Applying Twitter data to linguistic research: Case studies in sound, form and meaning

Michael Dow
Université de Montréal

Patterns and Singularities:
Department of French & Italian 2023 Graduate Conference

April 1, 2023

Introduction

Patterns

- Data are **patterns**: repetitions or divergences which we gather for a certain purpose.
- Sometimes, it's helpful to abstract away from "noise" and paint things as black and white—for instance, a process is active or not, phonetic or phonological, and so on.
- For better or worse, early generative linguistics, especially phonology, is often painted this way.
- Real-world variation is increasingly more central in theoretical frameworks, no doubt due in part to the rise in laboratory/experimental methods and "big data."
- See, for instance, optional rules (Labov, 1969), multi-stratal theories of phonology (e.g., Kaisse and Shaw, 1985), and direct mirroring of real-life rates of application in Optimality Theory (e.g., Anttila, 1997).

Singularities

- Variation is a Pandora's box, especially when those numbers are directly incorporated into our models.
- What truth are we trying to get at here? Should the mechanisms accounting for variation be considered linguistic entities, or are they extralinguistic, existing primarily for the researcher describing them (cf. Coetzee, 2016)?
- Such an endeavour requires a comprehensive understanding of **what types of processes** are variable **in which languages** and **to what extent**.
- Such questions strain the limits of our traditional empirical methods. What if we filled this gap using freely available, modern solutions?
- While these questions will remain very much open, these are the **singularities** I want us to think about.

Context

- My background is first as an experimental phonologist, with an eye for the phonetic-phonological interface.
- I started working with Twitter data around early 2020, somewhat as a necessity, but also due to an emerging interest in coding.
- The questions that come up in the two areas are not so different in essence:
  - For instance, if a vowel is 50% nasalized, can we call that *deliberate*? What if 50% of all vowels are 50%+ nasalized in a corpus? Is it an *active* process in the language?
  - Compare with: At what point can we say "la COVID" became the norm in Québec? When does the word "pronouns" become a political dogwhistle in the US?

Today's talk

- We'll be looking at case studies from my research using Twitter data in linguistics research, in various sub-fields.
  - **Form**: How novel words are assigned gender in French.
  - **Meaning**: How existing words can go from neutral to charged and expand in meaning, all this decidedly *not* in a void.
  - **Sound**: How social media can inform us about internal and external factors affecting pronunciation.
- I also want to use this talk to invite people to exploit this data source, as well as provide resources and advice to those already interested.

Main take-away

Twitter data—like any sort of "big data"—will never supplant traditional methods in linguistics, nor should they.

Doing single-variable percentages *à la* field studies or sociolinguistic experiments may not be viable.

But Twitter data do provide valuable, powerful and (for now) free tools for:

- Looking at change over time,
- Mapping variables geographically,
- Quantifying variation,
- Affirming pre-existing hypotheses, and
- Inspiring follow-up studies.

Outline

A brief overview of Twitter

Basics & history

- Social media site/app for posting text and/or visual media: tweets, replies, retweets, quote tweets.
- Text length per post for basic users has ranged from 140 to 280 characters over time, with 4,000 available as of recently to paid subscribers.
- Total users currently surpass 450 million.
- Estimated 500 million tweets per day
- Fairly even age distribution, with 25-34 representing the largest percentage of users, and world users skew overwhelmingly male on average.
- Users in the US are fairly evenly distributed for race.
- As of 2020, the majority of content published to Twitter comes from Democrat-leaning accounts, but...

Things we can do

Researchers with academic-level API access can:

- Get full counts for a query (without counting against monthly quota)
- Get tweets & metadata for a query (10M tweets/month)
- Get user timelines (limit 3,200 most recent)
- Specify language and custom time frame
- Search keyword, set of keywords or exact expressions
- Exclude according to keywords/expressions and other criteria (e.g., replies, verified accounts, etc.)
- Search a country, a point radius (set of long-lat coordinates) of 25 miles or geographical bounding boxes (set of 2 long-lat coordinates) 25×25 miles

Things we can't do

- Force accented characters (e.g., *amá* will return *ama* and *amá*, even as an exact expression)
- Perform substring searches or use "jokers" (e.g., no way of searching for "all words ending in x")
- Download everything in one chunk
- Specify *a priori* a language not supported by Twitter's in-house, automatic detection
- Predict what's going to happen next with Twitter's current... direction

Le COVID ou la COVID?

Dow and Drouin (in press)


Mathieu Avanzi
@MathieuAvanzi
👇👇👇👇
Arrêtez de dire LA covid
imgflip.com
3:07 PM · May 2, 2021 · Twitter for iPhone
1,693 Retweets 121 Quote Tweets 10K Likes

Introduction

- The arrival of new nouns in French necessarily brings with it questions of its grammatical gender, i.e., whether it is masculine or feminine.
- Despite a slight statistical bias for the masculine (giving rise to notions of "default gender"), especially when it comes to English borrowings, multiple factors (within or beyond the French lexicon) interact with gender and can influence a new word's gender.
- The sudden and well-documented emergence of the word "COVID-19" allows us to trace its evolution and variation over time, in many different varieties of French.

French gender

- French nouns belong to one of two classes which determines their behaviour with respect to several phenomena (e.g., definite article: masculine *le* vs. feminine *la*) = gender.
- Relatively opaque in French compared to other languages (e.g., Corbett, 1991) but follows some regularities:
  - Suffixes always contribute the same gender and overwrite that of the base noun (e.g., diminutive -*ette* = feminine ; *le cigare* > *la cigarette*) (Surridge, 1993)
  - Certain word-final sound sequences are very predictable for gender (e.g., Tucker et al., 1977), but often both morphology and orthography (Lyster, 2006) have to be considered (e.g., *le squelette* 'skeleton', *le diabète* 'diabetes')
- Evidence that French speakers pay attention to these cues during acquisition (e.g., Carroll, 1989), in processing lexical information and in assigning gender to new words (e.g., Holmes and Segui, 2004)

Gender & borrowings

- French lexicon fairly balanced for gender, but 85% of contemporary borrowings from languages without gender are masculine, contributing to perception of the masculine as "default" or "unmarked" gender (Roché, 1992, 114-116).
- English words often receive the gender of their French equivalents (Haden and Joliat, 1940; Lupu, 2005; Nymansson, 1995), e.g., *une* ***love affair*** < *une affaire*.
- Another important factor is ellipsis with an unexpressed French noun, e.g., *une (voiture)* ***Ford***.
- Phonetic analogy plays a smaller role (Belleau, 2016) and can conflict with the above factors, giving rise to variation, e.g., *le/la* ***new beat*** (f. by ellipsis, m. due to final [it]) (Nymansson, 1995).

Regional differences: Europe & North America

- No significant differences in gender of nouns shared between European and North American varieties of French, or in gender of English borrowings (Belleau, 2016; Haden and Joliat, 1940; Nymansson, 1995), with a few important exceptions.
- The gender of certain words (e.g., *party*) and morphemes (e.g., some words in *-ing*) aside, two main factors distinguish the two:
  - Vowel-final English words tend to be masculine and consonant-final words feminine in Québécois French, unlike in European French (Léard, 1995).
  - Monosyllabic words (e.g., *job*) tend to be masculine in European French, unlike in Canadian French (Belleau, 2016).

Regional differences: Africa

- Documented variation in gender distinction in learners of certain varieties (Biloa, 2003; Calvet and Dumont, 1969; Holtzer, 2004) and certain words in Chadian French (Ndjerassem, 2005).
- Omission of gender-signalling determiners documented in some varieties (Boutin, 2007; de Féral, 2006; Jabet, 2006; Telep, 2014), leading to variation (Ayewa, 2009; Herault and Vonrospach, 1967).
- English loanwords in African varieties of French are extensively documented (e.g., Schmidt, 1990), but no synthesis on gender available.
- Some feminine loans of note in Gabonese (Boucher and Lafage, 2000), Chadian (Ndjerassem, 2005) and Cameroonian (Nzesse, 2009) French, but not enough to derive any significant trends.

A brief history of "COVID"

- On February 11, 2020, the WHO gave the disease caused by the virus SRAS-CoV-2 the abbreviated name "COVID-19" (World Health Organization, 2020).
- The term was generally masculine in WHO French communications until March 6, when it explicitly opted for the feminine (Avanzi, 2020).
- The same day, Radio-Canada (Bonsaint, p.c.) and the Office québécois de la langue française (Darras, p.c.) updated their files to reflect this.
- No public recommendation of the feminine is given by the Académie française or any of its members until May 7 (Académie française, 2020).
- In all cases, the reasoning for the feminine is the same, i.e., that the referent is *une maladie*, or the 'D' in COVID (*disease*), whether expressed or not.

Methodology: Twitter

- "Rehydrated" tweets from the COVID-19-TweetIDs repository (Chen et al., 2020).
- French-language tweets from January 21-June 31, 2020 with explicit gender marking retained.
- Retweets and formulaic tweets removed
- Place extracted from both geotagged fields and user-provided place description, with manual verification of small subset.
- Only users located in North American, African and European continents retained.
- Users split up into three groups according to follower size.

Methodology: Traditional media

- Eureka.cc, a media aggregator, queried for gender-marked instances of "COVID" in French-language media for each of the three continents.
- The number of articles returned for each gender was recorded for each month in the same time period as the Twitter corpus.

Results

	Africa	N. America	Europe	**Total**
February	47	64	417	528
March	904	728	4636	6268
April	2270	1460	9367	13097
May	2385	1817	10169	14371
June	6349	5197	30244	41790
Total	11955	9266	54833	**76054**

Table 1: Number of tweets by continent, per month

Twitter results

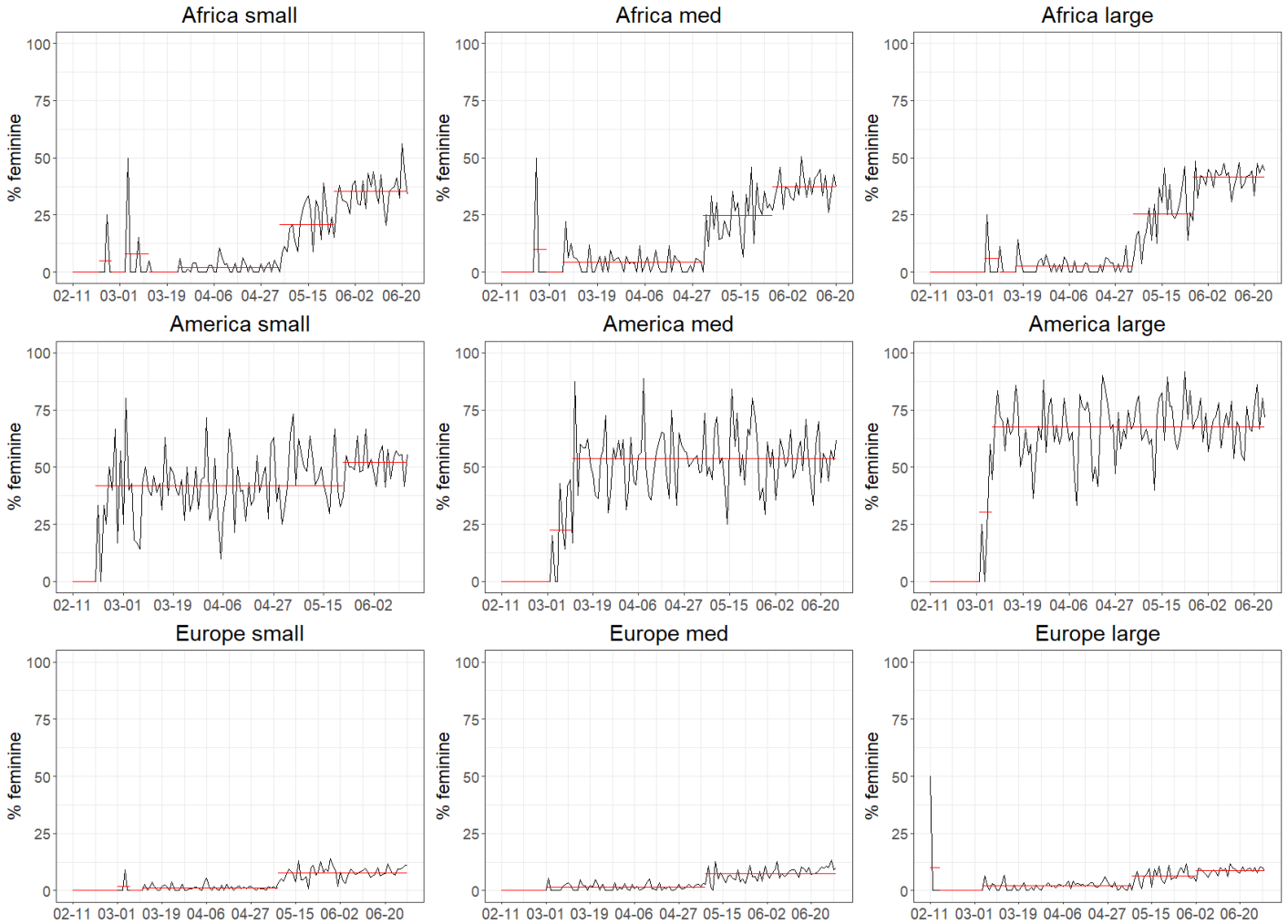


Fig. 1. % feminine, with breakpoints

Traditional media results

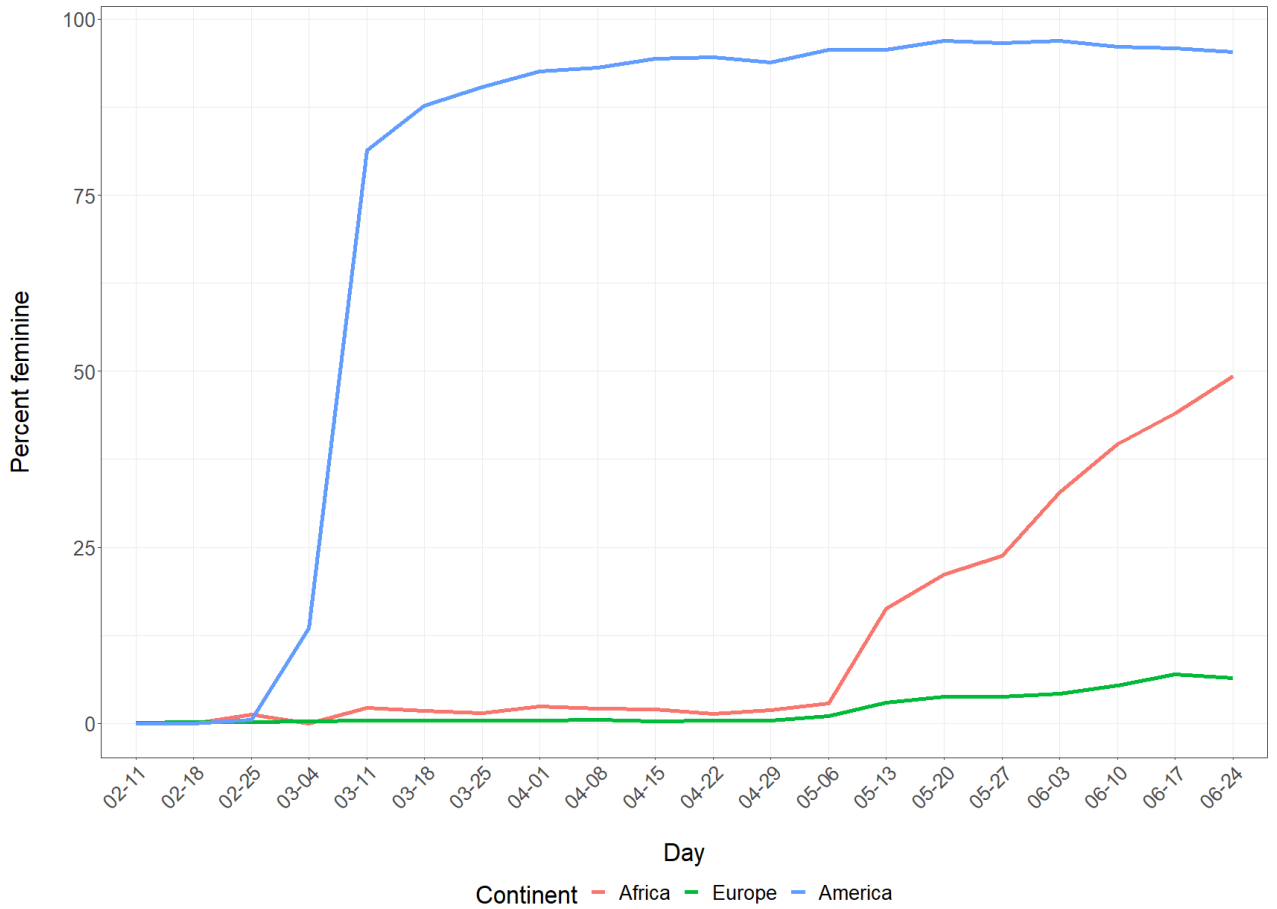


Fig. 2. % feminine per week, by continent

Results

- North American Twitter users and traditional media adopt very quickly the feminine, coinciding with the recommendations from the WHO, Radio-Canada and the OQLF in March 2020.
- Use of the feminine in African varieties sees a significant spike following the recommendation of the Académie française in May 2020.
- The masculine appears to be stable in European varities, regardless of date.

Discussion: Internal factors

- While we gather from Lexique (Gimenes et al., 2020) that [id] is a predominantly masculine ending in French, recall that Québécois French tends to assign the feminine to consonant-final English loans.
- It should be noted, however, that some express skepticism about the role of phonetic factors (Poplack, 2018; Poplack et al., 1982) and find that frequency is the determining factor in the establishment of a "fixed" gender.

Discussion: External factors

- Media outlets and linguistic authorities have a unique relationship in Québec, and little resistance in French Canadian media was met behind the scenes (Darras, p.c.; Bonsaint, p.c.).
- Québécois speakers tend to feel more positively towards and be more respectful of recommendations of local language authorities, in contrast with European speakers (Chalier, 2018, 2019; Kim, 2017; Maurais, 2008; Pöll, 2005; Pustka et al., 2019; Sebková et al., 2020; Tremblay, 1994).
- African francophone media (traditional and social) may appear to defer to the Académie Française for such matters. However, post-colonial, centralized language policy has largely proven ineffectual, and home-grown innovation is increasingly preferred (O'Mahony, 2019; Spolsky, 2018).

(*Psst*, what about now?)

Just looking at Canada, France and the Democratic Republic of Congo on Twitter (all users)...

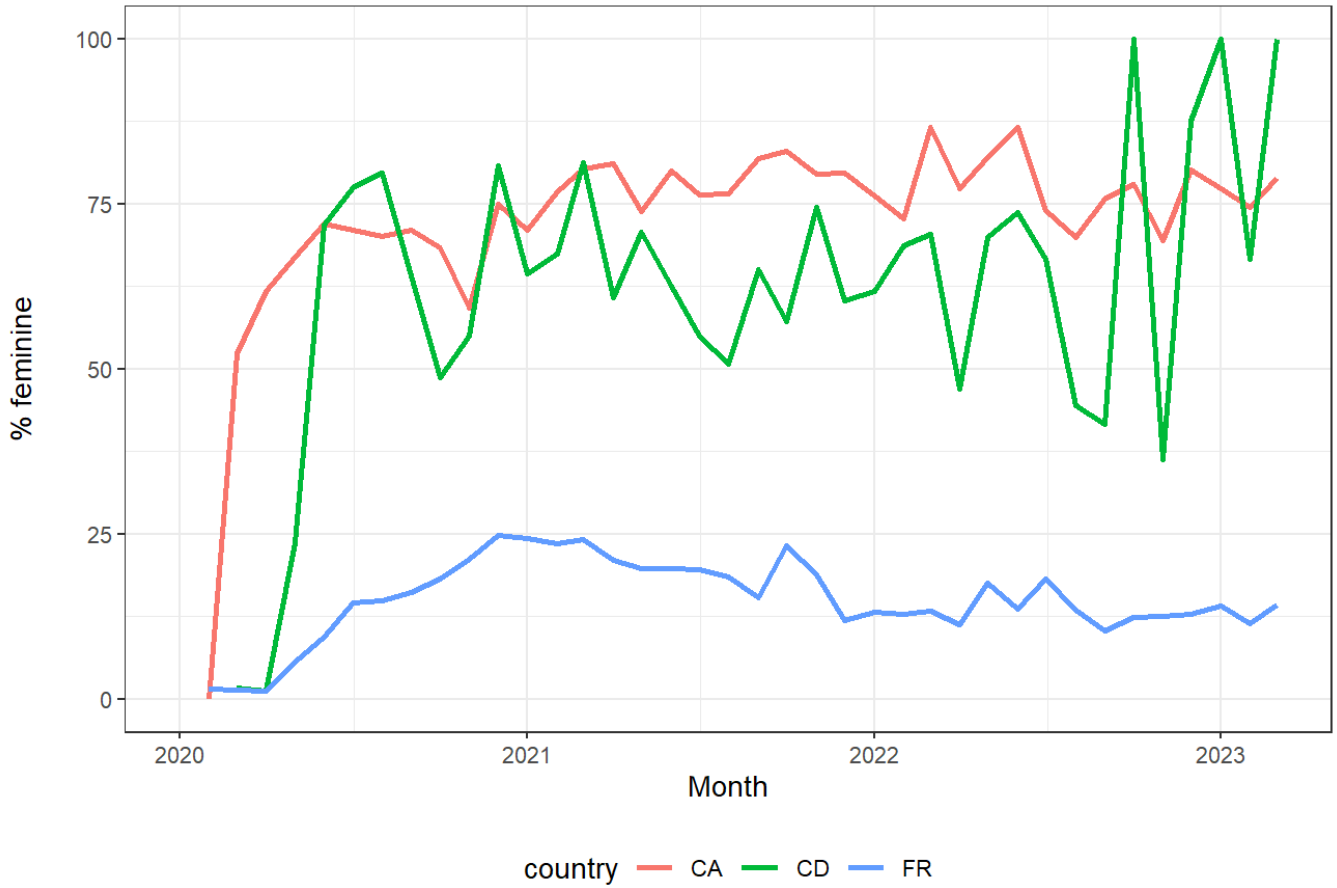


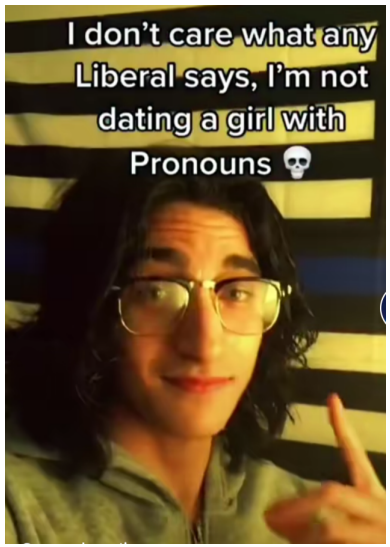
Fig. 3. % feminine use of "COVID" over time

Pronouns

Biers, Dow, Beuerlein et al. (2023)

Introduction

- What was once the neutral grammatical term "pronoun" has taken on new meaning recently.
- Beyond increasingly commonplace statements such as "my pronouns are he/him" in a variety of settings, we also see statements equating "pronouns" with gender inclusivity.
- In this project, we examine the evolution of the meaning of "pronouns" over time, with a particular eye for:
  - Amelioration vs. pejoration: Is the word gaining positive or negative associations?
  - Broadening vs. narrowing: Is the word becoming more specific or more general?



Corpus

- All English-language tweets containing "pronoun(s)" between 04/2006-11/2022 gathered $\longrightarrow$ corpus of ~9.3M tweets
- For 1 random day per month, count of English-language tweets estimated by querying high-frequency stopwords like "to," etc.
- Relative usage for each month: "pronoun(s)" tweets/estimated total English tweets

Methodology: Political affiliation

- User description extracted for random sample of 10% of profiles.
- Frequent, politically-charged keywords identified and defined as either right- (e.g., *MAGA*, *Republican*) or left-leaning (e.g., *non-binary*, *Democrat*), with values of -1 and 1, respectively.
- Values added up for each user and top 2000 users for each side selected—essentially, the most visibly political users were chosen.
- These users' tweets from the corpus + 30 most recent tweets from each user were compiled for training data.
- Logistic regression model trained using `sklearn` (Pedregosa et al. 2011, cf. also Paialunga 2021) with standard settings and an accuracy of 76%.
- This classifier was then run on the remaining users' tweets from the corpus to give them a "GOP probability" score.

Methodology: Sentiment analysis

- Sentiment analysis with VADER (Valence Aware Dictionary and sentiment Reasoner) (Hutto & Gilbert 2014)
- This algorithm functions similarly to more basic sentiment dictionaries, with the added bonus of being designed for social media and being trained for context.
- Each tweet given a composite score of -1 to 1.
- In the first graph, sentiment is discretized as either negative or positive.

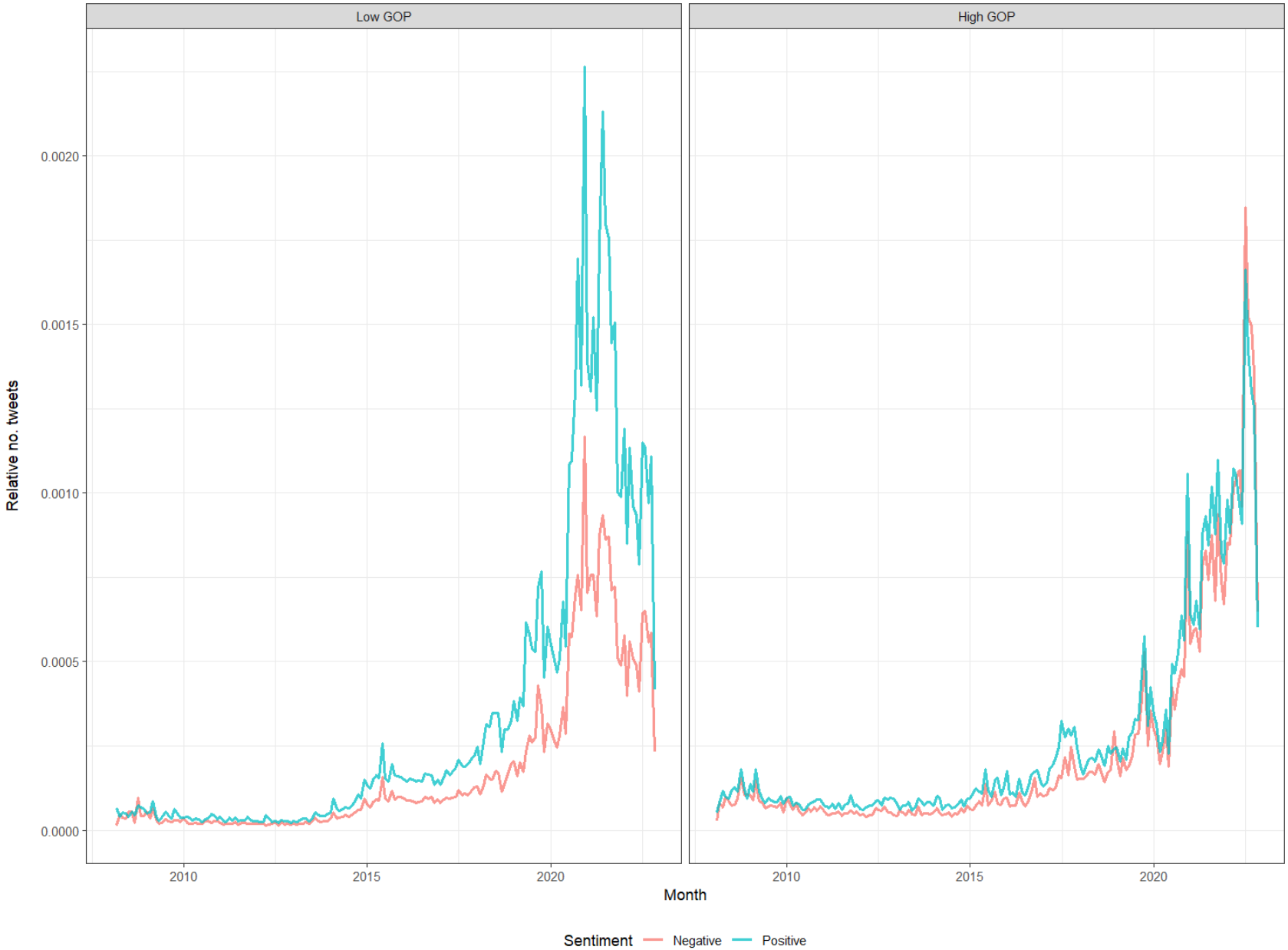


Fig. 4. Positive vs. negative sentiment over time by political affiliation, relative

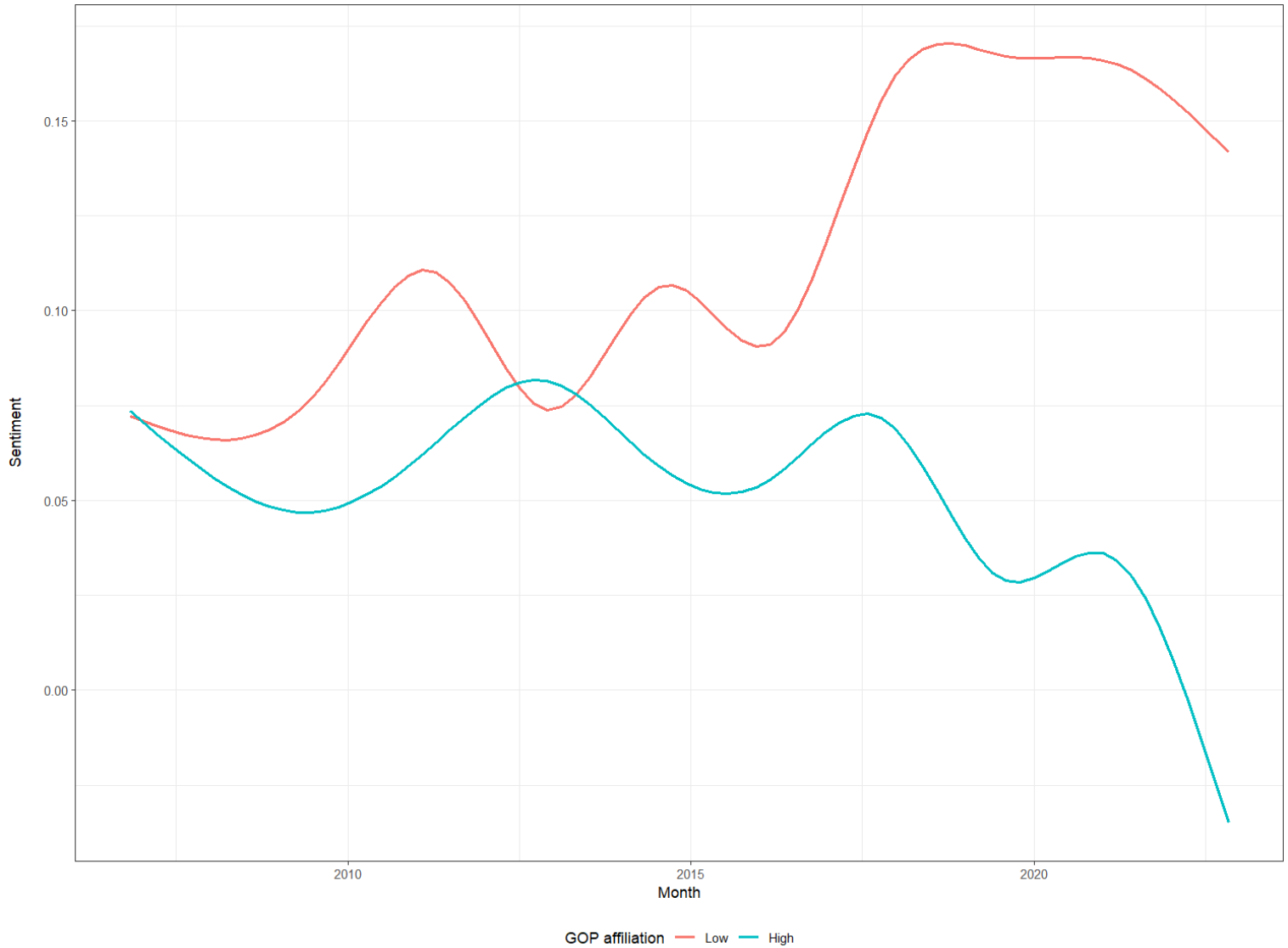


Fig. 5. Sentiment over time by political affiliation, smoothed conditional means

Discussion

- Relative number of tweets sees sharp rises coinciding with important moments in US politics.
- Positive tweets start to outstrip the negative in the low-probability GOP group.
- Clear bifurcation between the two groups that only seems to be widening.
- Next steps are examining the Twitter corpus in a more qualitative manner. This will include looking at co-occurrences and random samples.
- Themes we have already identified in qualitative data from Twitter and other platforms include: possession, ideology, threats to hegemonic hetero-masculinity, moral panic and absurdism.
- Keep resisting, share your joy and don't get baited into politically opportunistic messaging (cf. Kirby Conrod's work).

Twitter-MPhon pilot studies

Dow, Lareau and Drouin (upcoming)

Studying sounds with the written word

- As an exploratory venture, how can we use this particular medium to study phonetic or phonological processes?
- Philosophy from the start: The more languages, the better!
- Focus on open, multilingual and non-language-specific resources $\longrightarrow$ generalizable procedure simply by changing language ID and lexical/phonotactic criteria.
- Leave the bulk of human arbitration to verifying queries, interpreting the data and looking at qualitative samples.

Phenomena

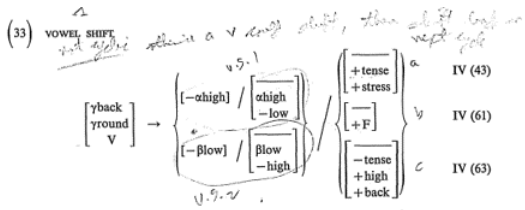
- Our initial plan was to look only at **morphophonological** or **lexical-phonological** processes. These are, *grosso modo* changes which:
  - Occur to sounds during the building of larger words or syntactical units,
  - Are commonly reflected in orthography.
- We also focused on phenomena which are known to vary along certain lines (esp. sociolinguistic and geographic or dialectal) and/or have a non-negligible degree of lexical exceptionality.

Compare. . .

- More general *flapping* in English, where *t* and *d* are made to sound alike (e.g., *a**t**om-A**d**am*). This process. . .
  - Occurs not only within words, but between units (*knotting-nodding*) and words (*get a life*),
  - Is known to be extremely regular in practice
  - Is phonetically "natural," and
  - Is not represented in orthography.
- . . . versus *learned backing* in French (Dell and Selkirk, 1978), where vowels move around (e.g., *fl**eu**r-fl**o**ral*)
  - But in a somewhat unintuitive or unsystematic way (e.g., *m**e**r-m**a**rin*), and
  - Only before certain suffixes, and only for certain words. New words are generally not affected.
  - But at least it's written!

Implications

- Classic generative linguistics treated both kinds as part of phonology, with hard-wired Grammatical explanations (including restrictions for the latter).
- Meanwhile, newer frameworks of phonology, especially theories of Government Phonology, may go so far as to intentionally exclude anything which is not regular or phonetically natural.
- This debate goes beyond the current presentation, but is something to keep in mind when thinking about which processes we are modelling.



Pilot studies

1. **English** *f*~*v* plurals (e.g., *elf-elves* but *roof-roofs/rooves* and *belief-beliefs*, **believes*)
2. **French** *a(i)l*~*aux* plurals (e.g., *journal-journaux* but *festival-festivals*, **festivaux*)
3. **English** *a*~*an* (e.g., *a book, an apple*)
4. **French** *h-aspiré* (e.g., *l'hôpital, le hibou*), with some variation (e.g., lisaison in *les haricots*)

We'll be focusing on 3 and 4 in this section, glossing over some of the methodology.

Expectations

- While non-normative use of *a* + vowel-initial words has long been noted (Wright, 1905), rates vary drastically according to sociolinguistic factors such as dialect (Orton and Halliday, 1963), race (Labov, 1972) and class (Lass, 2002), as well as stress pattern (Raymond et al., 2002).
- In more traditional, oral corpora, we see ranges of *a* + vowel usage from 5% (Fox, 2015) and 15% (Gabrielatos et al., 2010) up to 90% (Ash and Myhill, 1986).
- As for *h-aspiré*, Moisset (1996) finds application of external sandhi to 13% (92/686) of *h-aspiré* tokens, and higher rates are found by Gabriel and Meisenburg (2009) and Tessier and Jesney (2021), though on much fewer tokens.

General pipeline

Using English *a*∼*an* as an example:

1. Import & process lexicon: nouns & adjectives, categorized by initial segment, stress pattern...
2. Join with frequency data: limit to 20 most frequent words by group
3. Generate queries: exact phrases for "a + [word]", "an + [word]"
4. Collect counts ⟶ gather tweets ⟶ use v1.1 API key to gather device info (optional)
5. Process tweets: geographical data from user files, text from tweet files
6. Analyze, graph, map...

French results

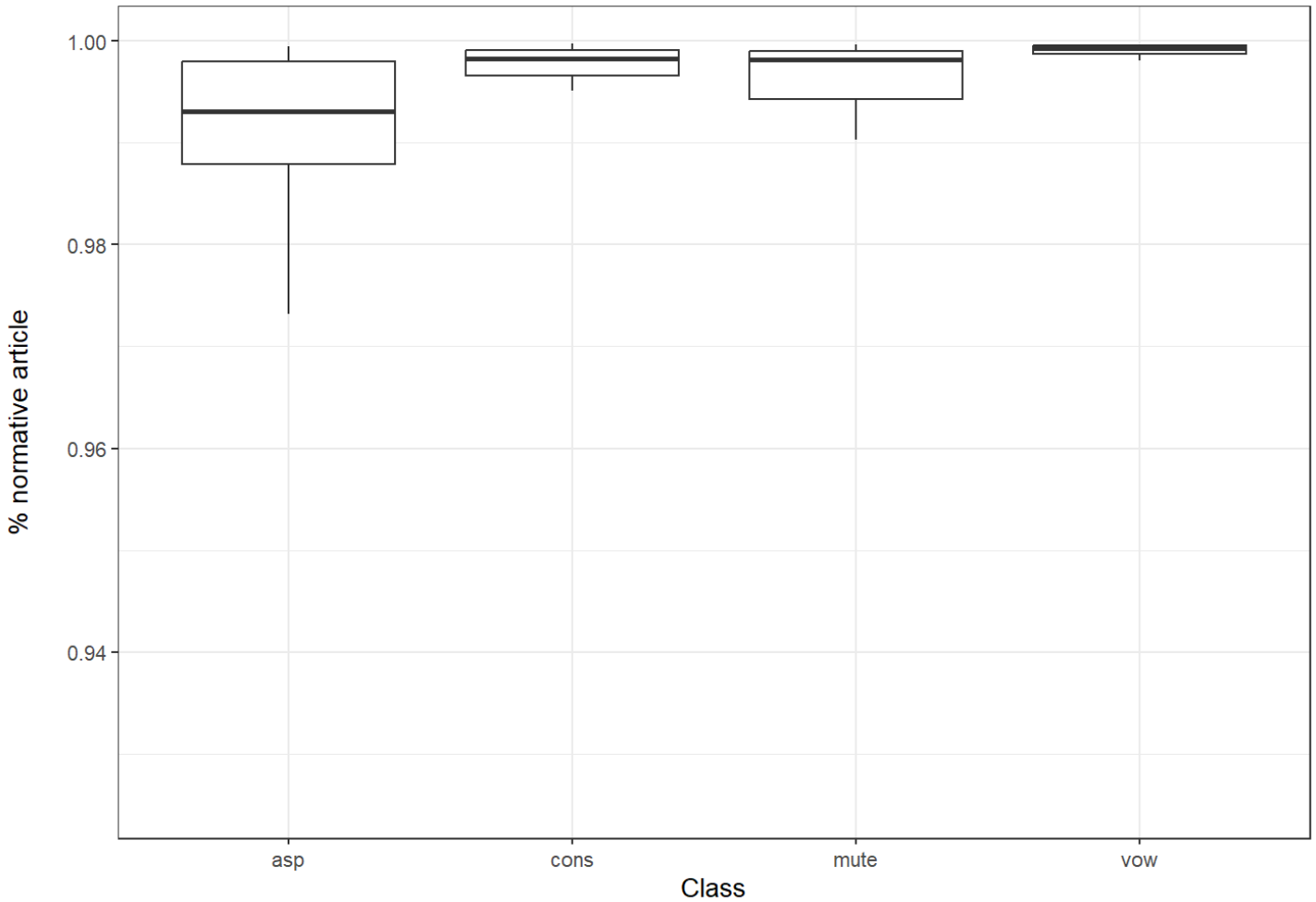


Fig. 6. % use of normative article by word class

English results

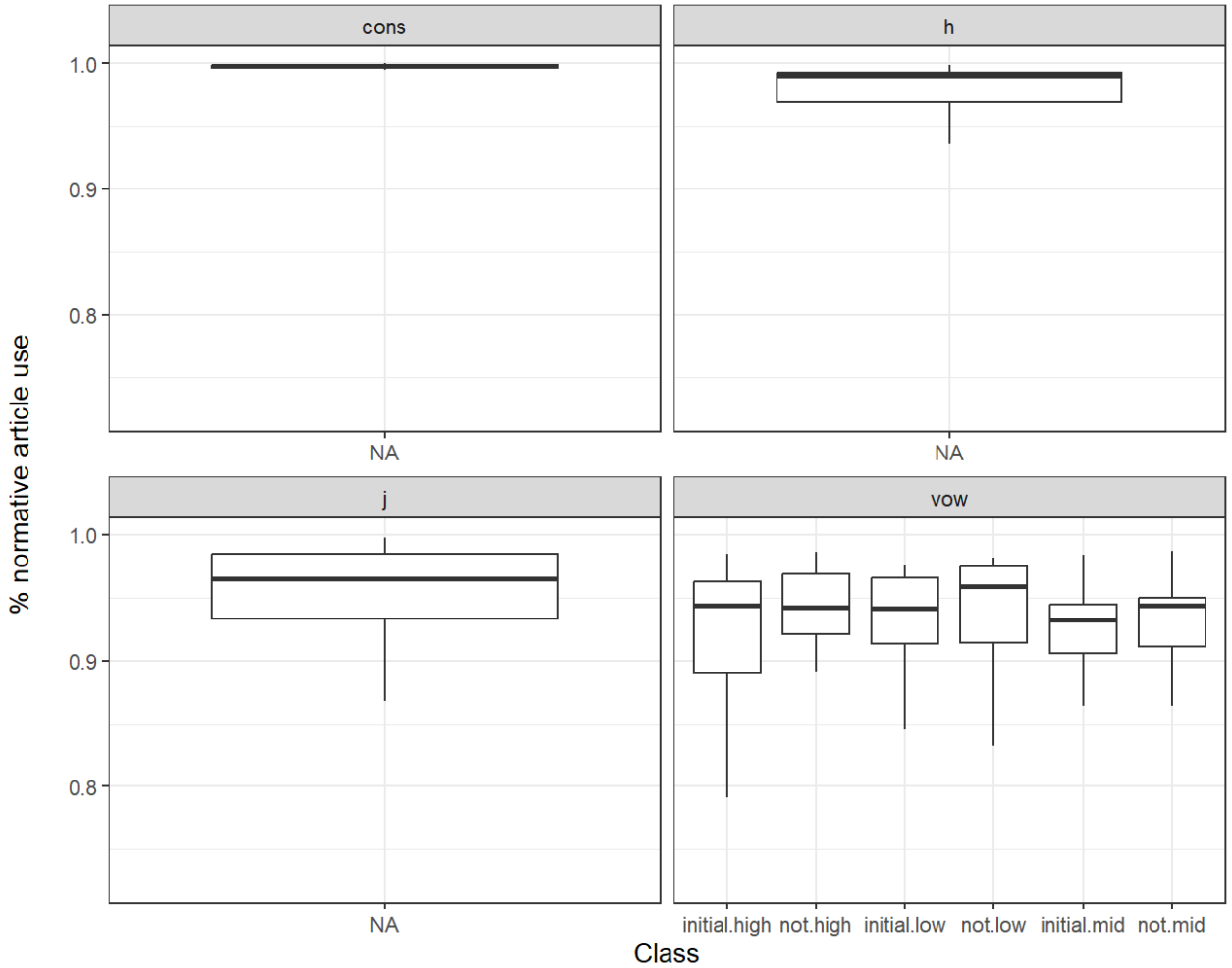


Fig. 7. % use of normative article by initial segment, stress and height

"Help, my results are boring!"

- So what if we plot variables geographically?
- Let's look at usages of *a* + V in the US, at the county level.
- Note that in the following graphs, a suspiciously high count was removed (Concho county, Texas).

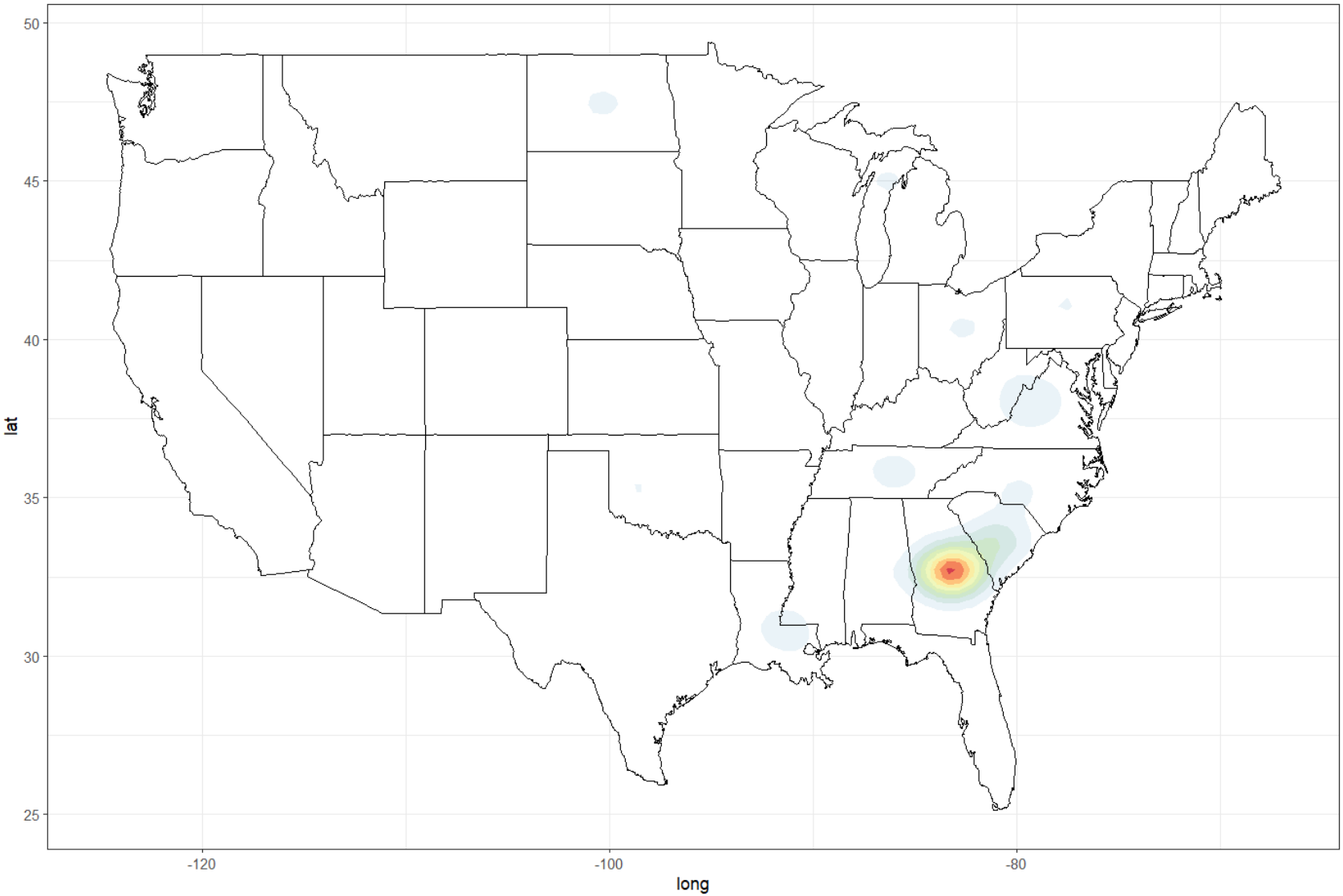


Fig. 8. Heatmap of *a* + vowel tokens per capita

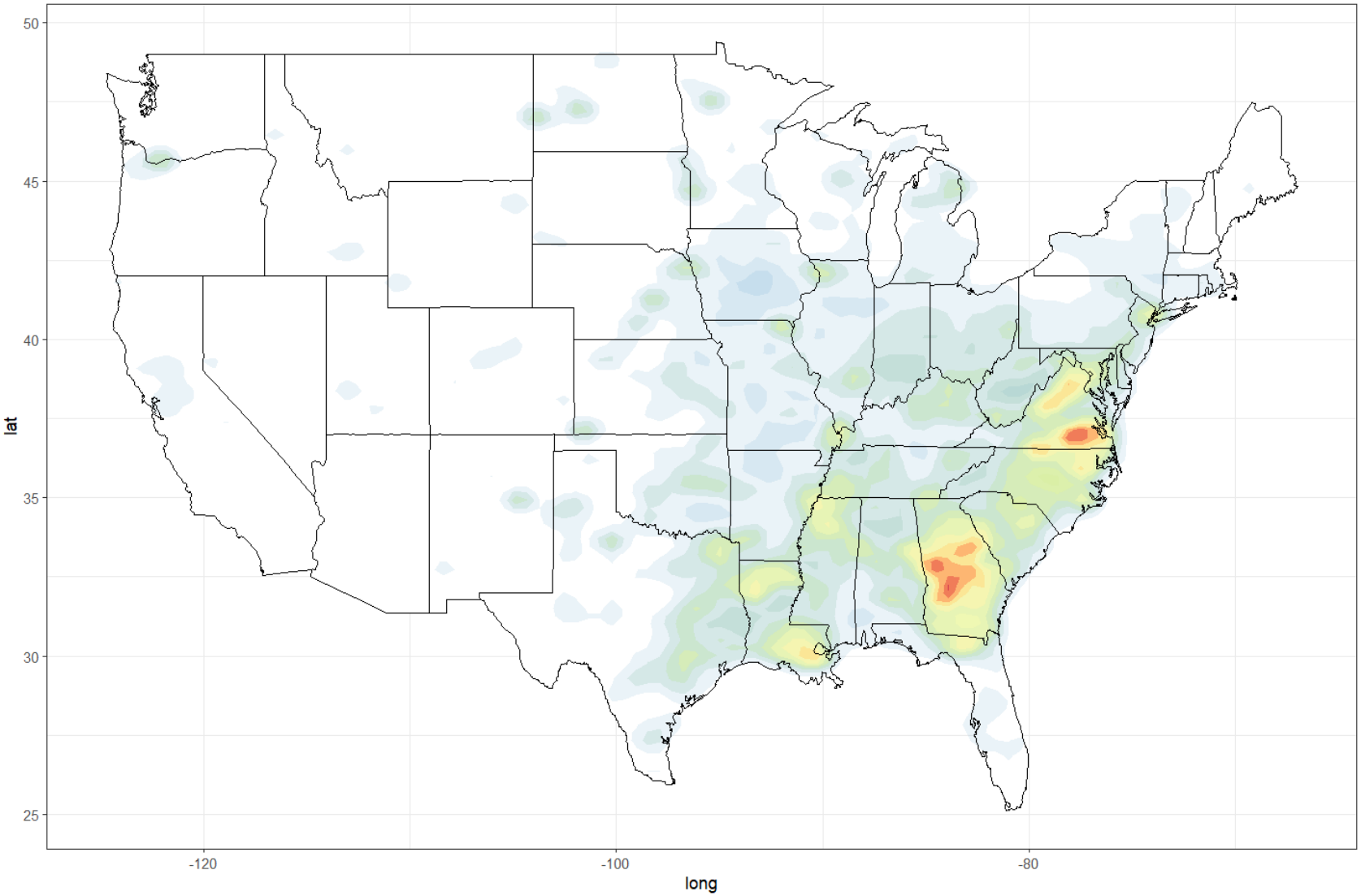


Fig. 9. Heatmap of percentage *a* + vowel tokens

Twitter-MPhon: Future

Branching out

- In the future, we want to look at more phonetically-driven, natural and/or regular phenomena.
- This requires a certain amount of faith, as it requires some sort of awareness of the process as well as a ready (and not-too-variable) way of transcribing the output with the language's orthography.
- Let's look briefly at some preliminary results for one process changing sounds (lenition in Tuscan Italian) and another process removing sounds (*r*-deletion in Brazilian Portuguese).
- The first process also gives us a fun geographical problem to solve!

Tuscan Italian

- Lenition process known as *gorgia toscana*: stop consonants /p, t, k/ commonly turn into close fricative (or less commonly, aspirated) equivalents, generally [ɸ, θ, h], in between vowels (e.g., *la casa* > *la hasa*)
- Documented factors:
  - **Geography**: Most common in North Central area, with Florence and Siena epicentres (Cravens, 2000; Hall, 1949)
  - **History**: /k/ likely affected first, then /t/ and finally /p/ (Izzo, 1972)
  - **Sociolinguistics**: Lenition of /p/ favoured over other consonants regardless of age, class or sex. Lenition increases disproportionately with age (Cravens, 2000)
  - **Structure**: /t/ more likely to pass to [h] in post-tonic position (unstressed syllable following a stressed one) (Russo, 2022)

Queries

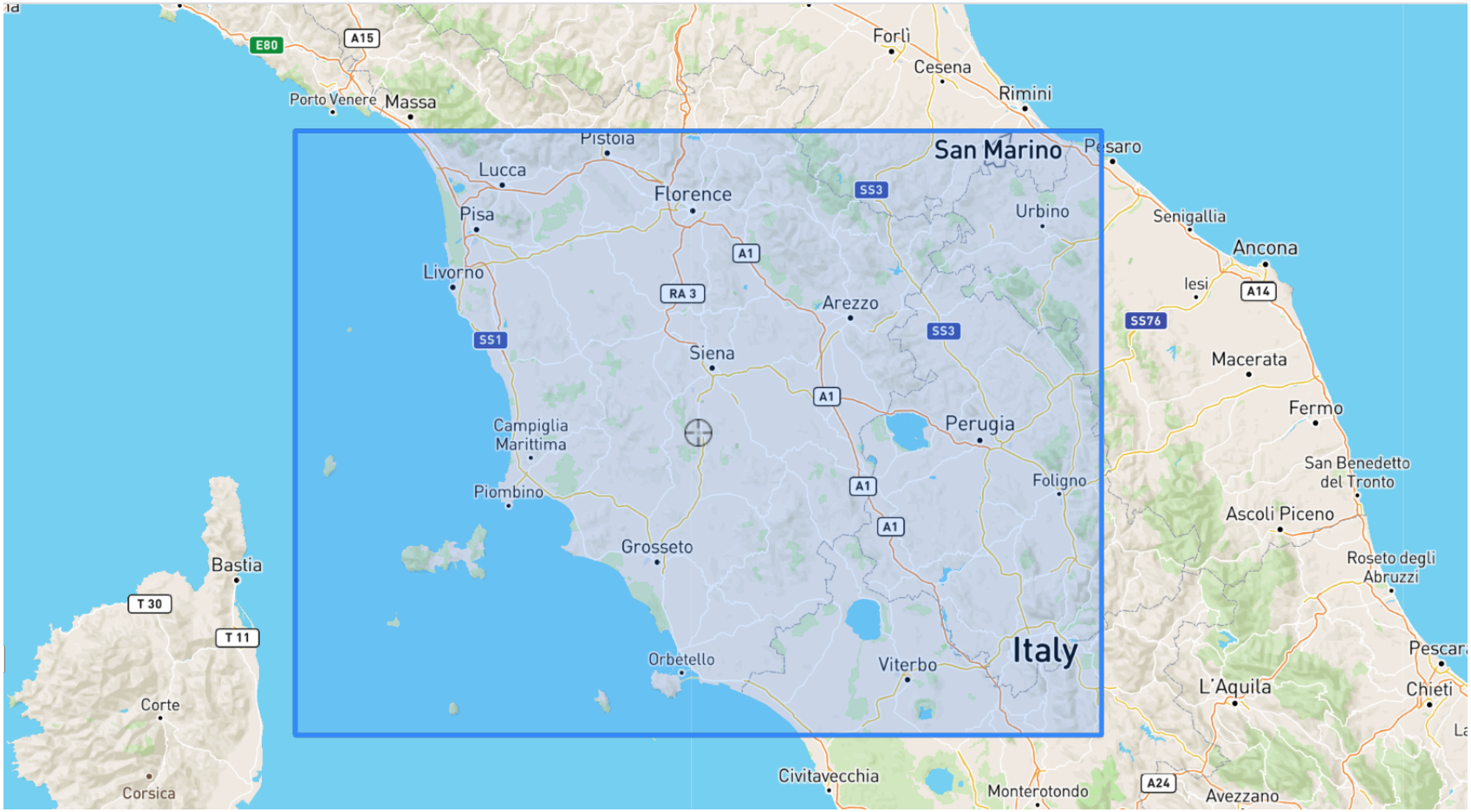
- Words with singleton onset or intervocalic /p, t, k/
- Stress conditions for intervocalic sites:
  - Post-tonic: 'V_V
  - Pre-tonic: V_'V
  - Unstressed: V_V
- Sites in phonetic transcription matched up with orthography & targeted consonant replaced with orthographic lenited variant(s), i.e., $t > th$ and h, $p > f$, $c(h) > h$
- Eliminate any potential competitors (homographs with lenited forms)
- 15 most frequent words selected by consonant and (stress) condition $\longrightarrow$ 110 base words

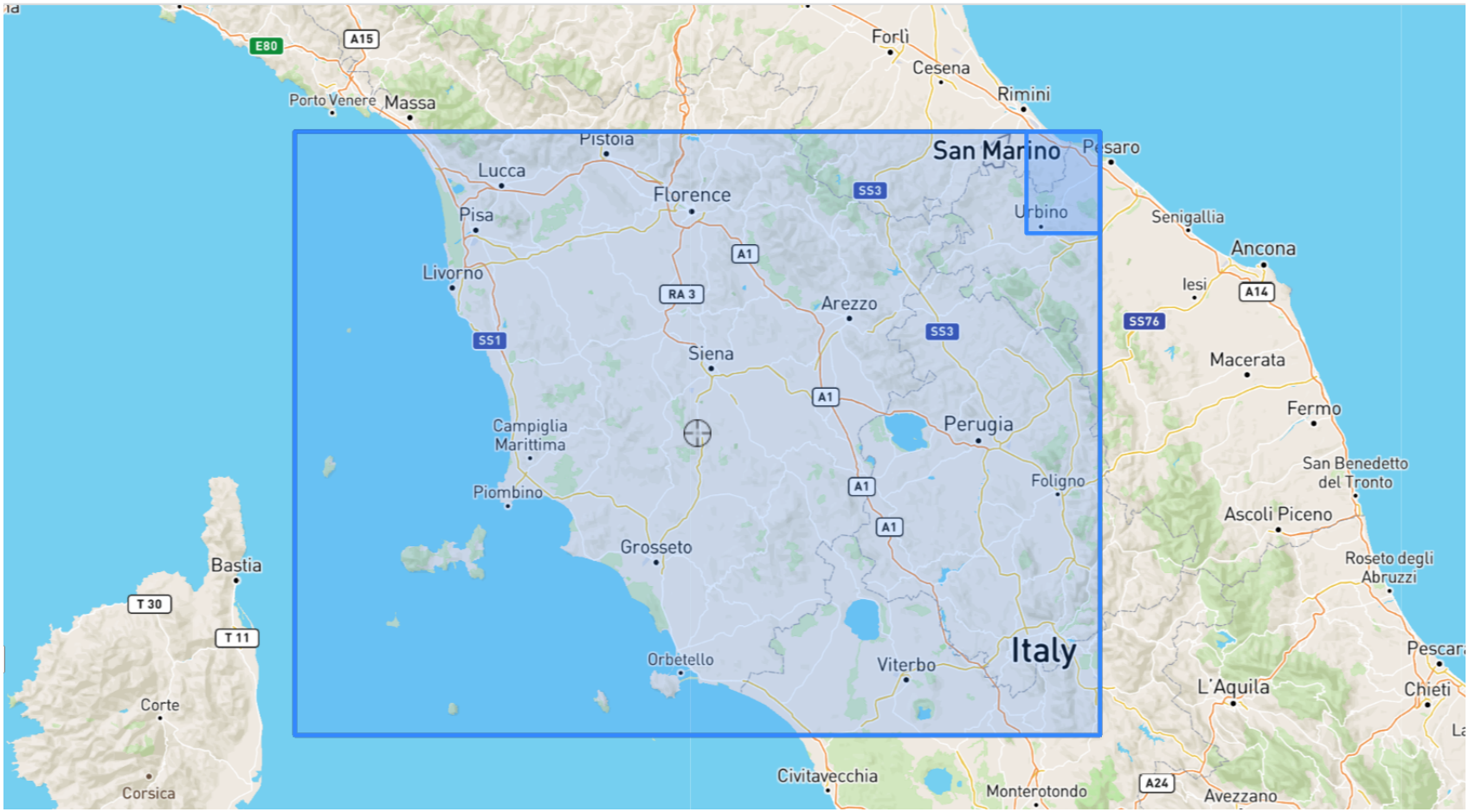
Geography

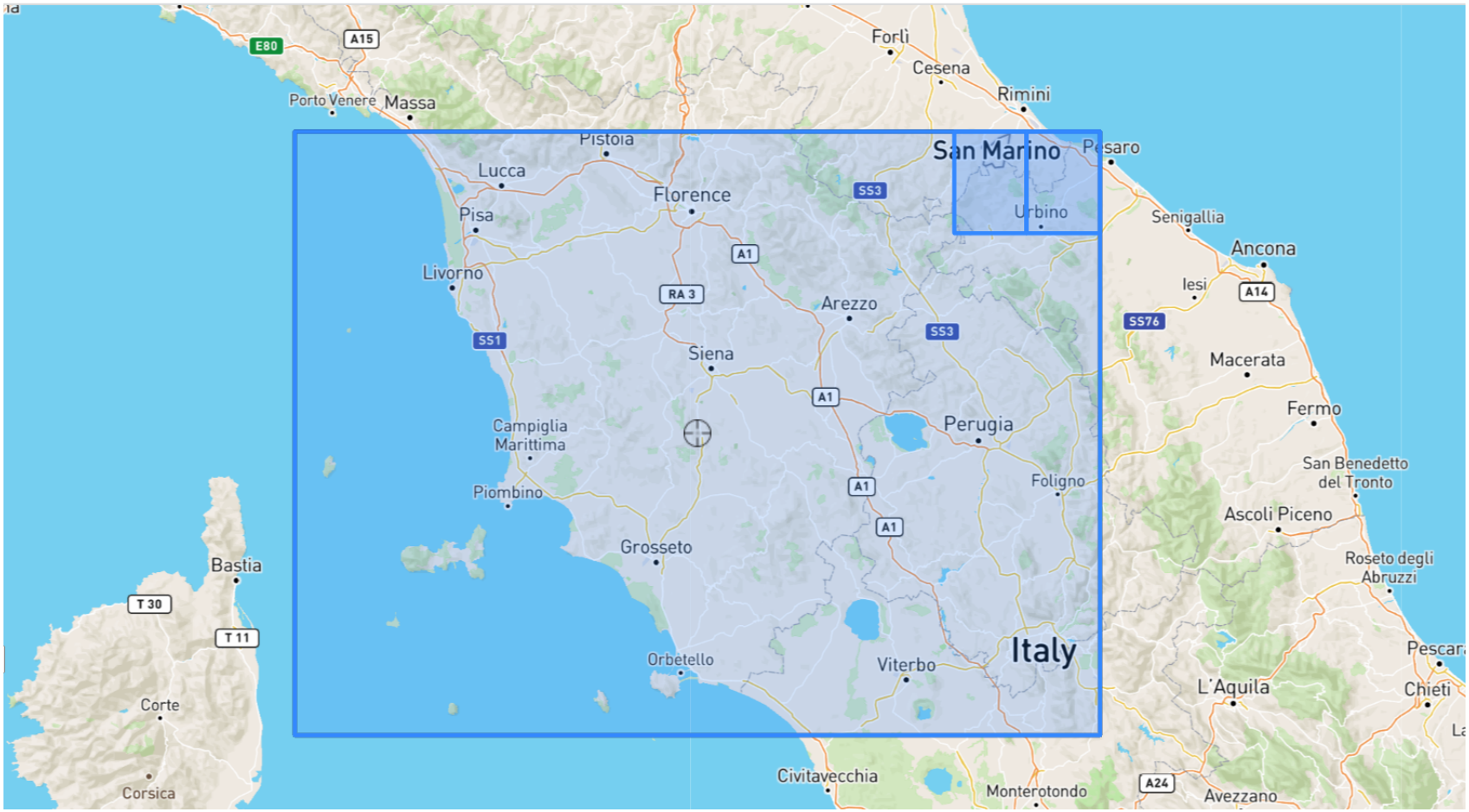
- **Problem**: Intermediate regions not supported by Twitter API, only country or equivalent of town/city/village.
- Two solutions exist:
  - Gather all tweets within Italy as a whole and pare it down from there, or
  - Use 25 × 25-mile bounding boxes to our advantage.
- While the latter is more complicated, the former can be prohibitive depending on the size of the corpus.

Solution

- Get two lat-long coordinates encompassing Tuscany.
- Using conversion formulae between degrees and miles, generate points at the extremities of a 20×20-mile box (just to be safe) in one corner.
- Keep going to generate a grid covering the entire larger box.
- Iterate over queries from box to box.
- Get centroid of boxes from results to pass single lat-long points to map.


E80
A15
Porto Venere
Massa
Forlì
Cesena
Rimini
Pistoia
San Marino
Pesaro
Lucca
Florence
SS3
Pisa
Urbino
Senigallia
A1
Ancona
Livorno
RA 3
Arezzo
Iesi
A14
SS1
SS3
SS76
Siena
Macerata
A1
Fermo
Campiglia
Marittima
Perugia
San Benedetto
del Tronto
Piombino
Foligno
A1
Ascoli Piceno
A1
Grosseto
Roseto degli
Abruzzi
Bastia
T 30
T 11
Orbetello
Viterbo
Italy
Pescara
L'Aquila
Chieti
Corte
Corsica
Civitavecchia
Monterotondo
A24
Avezzano


E80
A15
Porto Venere
Massa
Forlì
Cesena
Rimini
Pistoia
San Marino
Pesaro
Lucca
Florence
SS3
Pisa
Urbino
Senigallia
A1
Ancona
Livorno
RA 3
Arezzo
Iesi
A14
SS76
SS1
SS3
Siena
Macerata
A1
Fermo
Campiglia
Marittima
Perugia
San Benedetto
del Tronto
Piombino
Foligno
A1
Ascoli Piceno
A1
Grosseto
Roseto degli
Abruzzi
Bastia
T 30
T 11
Orbetello
Viterbo
Italy
Pescara
L'Aquila
Chieti
Corte
Corsica
Civitavecchia
Monterotondo
A24
Avezzano


E80
A15
Porto Venere
Massa
Forlì
Cesena
Rimini
Pistoia
San Marino
Pesaro
Lucca
Florence
SS3
Pisa
Urbino
Senigallia
A1
Ancona
Livorno
RA 3
Arezzo
Iesi
A14
SS76
SS1
SS3
Siena
Macerata
A1
Fermo
Campiglia
Marittima
Perugia
San Benedetto
del Tronto
Piombino
Foligno
A1
Ascoli Piceno
A1
Grosseto
Roseto degli
Abruzzi
Bastia
T 30
T 11
Orbetello
Viterbo
Italy
Pescara
L'Aquila
Chieti
Corte
Corsica
Civitavecchia
Monterotondo
A24
Avezzano

Results

- Noticeably small number of tweets (133)
- Let's go to an interactive map...
- Factors passed to map:
  - Relative frequency (relfreq)
  - Consonant (cons)
  - Stress condition (type)
  - Consonant × stress interaction (int)

Grain of salt: A point isn't necessarily a genuine usage – some may be metalinguistic commentary!

Già si sente l'accento toscano hoha hola halda ho la hannuccia horta horta pel mi amiho maremma maiala

Discussion

- Clustering of results around Florence & Siena
- Preference for /k/ lenition & initial position
- Dispreference for V_'V position
- Seemingly little effect of relative word frequency

Brazilian Portuguese

- Word-final deletion of r is characteristic of Brazilian Portuguese.
- The process is known to be coded for socioeconomic class, though in areas such as the North, it is less stigmatized. (De Oliveira, 1983)
- There is also precedent for representation of this phenomenon in writing. (Mateus and d'Andrade, 2000)

Methodology

- All words with orthographic final *r* were selected, and only words without *r*-less lexical competitors (as entries or inflected forms) were kept.
- Started by counting tokens with final *r* and without in both Portugal and Brazil for the period of 01/2007-01/2023. Brazil clearly had higher percentage of *r*-less tokens for many words.
- Set as queries top 50 words in Brazilian counts, *r*-less variants only. This yielded nearly 865,000 tweets.
- Device was gathered. Text was processed for following context (punctuation, vowel, various consonant manners).

Results

- The top 10 device sources account for 99.94% of the data, of which mobile devices represent 99.93%.
- 32% of *r*-deletion occurred before a juncture (punctuation or "tweet edge"), followed by 27% deletion before stop consonants.
- Pre-vocalic deletion was surprisingly high, at 21%, followed by fricatives (10%), nasals (8%) and other liquids (3%).
- Geography still needs some tinkering, but results suggest higher rates in centre and northwest.

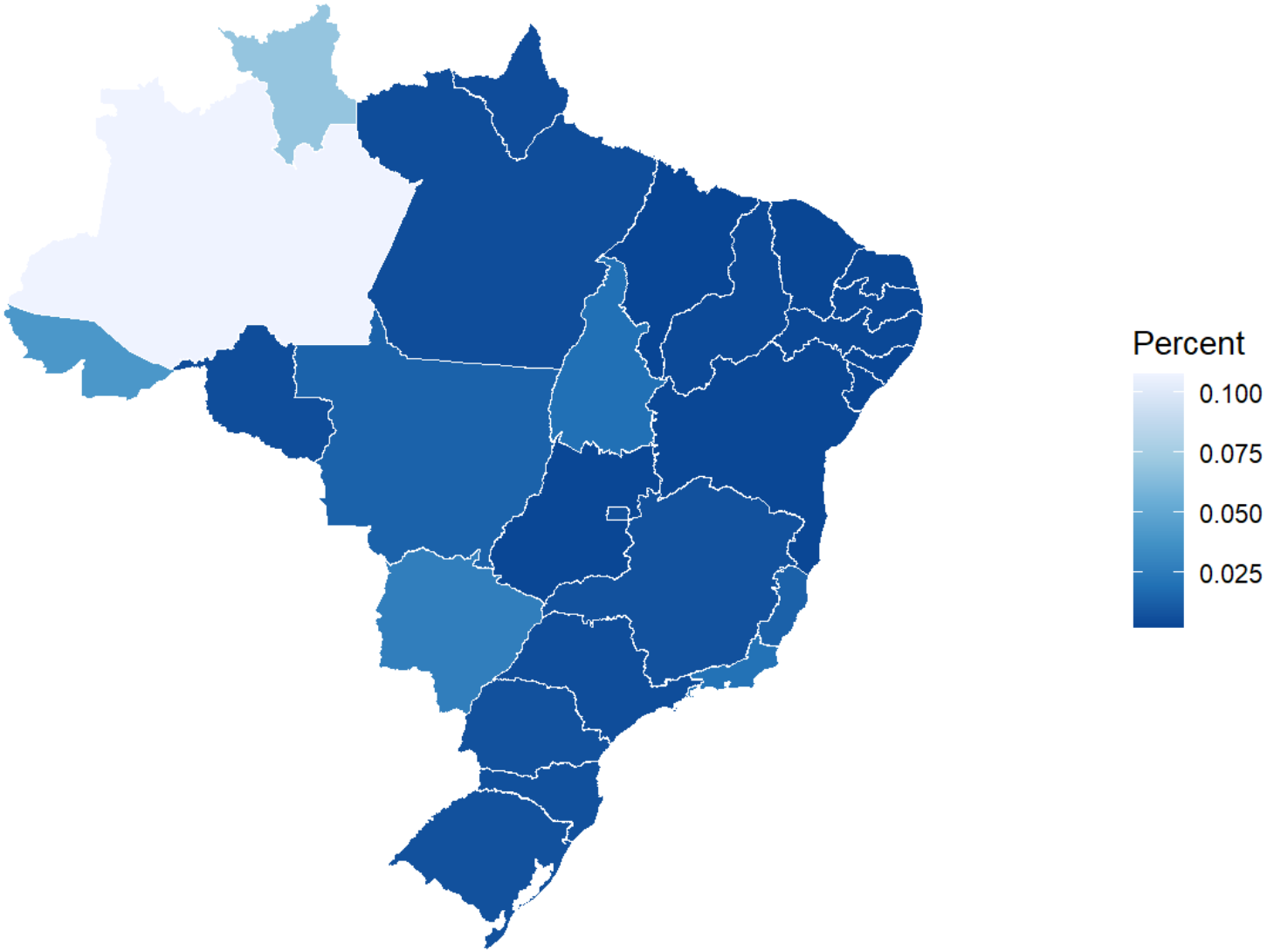


Fig. 10. Number of *r*-deletions per capita, by state

Discussion

- Certain types of processes (featural) may simply not be written as much as others (insertion/deletion).
- Cultural and/or sociolinguistic factors may also influence which phenomena are tractable.
- There's a lot of research on non-standard/oral-like orthography (e.g., volume 4/4 of *Journal of Sociolinguistics*, 2000), as well as research specific to social media (Eskander et al., 2013; Pottinger, 2021).

General discussion

So what?

- For getting tractable results, I doubt it's as easy as collecting counts—even with linguistic variables involved—calculating relative percentages and comparing with traditional corpora.
- Twitter data are presumably going to be much more "big tent" than an average survey, experiment, field corpus, etc. Appropriate scope should be kept in mind.
- It seems better to:
  - Compare geographic regions and/or time frames,
  - "Zoom in" on specific geographic regions, or
  - Focus on specific types of people (will likely need to be filtered *post hoc* from user descriptions).

Making sense of it all

- I imagine a sort of pipeline of on-the-ground observations and big data feeding each other, for instance:
  - People noticed fairly quickly about the difference between North American French and European French when it comes to the gender of "COVID".
  - Not many people in Western media were talking about other varieties of French, e.g., those spoken in Africa. Twitter data makes this easy to observe.
  - Twitter data help us see the rise in the feminine over time and, with some additional analysis, we could see who figures among the important actors in this change.
  - We can't probe the attitudes behind these evolving phenomena as easy – we could try, but something more sure would be survey data directly linking gender usage in the early days of the pandemic with attitudes towards local language authorities, for instance.

Taking it further

- We can test hypotheses that we can incorporate into experimental design. These questions have ready application to linguistic theory.
- Structural complexity or size (whether syllable contact or syntactic constituent), number of syllables playing into BP r-deletion – we can imagine judgment tasks, for instance.
- Internal vs. external domains, prominence effects (esp. stressed vs. unstressed syllable onset faithfulness) in Tuscan lenition, other effects like markedness (k vs. p, t) and faithfulness (word-initial vs. word-internal)
- We can test how variation in the gender of neologisms or borrowings in (presumably mostly) L1 speech mirrors what we know about variation in gender in borrowings in bilingual speech

Limitations

- Geographical information is very precise but only represents where tweet was sent from, or where user identifies as home.
- Extracting other common sociolinguistic metrics such as gender, race and age remains difficult and inaccurate for the time being.
- Without easy access to device data, autocorrect should be kept in mind as a confounding factor.
- Tokens are impressionistically more likely to be more spread out over users (i.e., fewer tokens per speaker).

Recommendations: Academic

- Twitter data especially shine in the chronological study of recent and sudden phenomena (like COVID and "pronouns").
- Marginal or variable phenomena are likely to be just as much so if not more in Twitter data, when compared to normative forms. Imagine additional angles (whether in Twitter or not) from the beginning.
- Qualitative analysis must be integrated throughout and not be sidelined.

Recommendations: Practical

- If interested, get an academic API key (or have someone else get one) and collect your data **ASAP**.
- Ensure ample storage before downloading tweets metadata. 10M tweets will likely run around 20 Gb for both tweets and users.
- Even with an impressive machine, you can't expect to import all your data at once and explore it. If possible, import bit by bit, process, save, then compile.
- Save often, maybe even after each iteration, because your cat *will* walk over your computer's power button 3/4 of the way into a 10-hour process.

Important resources

- Wiktionary data (kaikki.org): Includes orthographic form, IPA transcription, POS, etymological information, inflected forms, and so much more.
- Sketch Engine: Word (relative) frequency data, including online corpora.
- `academictwitteR`: For gathering counts, tweets and user information, etc. Has intelligent and quasi-foolproof handling of API refresh rates. Also has commands for gathering files which may be sufficient. Requires academic level access.
- `jsonlite`: Useful for manipulating Wiktionary data and raw tweet/user files.
- `tidyr`: Especially useful for unpacking nested lists in lexicon and data files (`unnest` family).
- Regular expressions, `stringr`/`stringi`: For manipulating text.
- Various mapping packages are available, some for data and visualization. The main ones used here are `maps`, `sf` and `mapview` (for interactive maps).
- `for` loops and/or the `apply` family: For iterating over large number of files or rows.

Thank you!

Acknowledgements

- Many thanks to my colleagues cited here and to my students Chang Chen, Nathan Samon, Yutaka Suzuki, Ariel Sosic, Georges Awaad.
- This work was financed by the Social Sciences and Humanities Research Council - Insight Development grants as well as internal funds from the Université de Montréal.

Works Cited I

Académie française. 2020. Le covid 19 ou la covid 19. Fiche terminologique. `http://www.academie-francaise.fr/le-covid-19-ou-la-covid-19`.

Anttila, Arto. 1997. Deriving variation from grammar. In *Variation, change, and phonological theory*, ed. Frans L. Hinskens, Roeland van Hout, and W. Leo Wetzels. John Benjamins, 35–68.

Ash, Sharon and John Myhill. 1986. Linguistic correlates of inter-ethnic contact. *Diversity and diachrony* 53: 33–44.

Avanzi, Mathieu. 2020. Le/la covid ? Réouvrir ou rouvrir ? Les leçons de grammaire du coronavirus. *The Conversation* URL `https://theconversation.com/le-la-covid-reouvrir-ou-rouvrir-les-lecons-de-grammaire-du-coronavir`

Ayewa, Kouassi Noël. 2009. Une enquête linguistique: le français, une langue ivoirienne. *Le français en Afrique* (25): 117–134.

Belleau, Rémi. 2016. *Attribution et variation du genre d'emprunts à l'anglais, à l'italien, au japonais et à l'arabe dans le lexique du français*. Master's thesis, Université Laval.

Works Cited II

Biers, Kelly Biers, Michael Dow, Eric Beuerlein, Kimaya Guthrie, Sam McIntosh, and Heather Roberts-VanSickle. 2023. Of prounoun havers and havers-not: The semantic shift of 'pronoun' in online political discourse. Blooming: Metamorphoses and Seasons of Queerness: 2023 UNCA Queer Studies Conference.

Biloa, Edmond. 2003. *La langue française au Cameroun: analyse linguistique et didactique*. Peter Lang.

Boucher, Karine and Suzanne Lafage. 2000. *Le lexique français du Gabon: entre tradition et modernité*. Institut de linguistique française.

Boutin, Akissi. 2007. Déterminant zéro ou omission du déterminant en français de Côte d'Ivoire. *Le français en Afrique* 22: 161–182.

Calvet, Maurice and Pierre Dumont. 1969. Le français au Sénégal: interférences du wolof dans le français des élèves sénégalais. *Collection IDERIC* 7(1): 71–90.

Carroll, Susanne. 1989. Second-language acquisition and the computational paradigm. *Language Learning* 39(4): 535–594.

Works Cited III

Chalier, Marc. 2018. Quelle norme de prononciation au Québec? Attitudes, représentations et perceptions. *Langage et société* 163(1): 121–144.

Chalier, Marc. 2019. La norme de prononciation québécoise en changement (1970-–2008)? L'affrication de /t, d/ et l'antériorisation de /$\tilde{a}$/ chez les présentateurs des journaux télévisés de Radio-Canada. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 64(3): 407–443. 407.

Chen, Emily, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance* 6(2): e19273.

Coetzee, Andries W. 2016. A comprehensive model of phonological variation: Grammatical and non-grammatical factors in variable nasal place assimilation. *Phonology* 33(2): 211.

Corbett, Greville G. 1991. *Gender*. Cambridge University Press.

Cravens, Thomas D. 2000. Sociolinguistic subversion of a phonological hierarchy. *Word* 51(1): 1–19.

Works Cited IV

De Oliveira, Marco Antônio. 1983. *Phonological variation and change in brazilian portuguese: the case of the liquids*. University of Pennsylvania.

Dell, François and Elisabeth Selkirk. 1978. On a morphologically governed vowel alternation in french. *Recent transformational studies in European languages* 1: 51.

Dow, Michael and Patrick Drouin. in press. Tracing the evolution of the gender of "covid-19" in the french of three continents: A traditional and social media study. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 68(3).

Dow, Michael, François Lareau, and Patrick Drouin. upcoming. Twitter-mphon: Studying morphophonological variation with twitter data. 2023 Annual Conference of the Canadian Linguistics Association.

Eskander, Ramy, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing spontaneous orthography. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 585–595.

Works Cited V

de Féral, Carole. 2006. Étudier le camfranglais: recueil des données et transcription. *Le français en Afrique* 21: 211–218.

Fox, Susan. 2015. *The New Cockney: New ethnicities and adolescent speech in the traditional East End of London*. Palgrave Macmillan.

Gabriel, Christoph and Trudel Meisenburg. 2009. Silent onsets? An optimality-theoretic approach to French *h-aspiré* words. *Variation and Gradience in Phonetics and Phonology* : 163–184.

Gabrielatos, Costas, Eivind Nessa Torgersen, Sebastian Hoffmann, and Susan Fox. 2010. A corpus-based sociolinguistic study of indefinite article forms in London English. *Journal of English Linguistics* 38(4): 297–334.

Gimenes, Manuel, Cyril Perret, and Boris New. 2020. Lexique-infra: Grapheme-phoneme, phoneme-grapheme regularity, consistency, and other sublexical statistics for 137,717 polysyllabic French words. *Behavior Research Methods* .

Haden, Ernest F and Eugene A Joliat. 1940. Le genre grammatical des substantifs en franco-canadien empruntés à l'anglais. *Publications of the Modern Language Association of America* 55(3): 839–854.

Works Cited VI

Hall, Robert A. 1949. A note on" gorgia toscana". *Italica* 26(1): 65–71.

Herault, Georges and Jean-Paul Vonrospach. 1967. *Étude phonétique et syntaxique du français d'élèves de cours préparatoire de la région d'Abidjan*, vol. 1. Institut de linguistique appliquée d'Abidjan.

Holmes, Virginia M and Juan Segui. 2004. Sublexical and lexical influences on gender assignment in French. *Journal of Psycholinguistic Research* 33(6): 425–457.

Holtzer, Gisèle. 2004. Savoirs et compétences en français écrit d'élèves guinéens: les enquêtes campus (1998-2001). *Le français en Afrique* 19: 35–73.

Izzo, Herbert J. 1972. *Tuscan and etruscan: the problem of linguistic substratum influence in central italy*. University of Toronto Press.

Jabet, Marita. 2006. Noms sans déterminant en français abidjanais: trait sociolinguistique, sémantique et/ou pragmatique? *Le français en Afrique* 21: 325–337.

Kaisse, Ellen M and Patricia A Shaw. 1985. On the theory of Lexical Phonology. *Phonology Yearbook* 2: 1–30.

Works Cited VII

Kim, Minchai. 2017. *Variation terminologique en francophonie : Élaboration d'un modèle d'analyse des facteurs d'implantation terminologique*. Ph.D. thesis, Université Paris Sorbonne.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* : 715–762.

Labov, William. 1972. *Language in the inner city: Studies in the Black English vernacular*. 3. University of Pennsylvania Press.

Lass, Roger. 2002. South African English. *Language in South Africa* 104126: 104–126.

Léard, Jean-Marcel. 1995. *Grammaire québécoise d'aujourd'hui: comprendre les québécismes*. Guérin universitaire.

Lupu, Mihaela. 2005. La masculinisation du lexique français: le rôle catalyseur des anglicismes. *Analele Universității „Alexandru Ioan Cuza" din Iași. Secțiunea IIIe. Lingvistică* .

Lyster, Roy. 2006. Predictability in french gender attribution: A corpus analysis. *Journal of French Language Studies* 16(1): 69–92.

Works Cited VIII

Mateus, Maria Helena and Ernesto d'Andrade. 2000. *The phonology of portuguese*. OUP Oxford.

Maurais, Jacques. 2008. *Les Québécois et la norme l'évaluation par les Québécois de leurs usages linguistiques / jacques maurais*.

Moisset, Christine. 1996. The status of 'h aspiré' in French today. *University of Pennsylvania Working Papers in Linguistics* 3(1): 17.

Ndjerassem, Mbai-Yelmia Ngabo. 2005. *Le français au tchad*. UFR Lettres, Arts et sciences humaines.

Nymansson, Karin. 1995. Le genre grammatical des anglicismes contemporains en français. *Cahiers de lexicologie* 66(1): 95–113.

Nzesse, Ladislas. 2009. *Le français au cameroun: d'une crise sociopolitique à la vitalité de la langue française (1990-2008)*. UFR Lettres, Arts et sciences humaines.

O'Mahony, Jennifer. 2019. Why the future of French is African. *BBC News* URL `https://www.bbc.com/news/world-africa-47790128`.

Works Cited IX

Orton, Harold and Wilfrid J. Halliday. 1963. *The survey of English dialects, volume 1: The six northern counties and the Isle of Man*. Leeds, UK.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Pöll, B. 2005. *Le français langue pluricentrique?: études sur la variation diatopique d'une langue standard*. P. Lang.

Poplack, Shana. 2018. *Borrowing: Loanwords in the speech community and in the grammar*. Oxford: Oxford University Press.

Poplack, Shana, Alicia Pousada, and David Sankoff. 1982. Competing influences on gender assignment: Variable process, stable outcome. *Lingua* 57(1): 1–28.

Pottinger, Hamish. 2021. Language standards in an unstandardised language: The orthographies and ideologies of scots users on twitter. *Journal of Languages, Texts, and Society* 5: 1–36.

Works Cited X

Pustka, Elissa, Jean-David Bellonie, Marc Chalier, and Luise Jansen. 2019. C'est toujours l'autre qui a un accent : Le prestige méconnu des accents du Sud, des Antilles et du Québec. *Glottopol* 31: 27–52.

Raymond, William D, Julia A Fisher, and Alice F Healy. 2002. Linguistic knowledge and language performance in English article variant preference. *Language and Cognitive Processes* 17(6): 613–662.

Roché, Michel. 1992. Le masculin est-il plus productif que le féminin? *Langue française* (96): 113–124.

Russo, Michela. 2022. Locality domains on lenition. spirantization (gorgia) and voicing in tuscan dialects. *Linx. Revue des linguistes de l'université Paris X Nanterre* (84).

Schmidt, Jean. 1990. Panorama des emprunts à l'anglais dans le français d'Afrique. *Bulletin du réseau des observatoires du français contemporain en Afrique noire* (7): 1987–88.

Works Cited XI

Sebková, Adéla, Kristin Reinke, and Suzie Beaulieu. 2020. À la rencontre des voix francophones dans la ville de Québec : les attitudes des Québécois à l'égard de diverses variétés de français. *SHS Web Conf.* 78: 02002.

Spolsky, Bernard. 2018. Language policy in French colonies and after independence. *Current Issues in Language Planning* 19(3): 231–315.

Surridge, Marie E. 1993. Gender assignment in French: The hierarchy of rules and the chronology of acquisition. *IRAL-International Review of Applied Linguistics in Language Teaching* 31(2): 77–96.

Telep, Suzie. 2014. Le camfranglais sur internet: pratiques et représentations. *Le français en Afrique* (28): pp–27.

Tessier, Anne-Michelle and Karen Jesney. 2021. Learning French liaison with gradient symbolic representations: Errors, predictions, consequences. In *Proceedings of the Annual Meetings on Phonology*, vol. 8.

Works Cited XII

Tremblay, Louis. 1994. *Convergence et divergence dans l'emploi de termes communs recommandés par l'office de la langue française*. Master's thesis, Université Laval.

Tucker, G. Richard, Wallace E. Lambert, and André A. Rigault. 1977. *The French speaker's skill with grammatical gender: An example of rule-governed behavior*, vol. 8. Walter de Gruyter GmbH & Co KG.

World Health Organization. 2020. Prévention et contrôle des infections (PCI) pour le virus de la COVID-19. `https://openwho.org/courses/COVID-19-IPC-FR`.

Wright, Joseph. 1905. *The English dialect grammar*. Oxford University Press.