984). These consonants are the voiced phonological fricatives (/v z ʒ/) and the rhotic (which we transcribe as /r/ and [r] as variation in phonetic realization is common, but not relevant to the current study), which patterns phonologically with the voiced phonological fricatives with respect to its effect on preceding vowels. When these segments are in a word-final simple coda (or the coda is /vr/, comprised of the sole rising-sonority combination of lengthening consonants to appear word-finally), the preceding vowel is acoustically longer than when other consonants are in coda. In Laurentian French, this lengthening can also trigger diphthongization, with a word like *neige* /nɛʒ/ 'snow' optionally being pronounced similarly to [najʒ] or [nɛjʒ]. Based on phonological lengthening patterns, we expect that repetition will more often target the vowel preceding a lengthening consonant than the

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

lengthening consonant itself (e.g. *amer* /amɛr/ 'bitter' would be written <ameeeeeeeer> more often than <amerrrrrrr> across varieties).

2.5. Defective <e> (schwa) in French

The second test case is what is commonly referred to as **defective <e>**, also called *e caduc* 'vanishing e' or *schwa graphique* 'orthographic schwa' because it is typically spelled <e>, most varieties no longer pronounce it in the majority of contexts, and it was historically pronounced as a phonetic schwa (for clarity, we will use a schwa to identify the defective <e> when it is relevant for underlying forms; e.g. Morin 1978). The letter of interest for this study is unaccented word-final <e> in particular; <é> maps onto /e/, <è> and <ê> would map onto /ɛ(ː)/ word-finally but do not occur in this position, and <ë> is only word-final in a handful of words to indicate that a preceding vowel is pronounced as though the <e> were absent (e.g. *ambiguë* /ɑ̃bigy/ 'ambiguous.FEM' and *aiguë* /egy/ 'acute.FEM', compare *vague* /vag/ 'wave' without the diaeresis and *aigu* /egy/ 'acute.MASC' without <ë>).

Defective <e> is found in *code* 'code', for instance, which is pronounced [kɔd] rather than [kɔdə]. The notable exception to defective <e> not being associated with a pronounced vowel is in Midi French, where defective <e> is consistently pronounced with a vowel quality similar to [ø] – making *code* pronounced [kɔdø] –, unless there is an adjacent vowel (Eychenne 2014). (e.g. *vue* 'seen' is pronounced [vy] rather than [vyø] in all dialects). Unlike in English, where *code* /ko͡ʊd/ cannot be pronounced as a disyllabic word, it is possible in all varieties of French to produce [kɔdə] even when they normally do not pronounce defective <e> – but doing so is a stylistic choice that may convey hyper-formality or increased emotional affect (as postulated by Biers 2017, for example).

Based on the phonological patterns across dialects, we expect defective <e> to be repeated less often than full vowels in Laurentian and Metropolitan varieties, but in Midi French defective <e> should be repeated at higher rates compared to the other varieties unless a vowel is adjacent to the defective <e>. This pattern would reflect a dispreference found in speech for lengthening lexical schwas, as well as being consistent with literature on French suggesting that schwa is phonologically deficient (e.g. it is incapable of hosting stress).

2.6. Prosodic conditioning

The third and final test case is **prosodic variation** in an effort to explain that phonological features may be probabilistically reflected through conscious decisions without being salient at the time of writing, and tests a context in which the literature provides direct insight into only one dialect under examination (here Laurentian French). This question arises from preliminary analyses of the current Twitter data (McCulloch & Lamontagne 2020) in which we found that the <ai> orthographic sequence in the last syllable (e.g. *irai* /ire/ 'go.IND. FUT.1SG', *irais* /irɛ/ 'go.COND. PRES.1SG') was associated with the penultimate syllable being targeted by repetition more often when the sequence

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

could be associated with /ɛ/ than when it could be associated with /e/ (namely in Laurentian French, where the phonemic contrast is robust; e.g. Côté 2012). This result was hypothesized to be a consequence of prominence shifts to the penult being more common when the final syllable is light (for weight effects on prominence Laurentian French, see Lamontagne et al. 2017; Lamontagne 2020; for observations of prominence shifts in other dialects, see e.g. Carton et al. 1983; and Goldman & Simon 2007). However, cross-dialectal patterns of phonetic duration and of phonological processes (e.g. diphthongization) do not suggest that an open final syllable containing /e/ is phonologically heavy. The phonemic moraic differences between heavy mid-high and light mid-low vowels is therefore only expected to be reflected through vowel quality, not other cues to weight in the surface form, and is consequently surprising. As such, either abstract weight influences writing even when weight is neutralized in speech or another factor that correlates with vowel quality is responsible for the result.

In our third test case, we investigate the question of prosody in a context where prominence shift rates are expected to be far more extreme (and not strictly underlying): open and closed final syllables (Lamontagne et al. 2017; Lamontagne 2020). Penult rhymes are longer when the final syllable is open, but final-syllable rhymes are longer when the final syllable is closed (controlling for additional segment durations). We therefore expect letter repetition to affect (orthographic) vowels in a non-final syllable more often when the final syllable is open than when it is closed. In other words, *fini* 'finished' will be more likely to be written as <fiiiiini> than *finir* as <fiiiiiiinir>.

We consider words ending in a nasal vowel (e.g. *enfant* /ɑ̃fɑ̃/ 'child') separately from words ending in a consonant or an oral vowel. This distinction in final open syllables arises because rhymes ending in nasal vowels appear to have phonological patterning intermediate between open and closed rhymes with respect to prominence assignment (Lamontagne et al. 2017; Lamontagne 2020) and because nasal vowels can optionally undergo fronting and diphthongization in Laurentian French in open syllables (Dumas 1974). Both of these patterns suggest that nasal vowels variably conserve their weight in final open syllables (unlike oral vowels), while all closed syllables pattern together independent of vowel nasalization. We consequently expect words ending in a nasal vowel to pattern as intermediate between words ending in a consonant and words ending in a vowel.

2.7. Summary

In summation, we have three test cases to examine whether a variety's phonological system influences its speakers' non-standard spellings in casual written media even when conveying those phonological traits is not expected to be a goal of the speaker. We expect that the likelihood of letter repetition will reflect traits found in the speech of native speakers from the area from which a tweet originates. We additionally test the hypothesis that prosody is what enables a phonological feature to be represented on Twitter in contexts where we suspect the author did not intend to convey that phonological feature directly. In the next section, we

Table 1. The collection regions for each variety

Variety	Latitude	Longitude	Radius
Laurentian	48.48658	−75.171	530 km
Metropolitan	48.46819	1.59806	220 km
Midi	44.47461	3.27347	250 km

will describe how the Twitter corpus was collected, as well as how the data were extracted and then analysed.

3. METHODS

This study first and foremost involved the creation of a corpus of tweets, which needed to be sufficiently large so that sufficient tokens of a non-standard orthographic pattern (here letter repetition) would be present to allow for statistical analysis. In section 3.1, we describe the corpus and its collection, as well as the methods used to extract tokens of letter repetition from that corpus. In section 3.2, we discuss the statistical analysis performed in order to compare the three varieties with respect to our three test cases (final consonants, defective <e> and prominence shift).

3.1. Corpus and extraction

The corpus of French tweets was collected using the Twitter Application Program Interface (API) through the twitteR package (version 1.1.9; Gentry 2015), with the tokens analysed as part of this study having been collected between 10 January and 26 April 2017. Latitudes, longitudes and radii for tweet collection were selected to have a maximal area within the target variety's area, focusing on larger population centres. The resulting values are presented in Table 1 and the approximate dialect areas are illustrated in Figure 1, with Laurentian French covering southern Quebec and north-eastern Ontario in Canada, Metropolitan French targeting northern France, and Midi French targeting southern France (in the latter cases avoiding the Franco-Provençal region in eastern France, where a different variety is spoken).

In total, over 65 million French tweets were collected during the collection period. We excluded tweets that contained urls as well as retweets (following Wang et al. 2012, these often are duplicates, corporate tweets or automated tweets), then extracted the words where at least one letter was repeated at least twice after the initial instance (our criterion for determining letter repetition and minimizing the inclusion of simple typographical errors) using the regular expression "\b(\w*(\w)\2{2,}\w*)\b" to search for words containing at least three consecutive instances of a letter. These words were frequently, but not exclusively, at the end of their source tweet. Finally, we compared the words extracted to those in the Lexique lexicon (New et al. 2001) lexicon of French to ensure that only French words were included in our corpus of repetition, with

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Table 2. The total number of Tweets and of tokens (words with repeated letters) by variety

Variety	Number of Tweets	Number of Tokens	Tokens per 1000 Tweets
Laurentian	38,918,898	17,183	0.4415
Metropolitan	12,410,003	28,079	2.2263
Midi	13,949,691	10,393	0.7450
Total	65,278,592	55,655	0.8526

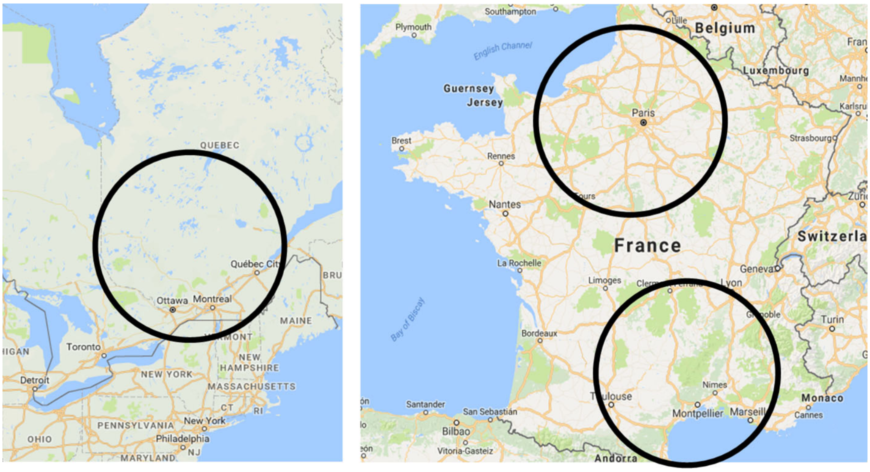

Figure 1. The approximate data collection regions for Laurentian French (left), Metropolitan French (upper right) and Midi French (lower right).

manual verification of words not in the Lexique. This last step was particularly necessary to ensure that borrowings were not included, since they may not pattern the same way as native words given the repetition patterns in their source language could be different, and to exclude acronyms or alphabetic abbreviations (e.g. *mdr* for *mort de rire*, literally 'deceased from laughing', which is a French equivalent of 'laughing out loud'). The Lexique transcriptions also served as the basis for phonological transcriptions of words, with manual verification of relevant transcription elements (e.g. the presence of a word-final consonant in pronunciation).

The resulting corpus of repetition included 55,655 tokens. Table 2 shows the number of tweets for each region, as well as the number of repetition tokens included and the frequency of repetition. Once the corpus of repetition tokens was generated, we could code the tokens for our test cases and perform the statistical analysis.

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

3.2. Statistical analysis

In order to ensure that apparent differences in repetition patterns are robust, we ran mixed-effects multinomial regressions with random intercepts for words using the MCMCglmm package (Hadfield 2010). This allowed us to control for words having considerably different frequencies in our corpus. Given that MCMCglmm implements Bayesian regression, we will interpret p-values indirectly; instead, we focus on our confidence that a conclusion is correct based on model outputs. Each model is based on 10,000 iterations of 1,000 samples, each based on a thinning of 10, after a burn-in of 30 iterations. The models predicted which letter would be repeated (but not how many times it would be repeated) given that the word includes repetition. For instance, if the word were *merci* 'thanks', the model would try to determine whether the final letter was repeated (e.g. <merciiiiii>), a non-final letter was repeated (e.g. <meeeeerci>) or whether non-final and final letters were repeated (e.g. <meeeeeerrrrccccciiiii>). We found during exploratory data analysis that having these three categories was sufficient overall to capture the patterns of interest because repetition occurs at the right edge of the word, but we will note relevant patterns where appropriate. All models are provided in the Appendix to preserve the flow of the text, where plots of the data will serve as the main illustration for the results we describe.

We computed separate models for each test case, as the parameters that were relevant to each one differed, but in all cases the predictors were recentered and rescaled by two standard deviations. Regardless of the test case, source dialect was included as a factor for main effects and for interactions, with Laurentian French as the model's baseline value both because of alphabetical ordering of factor levels and because the third test case (where preliminary work suggested Laurentian French differed) was of particular interest for direct comparisons. First, in the model testing **final consonants**, we examined the data where a final orthographic consonant was present and included as a predictor whether that final consonant mapped onto a consonant in pronunciation, as well as a predictor for the region and an interaction effect between the two predictors. We predict that final consonants will be repeated more often when they are associated with pronounced consonants than when they do not map onto a consonant in pronunciation. For example, we predict *zut* /zyt/ 'drats' to be spelled <zuttttt> more often than *salut* /saly/ 'hello' would be spelled <saluttttttt>.

We additionally created a model to test for effects specific to lengthening consonants (/v z ʒ r vr/), which are known to lengthen the preceding vowel in French speech and are therefore predicted to be associated with repetition of the previous vowel rather than of the lengthening consonant itself. For example, *mer* /mɛr/ 'sea' is predicted to be spelled <meeeeeeeer> more often than *mec* /mɛk/ 'man (slang)' is predicted to be spelled <meeeeeeeeeec> (and also more often than *aimer* /ɛme/ 'to love' is expected to be spelled <aimeeeeeeeeer>). For all final consonant cases, we predict little or no difference across varieties because the varieties all show similar patterns in production for whether an orthographic consonant is mapped onto a pronounced consonant or not.

Second, in our model testing **defective <e>**, we examined data that ended in a full final vowel (e.g. /e/ or /u/) or in a defective <e>. We included as the predictor of

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

whether the vowel was a full vowel or a defective <e>, whether the defective <e> followed a consonant or a vowel, as well as a predictor for the region and all interactions. Based on the differing phonological statuses of schwas across varieties, we predict that defective <e> will be repeated less often than other vowels and that Laurentian French and Metropolitan French will pattern together, but that Midi French will show an additional effect whereby defective <e> is less likely to be repeated when they follow another vowel. For example, we predict that Midi French authors are more likely to repeat the final <e> in *aime* /ɛm(ə)/ 'love.IND.PRES.3SG' than they are in *aimée* /ɛme/ 'loved.FEM', whereas Laurentian and Metropolitan authors are not predicted to show this difference. Additionally, the final <e> in *aime* is predicted to be repeated less often than the final <a> in *aima* /ɛma/ 'love.PST. PF.3SG' in all varieties because the defective <e> is phonologically defective in all dialects.

Finally, one model tests **prosodic conditioning**. This model includes only words of at least two syllables (prominence shifts are otherwise not possible) and classifies words based on whether their phonemic representation ends in an oral vowel (prominence shift most likely), in a coda (prominence shift least likely), or in a nasal vowel (intermediate rates of prominence shift). We predict that letter repetition will more often target the non-final syllable when prominence shifts are more likely because rhymes are significantly longer in the pronunciation of those words.

Overall, we predict that the phonological properties of each variety will significantly influence the repetition patterns produced by tweet authors of that variety and that these properties are mediated by prosody. If this is the case, casual writing like in many tweets could offer evidence for phonological variation in speech, even when tweet authors are not consciously and intentionally conveying the phonological trait under examination.

4. RESULTS

We find that, across our 55 000 tokens, final letters are often the target of repetition and that vowels are favoured for repetition over consonants. In all plots, three cases are distinguished: (a) *only* the non-final letter is repeated, like in <aaaaaaami>; (b) the final letter *and* at least one other letter is repeated, like in <aaaaaamiiiiiii>; and (c) *only* the final letter is repeated, like in <amiiiiiiii>. As shown in Figure 2, in all varieties, over 50% of tokens involve a final letter being repeated, of which about 13.4% of cases have the final letter *and* a non-final letter being repeated (e.g. <booonnnnnneeeeee> for *bonne* 'good.fem') and the rest have *only* the last letter repeating (e.g. <amiiiiii> for *ami* 'friend.MASC'). This proportion is highest in Laurentian French, where the rate of final repetition reaches about 67.7% and where about 15.8% of tokens involve repetition of both final and non-final letters.

Finality is only a preliminary consideration, and ignores the possible role of phonology. Table 3 presents the number of tokens with repetition for each combination of repeated letter type and of repeated letter position in the word. We observe that vowels are far more likely to repeat even though they are not

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

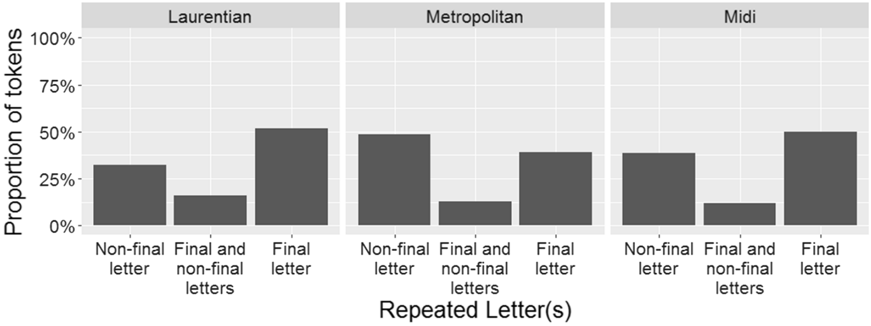

Figure 2. The repetition patterns by region and by target letter location.

the last letter of the word. This asymmetry is illustrated in Figure 3, vowels are repeated more often than consonants in all varieties, though the difference made by letter type is smallest in Laurentian French. We will show that this apparent difference in repetition preferences is largely motivated not by having different 'repetition grammars', but rather by applying similar or identical 'repetition grammars' to different phonological systems.

To demonstrate the effects of phonological systems on non-standard writing, we will analyse our test cases. In section 4.1 we examine the extent to which orthographic consonants mapping onto pronounced consonants influences their likelihood to be repeated. As the mapping onto pronunciation of final consonants is mostly consistent across varieties, we predicted that the varieties would show little or no difference with respect to consonant repetition. In section 4.2, we examine defective <e>, for which we predicted that Midi French would show distinct repetition patterns compared to the other two varieties as a result of their different phonological treatment of schwas in speech. Finally, in section 4.3 we directly test whether weight effects on prominence shifts in speech are also found for letter repetition.

4.1. Final consonants (Appendices A–B)

We first examine the behaviour of final consonants using the subsample of 27,309 tokens in which a final consonant is present. As illustrated in Figure 4, orthographic consonants that are associated with pronounced consonants (e.g. *zut* /zyt/ 'drat') are more likely to be the only repeated letter than orthographic consonants that do not map onto pronounced consonants (e.g. *coup* /ku/ 'hit' or *met* /mɛ/ 'puts') (for final position across contexts, E(·|x)=-1.3763, p<0.001). This means that <zuttttttt> is a more likely spelling than <coupppppp> is, and we additionally see that <couppppp> is an incredibly uncommon spelling as normally a vowel would be targeted instead of the silent consonant.

When we examine the results for each variety separately, we observe that they are fairly consistent across variety. As Figure 5 illustrates, there is no significant difference in repetition patterns when the final orthographic consonant is silent (a marginal effect is present, but will be discussed in Section 4.3). However,

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Table 3. The letter targeted by repetition according to position and type

	Final	Non-final	Final and non-final	Total
Consonant	12,681	1,113	2,357	16,151
Vowel	12,335	21,886	3,482	37,703
Consonant and vowel	0	158	1,643	1,801
Total	25,016	23,157	7,482	55,655

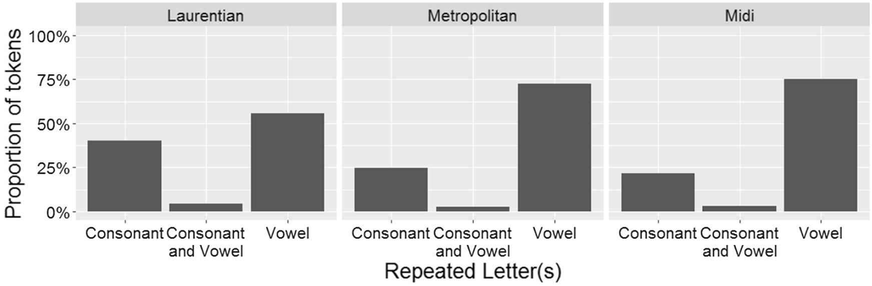

Figure 3. The repetition patterns by region and by target letter type.

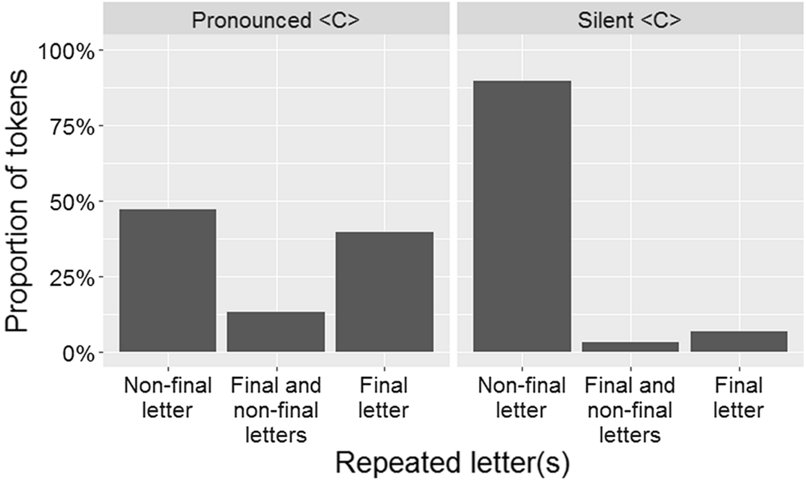

Figure 4. The target of letter repetition in words ending in orthographic consonants.

Laurentian French is distinct from the other two varieties when the consonant is pronounced (left panel); in these cases, Laurentian authors are more likely to repeat only the last letter (e.g. <zutttttttt>), whereas the two European varieties' authors are more likely to repeat only a non-final letter (e.g. <zuuuuuuuuut>;

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

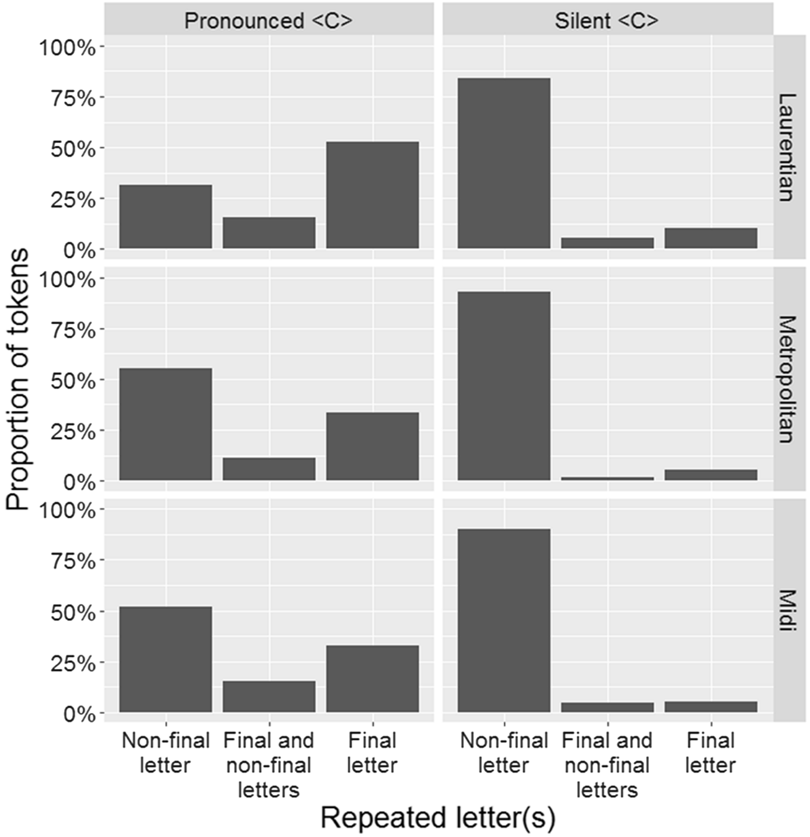

Figure 5. The target of letter repetition in words ending in orthographic consonants according to the variety.

for Midi French, E(·|x)=-0.4365, p<0.001; for Metropolitan French, E(·|x)=-0.1275, p=0.038) which data exploration shows is nearly categorically an orthographic vowel.

While a significant regional difference is found, data inspection suggests that the apparent size of the effect in Figure 5 is partly driven by words that allow a non-standard pronunciation with a final consonant, like *tout* ‘entirely’, which can be realized as [tu] or as [tʊt] in Laurentian French phrase-finally and before a consonant, and *lit* ‘night’, which can be realized as [li] or [lɪt] depending on the region within the Laurentian French dialect area.[3] As such, repetition

[3] Across varieties, *tout* can condition the pronunciation of a [t] in the onset of the following word if that word begins with a vowel (e.g. Côté 2010). This phenomenon, known as liaison, is distinct from the case of [tʊt] because liaison consonants do not consistently pattern like codas: liaison consonants do not trigger lengthening across varieties (e.g. the /i/ in *dis-en* ‘tell about (something)’ is not realized as long) and liaison consonants do not trigger closed-syllable laxing in Laurentian French (e.g. *petit* [p(ə)t͜si] ‘little.MASC’, *petite* [p(ə)t͜ssɪt] ‘little.FEM’, but *petit ami* [p(ə)t͜ssitami] ‘little friend.MASC’). *Lit* ‘bed’ does not condition liaison in [t].

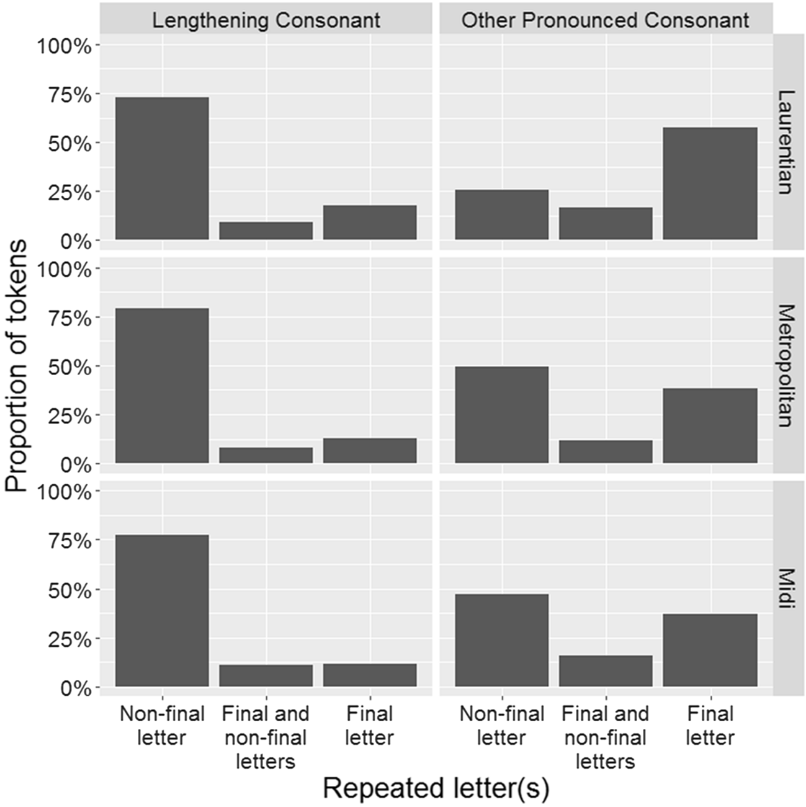

Figure 6. The target of letter repetition in words where the final orthographic consonant is pronounced according to the type of consonant pronounced.

patterns are suggestive – but not directly indicative – of variable lexical items across varieties.

There is an additional effect at play: the phonological properties of the final consonant significantly influence the repetition pattern. As shown in Figure 6, which depicts 5,732 tokens ending in a pronounced consonant, there is no significant difference in repetition patterns across varieties when the final consonant forms a lengthening sequence (/v z ʒ r vr/), in which case repeating the preceding vowel is preferred (matching the lengthening of the vowel found in speech; E(·|x)=-0.7625, p=0.026). Instead, the difference in repetition likelihood emerges when the final consonant that is pronounced is not a lengthening consonant, in which case Laurentian French authors are more likely to repeat only that final consonant compared to the European varieties' authors (for Midi French, E(·|x)=0.5360, p=0.094; for Metropolitan French, E(·|x)= 0.4988, p=0.042). This means that in all varieties <amouuuuuur> is favoured over <amourrrrrr> for *amour* 'love', which ends in a lengthening consonant,

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

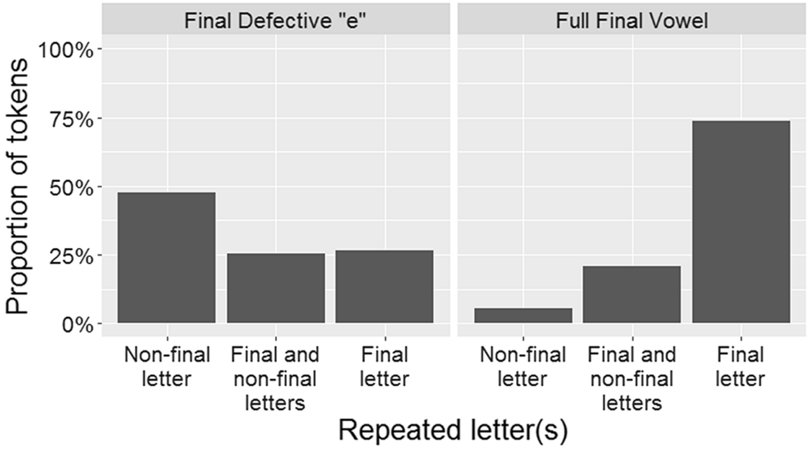

Figure 7. The rate of letter repetition in vowel-final words according to the type of final vowel.

but in Laurentian French <zutttttttt> is preferred over <zuuuuuut> compared to the opposite pattern being preferred in the European varieties.

Thus far, it appears that the two European varieties may pattern identically, but that Laurentian French may follow a slightly different 'grammar' of repetition with respect to non-lengthening final consonants. In the next section, we will examine the case that is expected to distinguish Midi French from the other two varieties, and we will determine whether this separation of European varieties from Laurentian French is still present.

4.2. Defective <e> (Appendix C)

In most varieties, defective <e> is not typically mapped onto pronounced vowels in French. For example, *robe* 'dress' is pronounced [rɔb] rather than [rɔbə]. In Midi French, however, defective <e> is nearly categorically pronounced as long as no vowel is adjacent. In the case of *robe*, this means the pronunciation will be similar to [rɔbø], while a word like *robée* 'wrapped in a robe (FEM)' is pronounced [rɔbe] (like in the other dialects) rather than [rɔbeø]. As a result, we predict that the likelihood of repeated defective <e> will be sensitive to the presence of an adjacent vowel in Midi French, but not in the other varieties.

We investigate this behaviour using a subsample of 39,851 tokens ending in a vowel (including defective <e>). As illustrated in Figure 7, there is a significant overall effect whereby final orthographic vowels almost always are the sole repeated letters when the vowel is not a defective <e> ($E(\cdot|x)=0.9493$, $p=0.016$). For example, a word like *robé* /rɔbe/ 'wrapped in a robe', the typical repetition pattern is <robéééééé>. However, in the case of final defective <e>, authors will target the last non-schwa vowel for repetition about half of the time, such that a word like *robe* would be spelled either <rooooooobeeeeeee> or <roooooobbbbbbeeeeeee> (both patterns occur frequently in the data). This represents the highest rate of repeating *both* the final letter and another letter in all cases examined.

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

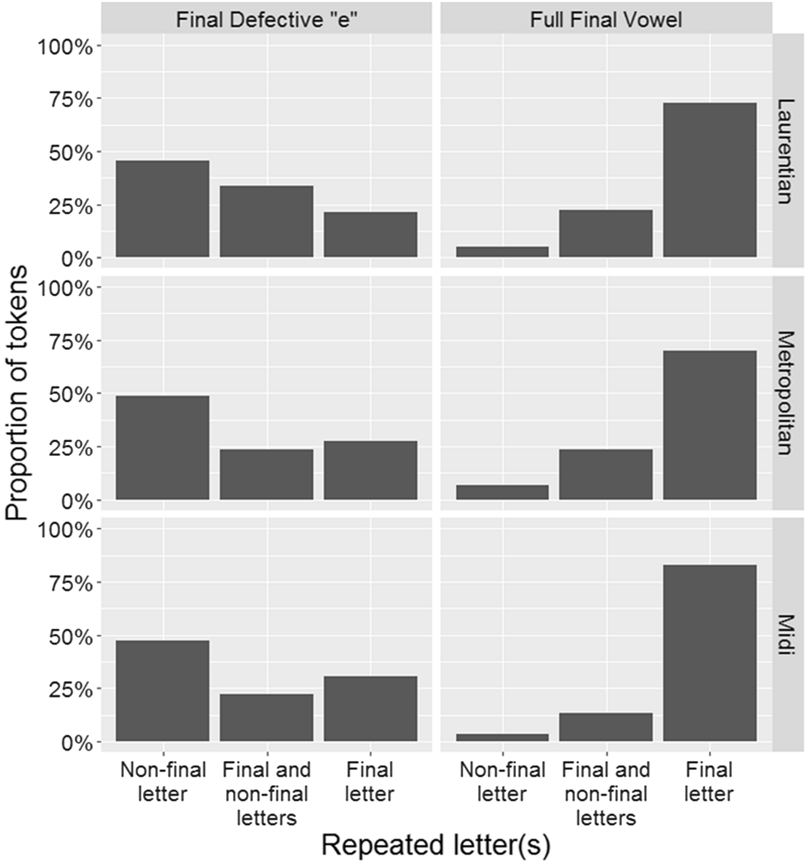

Figure 8. The rate of letter repetition by region in vowel-final words according to the type of final vowel.

When we examine individual varieties' patterns, as in Figure 8, it first appears that the primary difference is in Laurentian French once again, with a significant preference to repeat the defective <e> *and* the rightmost non-schwa vowel over only repeating the final defective <e> in Laurentian French (E(·|x)=1.2937, p<0.001; the posterior mean decreases significantly for Midi French and marginally for Metropolitan French). A precursory examination of the tokens suggests two explanations. The first explanation is that lengthening consonants between the defective <e> and the last full vowel increase the probability of repeating the last full vowel, with words of this shape not being equally frequent across varieties. The second explanation is that certain words have conventionalized preferences for repetition and those words also have different frequencies across varieties, similar to the effects captured in Usage-Based Phonology (Bybee 1999) or Exemplar Theory (Johnson 1997; Pierrehumbert 2003, 2016). Especially noteworthy for these cases are words like *neige* /nɛʒ/ 'snow', which was far more frequent in the Laurentian data and has an orthographically medial lengthening consonant, and *frette* /frɛt/ 'cold', a dialectal

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

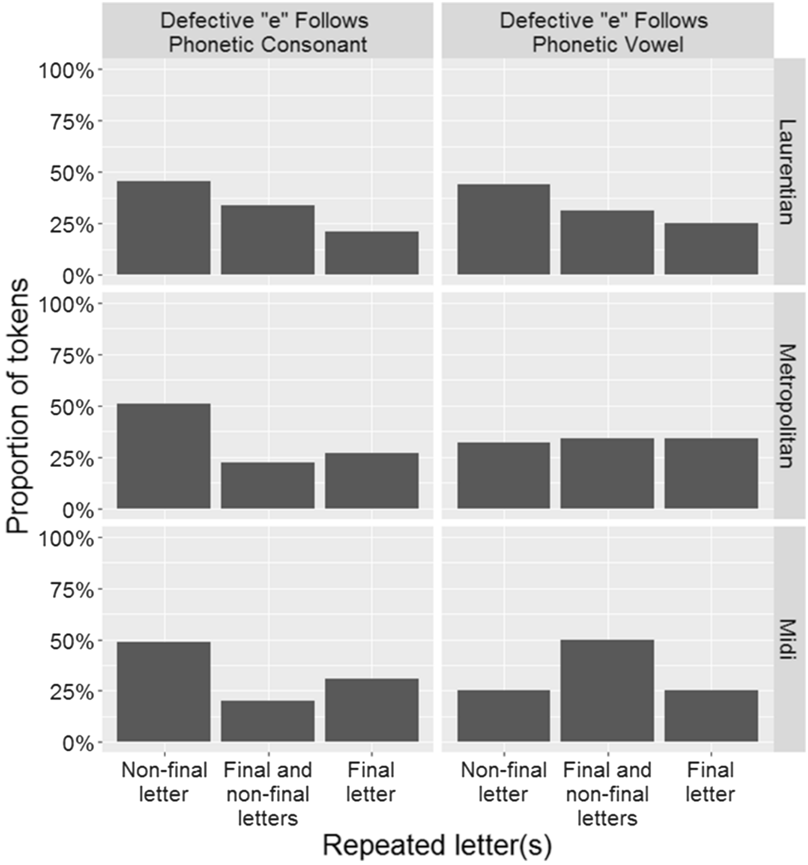

Figure 9. The rate of defective <e> repetition by region according to the preceding letter.

term in Laurentian French derived from *froid* 'cold' that had a strong preference for the first <e> to be repeated.

Finally, we turn to the effects of previous vowels and consonants on the likelihood for defective <e> to be repeated using the subsample of 1,437 tokens with final defective <e>. As shown in Figure 9, the only dialect where there is a near-significant dialectal difference for whether the defective <e> follows a vowel or a consonant in pronunciation is in Midi French ($E(\cdot|x)=3.8657$, $p=0.056$).[4] In these cases, authors of Midi French are more likely to target the

[4] Visual inspection of Figure 8 suggests that there could be an effect of the phonological context in Metropolitan French as well, albeit with a far smaller effect size than in Midi French. However, the associated interaction is not significant in our statistical models, which suggests that the effect may be driven by conventionalized preferences for individual words in addition to a preference to target the last full vowel directly when a consonant intervenes. We leave this for future work. If the effect is significant, it may reflect the role of defective <e> as a feminine morpheme marker or its potential to be analysed as the final member of a digraph or trigraph. We will return to the latter possibility in our discussion of the results for final consonants.

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

last full vowel and also repeat the defective <e>, meaning that for a word like *folie* 'madness', which is pronounced [fɔli] or [foli] in all dialects examined, spellings like <foliiiiieeeeee> are more common than ones like <foliiiiiiiie> or <folieeeeeee>.

Overall, we find evidence that full vowels are more likely to be the sole targets of repetition and that the phonological context of the schwa significantly influences its likelihood to be repeated in Midi French, where the schwa has a distinct phonological status. A case of a final defective <e> that was excluded from the models described in this section is of particular interest: in about 18% of cases, authors spelled the defective <e> as <euh> rather than <e>, as is typical in French orthography. This spelling, used to convey filled pauses in plays, for example, indicates that a defective <e> is intended to be pronounced – and at least one letter in the <euh> spelling was repeated in all cases where defective <e> is spelled <euh> and at least one letter undergoes repetition. For example, *bonne* 'good.FEM' would typically be spelled similarly to <bonneuhhhhhh>, <bonneeeuuuuuhhhhh> or <bonneuuuuuhhhhhhh>. This suggests that the schwa patterns more like a full vowel when the defective <e> is spelled in a way that reflects it being pronounced.

4.3. Prosodic conditioning (Appendix D)

Recall that earlier work (McCulloch & Lamontagne 2018, 2020) argued for the existence of prosodic conditioning of letter repetition whereby weight effects (that can trigger prominence shift and therefore penultimate lengthening in speech) may explain repetition in non-final syllables (where repetition is less common overall). In Laurentian French and likely in other varieties, closed final syllables are most likely to retain prominence, while the penult probabilistically attracts prominence away from open final syllables (Lamontagne et al. 2017; Lamontagne 2020). Figure 10 illustrates the syllable that is targeted by letter repetition based on 23,095 tokens of at least two syllables. Our results are initially surprising: for all varieties the lowest rate of repetition in non-final syllables is found for words with a final open syllable (for non-final generally, $E(\cdot|x)=-5.9258$, $p<0.001$; for interaction with final oral vowels, $E(\cdot|x)=-0.8379$, $p=0.018$). Furthermore, Laurentian French authors are less likely to produce repetition in non-final syllables as a baseline (see significantly higher rates of repetition in non-final and both non-final and final syllables for Metropolitan French as well as significantly higher rates for non-final in Midi French coupled with a trend for higher rates in both syllables in Midi French). Finally Midi French authors repeat letters in non-final syllables less frequently when the final syllable contains a (non-schwa) vowel ($E(\cdot|x)=-0.8338$, $p=0.032$), likely reflecting the importance of the contrast between schwa and other vowels in this variety.

Further data exploration elucidates the motivations for the results that run counter to our expectations. Regarding the case of nasal vowels, the suffix *–ment* /-mɑ̃/ 'ADV' is present in 9.7% of the 3617 tokens of words ending in a nasal vowel. We observe that over 56% of repetition tokens that are adverbs with this suffix exhibit repetition in non-final syllables, e.g. <vraiiiment> for *vraiment* 'very' or <laaaargement> for *largement* 'largely'. Similarly, only 16% of

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

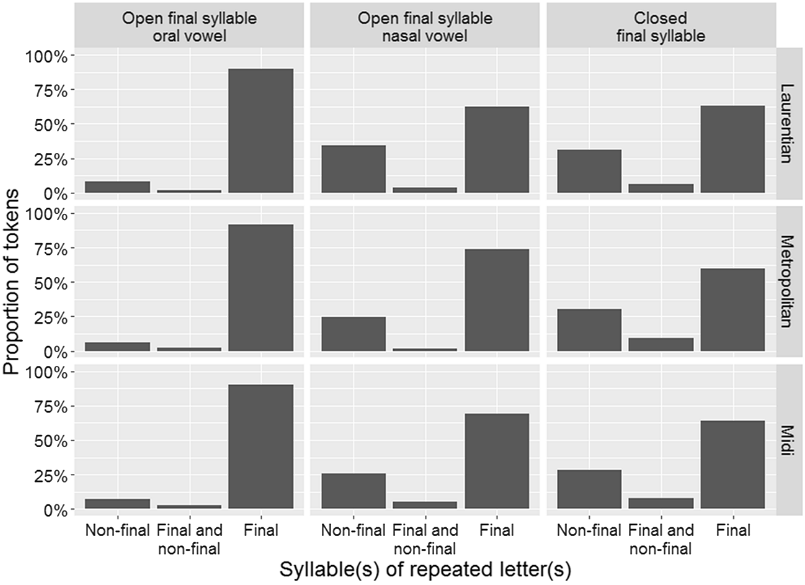

Figure 10. The rate of repetition by syllable according to the variety and the final syllable's profile.

adjectives in the repetition subcorpus end in an oral vowel. Turning to words ending in oral vowels, verbs are highly represented because of asymmetries in the French verbal inflection. While verbs may end in a consonant, this is typically the base-final consonant (e.g. /ɛm/ *aime* 'like.1sg') rather than a consonant in the inflexional suffix. Regarding consonant-final verbal affixes, in the first group of verbs we find only six in the full paradigm (three plurals in the simple past, and three forms in the subjunctive imperfect). The second group, a less frequent verb group, adds the infinitive (e.g. *finir*), the third person plural of the indicative present (e.g. *finissent*), and four forms in the subjunctive present (e.g. *finissent*). The result is that our vowel-final category has considerably more instances of overt inflexional affixes (for further discussion focusing on speech data, see Lamontagne 2020). Derivational suffixes, on the other hand, are often closed (e.g. *chanteur* 'singer.MASC', chanteuse 'singer.FEM'), with prominence optionally being assigned to the base-final syllable instead of or in addition to the final syllable in Laurentian French (Lamontagne et al. 2018). We will elaborate further on the importance of these asymmetries in the discussion, highlighting the relationship between letter repetition and the different patterns of acoustic cues to prominence, which suggest that duration is not the acoustic cue that best correlates with letter repetition. We here note that the phonological trends we find in retraction of repetition to the penult appear to be better predicted by morphology than by phonology, despite suggestive tendencies for the latter to also play a role.

5. DISCUSSION

Overall, our results from three varieties of French demonstrate that dialectal differences in letter repetition are present. Of greater interest, however, is the source of those differences: while in certain cases the repetition patterns may be governed by distinct 'grammars' of repetition, the different dialects actually show considerable similarity with respect to the factors that influence which letters will be repeated. As we discuss in section 5.1, we will therefore suggest that they follow the same general preferences for repetition, though the exact effect size for a given factor can differ across varieties. The presence of variation in repeated letters can then provide evidence for differences in the phonological systems (e.g. different phonological statuses and patterns for the schwa), as it seems likely that the patterns we observe result from influence from dialects' phonological systems or different mapping between orthographic and phonological systems. In section 5.2 we pose the question of where the relatively consistent 'grammar' of repetition comes from, and circumscribe the role prosody plays in repetition. We conclude with open questions, directions for future work and the contributions of large-scale corpora from social media (section 5.3).

5.1. General repetition patterns

The first of these preferences is to repeat letters that map directly and consistently onto pronunciation, found both for vowels (e.g. defective <e> compared to full vowels) and for consonants (e.g. silent consonants compared to pronounced ones). This is found in all three varieties, albeit to varying degrees. The main competition for this factor is with another shared preference to repeat vowels instead of – or in addition to – consonants. The preference for repeating vowels over consonants is likely to itself be a pronunciation-based factor. Whereas letter repetition in English shows much smaller effects of whether the letter is a consonant or a vowel (Lamontagne & McCulloch 2017), this was a primary effect across varieties of French.

Laurentian French appears to pattern slightly differently in that silent final consonants are targeted by letter repetition more often than is the case in other varieties. One explanation is that orthographic final consonants in Laurentian French are associated with pronounced consonants more often than in other varieties (e.g. *tout*, sometimes *lit* and *nuit*). However, the effect would be surprisingly large for the relatively small number of words that the authors know to show this dialectal variation, particularly given the use of random intercepts in our statistical models. We propose that the different treatment of (silent) final consonants in writing follows from their distinct roles in conveying vowel qualities across dialects. More specifically, a "silent consonant" may affect the pronunciation of the preceding vowel – for example, the <ai> in *irai* 'go.FUT.1SG' reflects /e/, while the <ait> in *irait* 'go.FUT.3SG' and the <ais> in *irais* 'go.FUT.2SG' both reflect /ɛ/.[5] This is not consistently true across varieties: while such contrasts remain robust in Laurentian French (e.g. Côté 2012; Walker 1984), they are absent in Midi French (Coquillon & Turscan 2012) and are rapidly undergoing neutralization in many Metropolitan varieties (e.g. Berit Hansen 2012). This suggests that the grapheme-to-phoneme correspondences differ across varieties (and across speakers, where varieties exhibit such variation); in Laurentian French, that final consonant may in fact be the final part of a vowel digraph or trigraph because of its effect on the vowel pronunciation, while in Midi French that consonant is likely just a silent letter. This potential analysis predicts that we have collapsed two types of "silent consonants": those that affect the preceding vowel quality and those that do not. Future work may probe this possibility by comparing contexts in which "silent" consonants are associated with a different vowel phoneme to contexts in which silent consonants do not do so (e.g. *prix* 'prize' is pronounced no differently than would be expected without the final <x>). Future work may, in addition, investigate whether consonants that surface elsewhere in the paradigm (e.g. the <t> in *petit* 'small.MASC' may be associated with the /t/ in petite 'small.FEM' and therefore be repeated more often) are repeated more often, which would suggest the presence of a potential latent or floating consonant in the representation (for spoken French, e.g. Tranel 1995).

5.2. Source of repetition patterns

One of the noteworthy results we found is that repetition appears to target the right edge of the word in most contexts (but see below for further discussion of exceptions). There are two main sources that could explain the right-edge preference we find in French and that these are prosody and word recognizability. The first possible source is that prosodic prominence in French also targets the right edge of the word (e.g. Jun & Fougeron 1995), and that this prosodic prominence affects repetition preferences, consistent with Werry's (1996) proposal that letter repetition compensates for the lack of prosody in writing. The second possible source is that it could be easier to write or parse words when the repeated letters – which could hinder comprehension – are at the right edge of the word. The left edge of the word is particularly important to speech perception (Connine, Blasko & Titone 1993; Marslen-Wilson & Zwitserlood 1989; Salasoo & Pisoni 1985) and lexical access (Astheimer & Sanders 2011; Mitterer 2011; Pisoni et al. 1985), and we expect applies in contexts where non-standard spellings like those examined limit the reader's ability to use only whole-word pattern recognition to trigger lexical access (as would often be the case for reading; Ehri & McCormick 1998).

While both sources likely play a role in creating the patterns we find in French, we will focus particularly on the prosodic explanation as non-final letters are repeated so often. First, however, we will return to the question of which types

[5] There are exceptions to this distinction. For instance, plural <s> does not reflect a change in the preceding vowel (e.g. *mais* 'May.PL (the month)' is /me/) and certain words are exceptional (e.g. *vrai* 'true' is /vrɛ/ despite <ai> being word-final).

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

of letters – rather than the positions in the word – are favoured for repetition. On one hand, the preference for repeating vowels over consonants could reflect a larger principle in phonology: that of sonority. Vowels are inherently more sonorous than consonants, meaning that they could be more likely to be repeated. This suggests that sonority (a phonological consideration) could be the dominating factor in French letter repetition. On the other hand, another consideration may be more important than sonority: prosody. It is known that speakers access intonation when reading sentences, and this also appears to be the case when speakers are writing (e.g. Chafe 1988). As such, we suggest that intonation could, in fact, be a primary and overarching consideration when French speakers decide which letters to repeat. This may not be a conscious and introspective decision, however; repeating the letters that are associated with prosodically strong sounds (e.g. vowels or prominent syllables) in production could be a by-product of prosody influencing writing speed (Fuchs & Krivokapić 2016) and therefore authors repeating where duration is increased in speech or of authors following speech-based intuitions about what is a more intuitive location to have repetition (with duration being a likely consideration, though see discussion below). A prosodic explanation would also be consistent with a preference for repetition to be on the right edge of the word, since that is the direction of prominence assignment and of phrase-final lengthening in spoken French (e.g. Jun & Fougeron 1995; Nakata & Meynadier 2008).

Turning to schwa, it is well-known that the vowel is disfavoured for prominence assignment in French (e.g. Vaissière 1974), which could explain the reduced likelihood of repeating defective <e>. This pattern is found in all varieties, but is especially well-studied in Midi French as the resulting penultimate prominence creates a recognizable phrasal melody for which the variety is known. More varieties should be tested and varieties' prominence patterns should be compared to their repetition patterns, but this could be a promising direction of research, suggesting that it is – in specific circumstances – possible to indirectly study prosody through text, and that it may be possible to study phonological or phonetic alternations that influence prosody.

We finally turn to the initially surprising results of final syllable shape on the syllable targeted by repetition. These results would appear to undermine a phonological explanation for writers' choice of syllable in which to repeat letters. However, we propose that our expectations were too simplistic in two ways. First, we only considered prominence shift as a motivation for letter repetition, but (emphatic) focus would be a reasonable motivation to repeat letters. Dahan and Bernard (1996) observe not only that emphatic focus is associated both with the enhancement of initial (and, for adverbs, specifically pre-derivational) syllables in spoken French, but furthermore that adverbs are particularly likely to be targeted by emphatic focus. Second, while prominence shift may motivate a change in the syllable of repetition, we assumed that increased spoken duration would be the acoustic cue that best predicts letter repetition in writing. The results correlated with verbs are the best counterargument to this assumption: in previous work on the interaction between morphology and phonology in Laurentian French prominence assignment (Lamontagne et al. 2018), it was found that inflectional affixes often triggered a separation of acoustic cues such

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

that the penult is marked with increased amplitude and duration, whereas the final syllable is marked with a pitch accent. We leave for future work further exploration of the implications for Midi French: in this dialect, the word-final schwa may be analysed as an inflexional suffix in verbs (e.g. *aime* 'like.3sg'), meaning somewhat distinct patterning would be possible. However, our results suggest that the phonologically defective nature of the schwa may play a more important role than morphology does, unless the schwa results already reflect its frequent morphological role (e.g. also surfacing as a gender marker in many nouns and adjectives). Derivational affixes, on the other hand, allow all cues to be realized on the base-final penult, with the additional possibility to mark prominence on both the word-final and the affix-final syllables simultaneously. Our finding that final open syllables have exceptionally high rates of final-syllable repetition therefore suggests that pitch is the acoustic correlate of prominence most associated with letter repetition, rather than duration, and therefore that letter repetition is sensitive to more abstract prominence than what first may have been assumed because of the intuitive similarity between lengthening in speech and letter repetition in writing. Furthermore, our result of frequent letter repetition in non-final syllables for contexts where derivational affixes are more likely to be present (closed syllables, adverbs, certain adjectives) suggests that letter repetition reflects the availability of prominence retraction in speech, reflects the increased likelihood to place focus on these words (particularly adjuncts), or both.

As such, we infer that letter repetition is conditioned by morphology and phonology, which are both mediated by prosody. Our results additionally hint at the possibility that Laurentian French and the two European varieties share prosodic conditioning of prominence retraction to non-final syllables. The phonological (Lamontagne et al. 2017; Lamontagne 2020) and morphological (Lamontagne et al. 2018) conditioning of prominence retraction has predominantly been examined in Laurentian French, and we infer that the conditioning is likely to be similar in European varieties under examination (but likely at lower rates, based on repetition) based on the minimal significant differences in the conditioning of the syllable affected by letter repetition.

5.3. Concluding discussion

In summary, Twitter – and other sources of massive corpora of non-standard writing – allow for the analysis of phonological variables, even in cases where the writer is not presumed to have intended to convey a phonological variant. In this study, we found evidence that French varieties having differing phonological statuses for the schwa and having different vowel inventories showed evidence of these phonological traits through the patterns of letter repetition written by speakers of those varieties. In essence, the use of non-standard spelling allows French orthography – which reflects pronunciations from several centuries ago – to convey more modern phonological patterns. It further suggests that speakers' interpretation of the spelling system may vary between dialects in important ways (e.g. silent letters vs. digraphs and trigraphs) that may inform both how we processes grapheme-to-phoneme correspondences and how best to teach literacy.

While we cannot be certain in studies such as this one that the speakers are native speakers of the language under examination or that they have not immigrated to the dialect area, the relatively consistent behaviour of tweet authors within a dialect area suggests that outliers are not a problem given the quantity of data available for analysis. Furthermore, the relationship between authors' dialect areas and their repetition patterns suggests that the problem of tweet authors not necessarily tweeting from their native dialect area is not a sufficiently large problem to make the analysis of Twitter data impossible. Future work may seek to test for language contact and dialect contact in mobile authors or in authors who tweet in multiple languages, given the considerable differences in communicative norms across languages (e.g. between English and French for letter repetition, cf. Lamontagne & McCulloch 2017). Future work may also seek to confirm the relationship between informal writing and speech by examining how speakers read texts with non-standard elements (e.g. letter repetition, capitalization, tildes, clear instances of pronunciation-based spelling).

With resources made available as a result of new media platforms like Twitter, researchers have an increased opportunity to examine linguistic behaviour in contexts where speakers (or writers) are not adapting their styles to standardizing pressures like being in a laboratory setting. Additionally, the massive amount of data available makes it possible to create corpora that are suitable to examine exceptionally rare phenomena. In the case of this study, the availability of resources like Twitter enables us to examine multiple dialects of the same language and, with time, run comparable studies on numerous more languages to determine which patterns are robust to changes not only in dialect, but also in language or writing conventions. This is reflected in the growing body of sociolinguistic work using Twitter as its data source, including studies of lexical diffusion (e.g. Maybaum 2013), studies of dialect change and formation (e.g. Willis et al. 2019), and studies of social networks and how they influence language use and language change (e.g. Hale 2014).

The current study, for instance, is part of a larger project to determine the extent to which spelling systems and phonological systems interact cross-linguistically in creating non-standard spellings, and introducing data from other languages (e.g. Italian to examine patterns in cross-word germination). With respect to French, our results suggest that the varieties of French under examination have similar prosodic systems, for example with respect to the treatment of schwas (confirmed by existing work, where schwa is disfavoured in prominence assignment; Garde 1968; Prieto et al. 2005). The results also suggest that the prosodic system of European varieties should be probed further with respect to the conditioning of non-final prominence (which is a well-attested phenomenon, see e.g. Avanzi et al. 2011a, 2011b; Bardiaux & Mertens 2014; Carton et al. 1983; Goldman & Simon 2007; Lamontagne et al. 2017, 2018; Lamontagne 2020; Simon 2004, 2011). The results we obtained suggest similar conditioning for prominence retraction across varieties, but that the rate of prominence retraction would vary significantly by dialect.

Future work could seek to investigate register effects more broadly. In the current study, it is expected that the main variable (i.e. intentional letter repetition) is likely restricted to lower registers or otherwise to informal writing. However, not all variables are as limited in their availability and therefore could surface across written registers, for example morphosyntactic variables such as the order of ditransitive verb arguments in English (also called the dative shift or the dative alternation, e.g. Kendall et al., 2011). Alternatively, the patterns underlying (potentially accidental) spelling errors could be compared to those underlying intentional spelling variants. In either case, it could be enlightening to compare variation in written and spoken media in future studies in order to better understand what constraints the medium of communication imposes.

REFERENCES

Armstrong, S. (1999). *Stress and weight in Québec French*. M.A. thesis, University of Calgary.

Astheimer, L. B. & L. D. Sanders. (2011). Predictability affects early perceptual processing of word onsets in continuous speech. *Neuropsychologia* **49** (12): 3512–3516.

Avanzi, M., N. Obin, A. Bardiaux, & G. Bordal. (2011a). “Données et hypothèses sur la variation prosodique de 6 variétés de français parlées en France”. Journées PFC, Paris.

Avanzi, M., S. Schwab, J.-P. Goldman, P. Montchaud, I. Racine & H. Andreassen. (2011b). “Étude acoustique de l’accentuation pénultième dans trois variétés de français”. Journées PFC, Paris.

Bardiaux, A. & P. Mertens. (2014). Normalisation des contours intonatifs et étude de la variation régionale en français. *Nouveaux cahiers de linguistique française* **31**, 273–284.

Berit Hansen, A. (2012). A study of young Parisian speech: Some trends in pronunciation. In: R. Gess, C. Lyche and T. Meisenburg (eds), *Phonological variation in French: illustrations from three continents*. Amsterdam: John Benjamins, pp. 151–172.

Biber, D. (1995). *Dimensions of Register Variation. A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biers, Kelly. (2017). Vowel Epithesis Variation in French. *Cahiers of the Association for French Language Studies* **21**(1): 1–34.

Boula de Mareüil, P., B. Vieru-Dimulescu, C. Woehrling & M. Adda-Decker. (2008). Accents étrangers et régionaux en français. *Caractérisation et identification, Traitement Automatique des Langues*, **49**(3): 135–162.

Bybee, J. (1999). Usage based phonology. In: M. Darnell, E. Moravcsik, F. Newmeyer, M. Noonan & K. Wheatley, (eds.), *Functionalism and formalism in linguistics, volume I: General papers*. Amsterdam: John Benjamins. 211–242.

Carton, F., M. Rossi, D. Autesserre & P. Léon. (1983). *Les accents du français*. Paris: Hachette.

Chafe, W. (1988). Punctuation and the prosody of written language. *Written Communication*, **5**: 396–426.

Connine C., Blasko D & Titone D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *J Mem Lang.* **32** (2): 193–210.

Côté, M.-H. (2012). Laurentian French (Québec): extra vowels, missing schwas and surprising liaison consonants. In: R. Gess, C. Lyche, and T. Meisenburg (eds), *Phonological variation in French: illustrations from three continents*. Amsterdam: John Benjamins, pp. 235–274.

Coquillon, A. & G. Turcsan. (2012). An overview of the phonological and phonetic properties of Southern French. In: R. Guess, C. Lyche, T. Meisenburg (eds), *Phonological variation in French: Illustrations from three continents*. Amsterdam: John Benjamins Publishing Company, pp. 105–128.

Dalola, A. (2017*). #YouAreWhatYouTweet: Identity and vowel devoicing in French-language tweets*. Poster presented at New Ways of Analyzing Variation (NWAV) 46 (Madison, Wisconsin, USA – November 2–5, 2017).

Dumas, D. (1974). Durée vocalique et diphtongaison en français québécois. *Cahier de linguistique*, **4**: 13–55.

Ehri, L. C. & McCormick, S. (1998). Phases of word learning: Implications for instruction with delayed and disabled readers. *Reading & Writing Quarterly* **14**(2), 135–163.

Eisenstein, J. (2013). Phonological factors in social media writing. *Proceedings of the Workshop on Language Analysis in Social Media*: 11–19.

Eychenne, J. (2014). Schwa and the loi de position in Southern French. *Journal of French Language Studies*, **24**: 223–53.

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Fuchs, S. & J. Krivokapić. (2016). Prosodic Boundaries in Writing: Evidence from a Keystroke Analysis. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.01678.

Garde, P. (1968). *L'accent*. Paris: Presses Universitaires de France.

Gentry, J. (2015). *twitteR: R Based Twitter Client*. R package version 1.1.9.

Goad, H. & A.-E. Prévost. (2011). *A test case for markedness: The acquisition of Québec French*. Ms., McGill University.

Goldman, J.-P. & A.-C. Simon. (2007). *La variation prosodique régionale en français (Liège, Vaud, Tournai, Lyon)*. Regards croisés sur la phonologie du français contemporain (Paris, December 6–8, 2007).

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, **33.2**: 1–22.

Hale, S. A. (2014). "Global connectivity and multilinguals in the Twitter network". *CHI*. 10 pp.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix, (Eds.). *Talker variability in speech processing*. San Diego: Academic Press, 145–165.

Jun, S-A & C. Fougeron. (1995). The Accentual Phrase and the Prosodic structure of French. *Proceedings of ICPhS (Stockholm, Sweden)*, **2**: 722–725.

Kendall, T, J. Bresnan & G. Van Herk. (2011). The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory* **7**(2): 229–244.

Koch, P. & W. Oesterreicher. (2012). Language of Immediacy – Language of Distance. Orality and Literacy from the Perspective of Language Theory and Linguistic History. In: C. Lange, B. Weber & G. Wolf, (eds.), *Communicative Spaces. Variation, Contact, and Change. Papers in Honour of Ursula Schaefer*. Frankfurt: Lang. 441—473.

Kotzur, G. (2015). Nothing but "typed talk"? Analysing discourse in computer-mediated communication by employing Koch and Oesterreicher's framework model. *Münchener Beiträge zur Allgemeinen und Historischen Sprachwissenschaft* **4**: 46–56.

Lamontagne, J. (2020). *Interaction in Phonological Variation: Grammatical Insights from a Corpus-Based Approach*. Doctoral thesis, McGill University.

Lamontagne, J. & G. McCulloch. (2017). *Wayyy longgg: Orthotactics and phonology in lengthening on Twitter*. Presented at the 91st Annual Meeting of the Linguistic Society of America in Austin, Texas (USA).

Lamontagne, J., H. Goad & M. Sonderegger. (2017). *Evidence of weight sensitivity in Laurentian French prominence assignment*. Presented at the Annual meeting of the Canadian Linguistics Association in Toronto (Canada).

Lamontagne, J., H. Goad & M. Sonderegger. (2018). *Morphological and Phonological Motivations for Prominence Shifts in French*. Presented at the Montreal-Ottawa-Toronto Phonology Workshop in Hamilton (Canada).

Law, J. (2017). *"Les jons fon skil veulent": Reflections of the French nasal vowel shift in variant orthography*. Presented at New Ways of Analyzing Variation (NWAV) 46 (Madison, Wisconsin, November 2–5, 2017).

Marslen-Wilson W. & P. Zwitserlood. (1989). Accessing spoken words: the importance of word onsets. *J Exp Psychol: Hum Percept Perform* **15** (3): 576–585.

Maybaum, R. (2013). Language change as a social process: Diffusion patterns of lexical innovations in Twitter. *Berkeley Linguistics Society* **39**: 152–166.

McCulloch, G. & J. Lamontagne. 2020. La phonologie du français sur Twitter. In: D. Bigot, D. Liakin, R. A. Papen, A. Jebali & M. Tremblay, (eds), *Les français d'ici en perspective*. Québec: PUL. 171–194.

McCulloch, G. & J. Lamontagne. 2018. *Troppppppp Loooongueuuhhhh: la phonologie du français sur Twitter*. Presented at Les français d'ici 2018 in Montreal (Canada).

Mitterer, H. (2011). Recognizing reduced forms: Different processing mechanisms for similar reductions. *Journal of Phonetics* **39**: 298–303.

Morin, Y. C. (1978). The status of mute 'e'. *Studies in French Linguistics*, **1**: 79–140.

Nakata, S. & Y. Meynadier. (2008). Final accent and lengthening in French. *Conference on SpeechProsody (Campinas, Brazil)*: 567–570.

New B., C. Pallier, L. Ferrand & R. Matos. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, **101**: 447–462.

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Pierrehumbert, J. (2003). Probabilistic Phonology. In: R. Bod, J. Hay & S. Jannedy (Eds.), *Probability theory in linguistics*. The MIT Press, 177–228.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics* **2**: 33–52.

Pisoni, D. B., H. C. Nusbaum, P. A. Luce & L. M. Slowiaczek. (1985). Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech Communication* **4**(1–3): 75–95.

Prieto, P., M. D'Imperio & B. Gili Fivela. (2005). Pitch Accent Alignment in Romance: Primary and Secondary Associations with Metrical Structure. *Language and Speech* **48**(4): 359–396.

Rose, Y. & C. dos Santos. (2008). Stress Domain Effects in French Phonology and Phonological Development. *Roman Linguist*: 89–104.

Salasoo A & D. Pisoni. (1985). Interaction of knowledge sources in spoken word identification. *J Mem Lang.* **24** (2): 210–231.

Schnoebelen, T. (2012). Do you smile with your nose? Stylistic variation in Twitter emoticons. *University of Pennsylvania Working Papers in Linguistics*, 18.2: http://repository.upenn.edu/pwpl/vol18/iss2/14.

Simon, C. (2011). "La prosodie des accents régionaux en français. État des lieux". Journées PFC, Paris.

Simon, A-C. (2004). *La structuration prosodique du discours en français*. Bern: Peter Lang.

Tatman, R. (2015). #go awn: Sociophonetic variation in variant spellings on Twitter. In: S. Onosson and M. Huijsmans (eds), *Working Papers of the Linguistics Circle 25.2: Proceedings of the 31st annual North West Linguistics Conference*, 97–108.

Tatman, R. (2016). 'I'm a spawts guay': Comparing the use of sociophonetic variables in speech and Twitter. *Selected Papers from NWAV 44*, **22.2**: 161–170.

Thurlow, C. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online* **1**(1).

Tranel, B. (1995). French final consonants and nonlinear phonology. *Lingua* **95**(1–3): 131–167.

Vaissière, J. (1974). On French prosody. *Res. Lab. Electr. Prog. Report, MIT.*, **115**: 212–23.

Walker, D. C. (1984). *The Pronunciation of Canadian French*. Ottawa, Canada: University of Ottawa Press.

Wang, A., T. Chen & M.-Y. Kan. (2012). Re-tweeting from a Linguistic Perspective. *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*: 46–55.

Werry, C. C. (1996). Linguistic and Interactional Features of Internet Relay Chat. In: Susan C. Herring (ed), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*: 47–64.

Wikström, P. (2017). *I tweet like I talk: Aspects of speech and writing on Twitter*. Doctoral thesis, Karlstad University.

Willis, D., D. Gopal, T. Blaxter & A. Leemann. (2019). *Big data for a small language: Mapping variation in Welsh on social media*. New Ways of Analyzing Variation (NWAV) 48 (Eugene, Oregon, Oct. 10–12, 2019).

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Appendix A The model output for final consonants based on 27309 tokens

	Post. Mean	CI 95% (min.)	CI 95% (max.)	p-value (MCMC)
Final Cons.	0.9276	0.5650	1.1989	<0.001***
Non-final letter	1.0692	0.5181	1.4626	0.004**
Final Cons. : Midi	−0.4365	−0.5699	−0.2586	<0.001***
Non-final letter : Midi	0.4251	0.2700	0.5842	<0.001***
Final Cons. : Metropolitan	−0.1275	−0.2444	−0.0077	0.038*
Non-final letter : Metropolitan	0.7334	0.6133	0.8627	<0.001***
Final Cons. : Silent	−1.3763	−1.9998	−0.8659	<0.001***
Non-final letter : Silent	10.2202	6.7253	12.6277	<0.001***
Final Cons. : Midi : Silent	0.8449	−0.0770	2.0322	0.108
Non-final letter : Midi : Silent	−0.8504	−1.9658	0.1192	0.070
Final Cons. : Metropolitan : Silent	0.2906	−0.3133	0.8492	0.306
Non-final letter : Metropolitan : Silent	−0.8986	−2.2849	0.4732	0.322

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Appendix B The model output for lengthening consonants compared to other pronounced consonants based on 5732 tokens

	Post. Mean	CI 95% (min.)	CI 95% (max.)	p-value (MCMC)
Non-final letter	2.1710	1.5226	2.6788	<0.001
Final and non-final	−0.8495	−1.5724	−0.1871	0.010**
Non-final letter : Midi	0.0436	−0.3879	0.4722	0.844
Final and non-final : Midi	0.0918	−0.4399	0.7603	0.798
Non-final letter : Metropolitan	−0.0105	−0.3733	0.2898	0.958
Final and non-final : Metropolitan	−0.2539	−0.7301	0.2063	0.274
Non-final letter : Pronounced	−0.7625	−1.4059	−0.0318	0.026*
Final and non-final : Pronounced	−0.2401	−1.0618	0.5591	0.554
Non-final letter : Midi : Pronounced	0.5360	−0.1243	1.1435	0.094
Final and non-final : Midi : Pronounced	0.6158	−0.2026	1.5011	0.148
Non-final letter : Metropolitan : Pronounced	0.4988	0.0300	0.9819	0.042*
Final and non-final : Metropolitan : Pronounced	0.4718	−0.2301	1.1331	0.178

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Appendix C The model output for final schwas based on 39851 tokens

	Post. Mean	CI 95% (min.)	CI 95% (max.)	p-value (MCMC)
Non-final letter	0.5736	0.3055	0.8404	0.002**
Final and non-final	−0.9045	−1.1374	−0.6777	<0.001***
Non-final letter : Midi	0.7503	0.6342	0.8539	<0.001***
Final and non-final : Midi	−0.1843	−0.3062	−0.0781	0.006**
Non-final letter : Metropolitan	0.8116	0.7278	0.8861	<0.001***
Final and non-final : Metropolitan	0.0407	−0.0492	0.1253	0.392
Non-final letter : Schwa	0.9493	0.2292	1.7331	0.016*
Final and non-final : Schwa	1.2937	0.7593	1.8212	<0.001***
Non-final letter : Midi : Schwa	−2.2213	−2.8364	−1.6461	<0.001***
Final and non-final : Midi : Schwa	−1.3679	−1.9477	−0.7545	<0.001***
Non-final letter : Metropolitan : Schwa	−1.4642	−1.9736	−0.9555	<0.001***
Final and non-final : Metropolitan : Schwa	−0.4870	−0.9943	−0.0008	0.054

Appendix D The model output for final schwas based on adjacent segments based on 1437 tokens

	Post. Mean	CI 95% (min.)	CI 95% (max.)	p-value (MCMC)
Non-final letter	2.1341	1.3038	3.0657	<0.001***
Final and non-final	0.5405	−0.4628	1.4909	0.258
Non-final letter : Midi	−1.5363	−2.1568	−0.9205	<0.001***
Final and non-final : Midi	−1.8401	−2.4967	−1.2018	<0.001***
Non-final letter : Metropolitan	−0.5542	−1.0777	−0.0120	0.038*
Final and non-final : Metropolitan	−0.4843	−1.0295	0.0538	0.094
Non-final letter : Vowel	0.3905	−3.1522	3.7475	0.828
Final and non-final : Vowel	0.6850	−2.4083	4.0521	0.676
Non-final letter : Midi : Vowel	1.1846	−3.5134	5.6312	0.624
Final and non-final : Midi : Vowel	3.8657	0.1944	8.2870	0.056
Non-final letter : Metropolitan : Vowel	−2.3111	−6.8522	1.3639	0.254
Final and non-final : Metropolitan : Vowel	−0.1862	−3.9204	3.2597	0.934

https://doi.org/10.1017/S0959269521000223 Published online by Cambridge University Press

Appendix E The model output for the syllable undergoing repetition based on 23,095 words of two or more syllables

	Post. Mean	CI 95% (min.)	CI 95% (max.)	p-value (MCMC)
Both	−5.9258	−7.3460	−4.2574	<0.001***
Non-final	−3.9377	−5.1417	−3.0240	<0.001***
Both:Midi	0.8009	−0.2407	1.8412	0.192
Non-final:Midi	0.6486	0.1121	1.2014	0.006**
Both:Hexagonal	1.1802	0.4181	2.1207	<0.001***
Non-final:Hexagonal	0.6261	0.2120	1.0927	<0.001***
Both:Open	0.4324	−0.5115	1.5934	0.514
Non-final:Open	−0.8379	−1.6144	−0.1017	0.018*
Both:Nasal	0.7280	−0.3940	1.9557	0.224
Non-final:Nasal	−0.0271	−0.8665	0.9046	0.896
Both : Midi : Open	−0.1391	−1.3020	1.4333	0.776
Non-final : Midi : Open	−0.8338	−1.7244	−0.0362	0.032*

Cite this article: Lamontagne J and McCulloch G (2022). Phonological variation on Twitter: Evidence from letter repetition in three French dialects. *Journal of French Language Studies* **32**, 165–196. https://doi.org/10.1017/S0959269521000223