Twitter-MPhon: Studying morphophonological variation with Twitter data

Michael Dow, François Lareau & Patrick Drouin
Université de Montréal

Annual conference of the Canadian Linguistics Association

June 2, 2023

Introduction

Introduction (as of 2021)

- Twitter's Academic API access tier: a free, amazing and extensive data source for researchers.
- Starting to be exploited by linguists but relatively underused in phonology, compared to other areas (but see, e.g., Lamontagne and McCulloch, 2022, plus work by Kie Zuraw, Rachael Tatman, etc.).

Let's develop a general and open-source methodology for studying phonological phenomena in any of the 30+ languages supported by Twitter's in-house language detection, from generating queries to measuring variation to plotting variables geospatially!

Introduction (as of 2023)

Introduction (as of 2023)

it-it- the f - it -flam - flames.
Flames, on the side of my face...

Fig. 1. Breathing—breathl—heaving breaths. Heaving breaths. . . Heathing

This is a cautionary tale.

- Recent changes to Twitter's management make its future as a data source untenable for academics.
- Today, we're focusing on pilot data from English and French, but we've gathered as much data as possible from many languages.
- One of our three API keys still works (*shhh!*), but we are anticipating needing to move on.
- Generic, preliminary scripts for our workflow are available at `https://github.com/mcdowlinguist/twitter-mphon`.

Roadmap

1. Introduction
2. Background
3. Phenomena
4. Methodology
5. Results
6. Discussion

Background

Project philosophy

- Emphasis on open-access, non-language-specific resources
- Variation is key!
- Starting hypothesis that morphophonological phenomena would be better attested and easier to target, testing:
  - French: Definite article elision before *h-aspiré* words
  - English: Indefinite article $a \sim an$ alternation.
- Both phenomena are fairly well-studied from a quantitative standpoint, allowing for comparison.

Basics of Twitter

- Social media "microblogging" platform offering different types of posts (original tweets, replies, retweets, and quote tweets)
- Prior to the end of 2022...
  - Accounts numbered approximately 450 million, with an estimated 500 million tweets sent per day.
  - Users are distributed fairly normally for age, (with 25-34 representing largest demographic), are overwhelmingly male but (in US) evenly distributed for race.
  - Democratic-leaning accounts in the US produced the highest amount of content.
- We exploited the Academic access API tier provided by Twitter till spring 2023 for free data access.
- We focus on original text tweets, ignoring media and links.

Things we can do

- 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., retweets, 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

Phenomena

French: *h-aspiré*

	Type	Example	Elision?
a.	*h-aspiré*: <h>	*hibou* ‘owl’	*le hiboux* [lə.(ʔ)ibu]
b.	*h-aspiré*: other	*yacht* ‘yacht’	*le yacht* [lə.jɔt]
c.	*h-muet*	*hôtel* ‘hotel’	*l’hôtel* [lɔ.tɛl]
d.	semivowel	*oie* ‘goose’	*l’oie* [lwa]

Table 1: Definite article before various word types

- *h-aspiré*: word phonetically begins with a vowel but exceptionally blocks external sandhi.
- *h-muet*: word begins with <h> (orthographic) and a vowel (phonetic) but does not block sandhi.
- Compare with consonant-initial words (no elision) and vowel-initial words (regular elision)

English: *a* ∼ *an*

- The prescriptive distribution of each allomorph is a before consonant-initial words and an before vowel-initial words, without lexical exception (e.g., *a book*, *an apple*).
- Words with an orthographic initial <h> may vary with respect to the presence of [h] in its phonetic form and thus the article it selects (e.g., *historic*).

Previous studies: French

- Experimental and corpus studies on *h-aspiré* finds variable but generally low rates of external sandhi.
- Moisset (1996) finds gradation among 18 speakers, with aggregated application of external sandhi for 13% (92/686) of *h-aspiré* tokens.
- Application of sandhi higher for lower class, and some word-level effects obtain (esp. *handicapé*).
- Higher rates are found by Gabriel and Meisenburg (2009) and Tessier and Jesney (2021), though on fewer tokens.

Previous studies: English

- Non-normative use of *a* + vowel-initial words has long been documented (Wright, 1905), especially in English (UK) dialectology.
- 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 these 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), depending on these variables.

Methodology

Queries: General

- Same general process for generating queries in both languages:
  1. Downloading the Wiktionary lexicon from kaikki.org,
  2. Merging with relative frequency data (obtained from SketchEngine), and
  3. Applying lexical and phonological filters.
- Additional subclasses were created and within each subclass, the n most frequent words were selected.
- For each word, a normative and non-normative variant provided our queries.

Queries: French

- Words limited to singular nouns, classified based on the initial segment of IPA transcription (orthographic <h>, vowel (without <h>), semivowel, and consonant).
- Semivowel- and <h>-initial words were further divided according to their “aspiration” behaviour from the relevant Wiktionary field and manually verified with the Multidictionnaire de la langue française (De Villers, 2009).
- Selected the maximum 25 most frequent words within each category combined them with a full definite determiner and the elided form (e.g., *le système*, **l'systéme* ‘the system’).
- Yielded 135 words, or **270 queries**.

Queries: English

- Words limited to singular nouns and adjectives, separated according to the initial segment in IPA transcription : /h/, /j/, other consonants, and vowels.
- Stress of the initial syllable for vowel-initial words was extracted and given a binary code.
- Initial vowels were further classified according to height (high, mid and low).
- The maximum 20 most frequent words according to initial segment type, stress pattern and vowel height were selected.
- Each of the two forms of the indefinite article was combined with each word (e.g., **a effort*, *an effort*).
- Yielded 392 words, or **784 queries**.

Data Collection & Processing

- We performed data collection in R using the `academictwitteR` package (Barrie and ting Ho, 2021).
- For each query, the relevant language was specified, and each expression was queried as an exact phrase. Retweets were excluded.
- The time frame specified was April 2006 until December 2022, with no limit on number of pages returned.
- Counts and percentages of non-normative to normative variants were gathered at various levels and plotted as boxplots.

Mapping

- We also gathered tweets and metadata for a subset of the 50 most variable vowel-initial English words in the US which contained geographic information.
- The centroids of bounding boxes provided were calculated, and using a variety of packages, each tweet's coordinates were matched to a county and processed as geometric objects.
- Percentage of *a* + vowel-initial (vs. *an*) tokens were performed within each county and divided by its population to provide a density score, then plotted as a heatmap with the `stat_density2d` function from the `ggplot2` (Wickham, 2016) package, passing density to a gradient.

Results

Results

- Aggregated use of normative forms is high for both phenomena, with some exceptions.
- No signs of a word frequency effect in either language.
- In French, only a handful of plausible words like *harcèlement* and *handicap* showed rates of the full article beneath 95%.
- Of the vowel-initial English words, mid vowels may stand out as having more *a* tokens, but appears minimal.
- The middle of Georgia stands out as the most prominent location for a + V per capita in the US.

French results

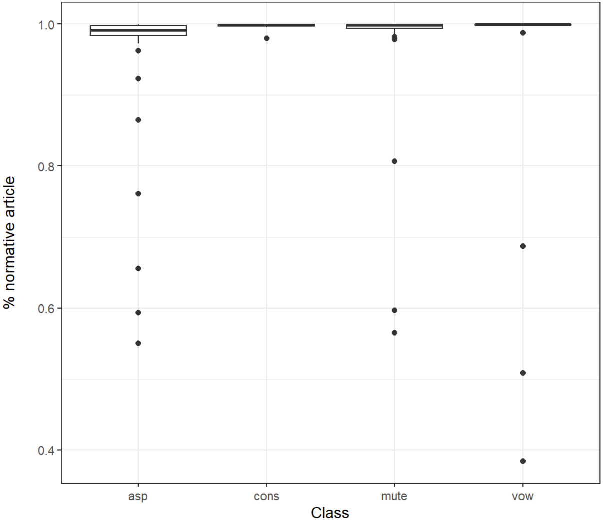

Fig. 2. Percent use of normative article, by class

English results (V-initial)

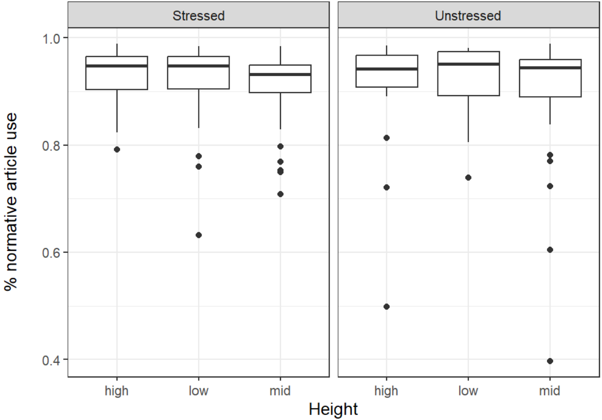

Fig. 3. Percent use of *an*, by height and stress

Map

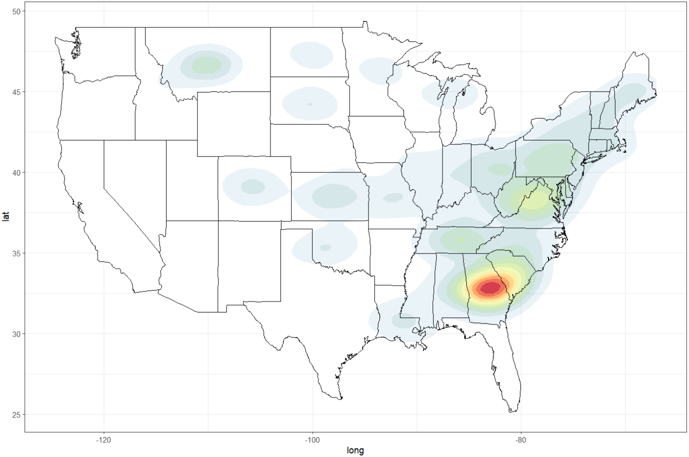

Fig. 4. Heatmap of % *a* + vowel tokens, per capita

Discussion

Discussion

- A little boring, and variation is much less pronounced than in the literature.
- We have nearly 355 million data points for French and more than 888 million for English. That's a lot of noise.
- We can also reasonably expect in any similar, large-scale Twitter corpus (based on a random sample of 25% of files in the English geographic dataset) only 1.5 tokens on average per user.
- It may also be that these phenomena are simply too regular in such a written (albeit often informal) medium.

Limitations

- The “participant” problem may be difficult to solve, at least in the microblogging format. (Other networks with longer text posts may be more reasonable.)
- Factors important in traditional corpus studies (e.g., age, gender, race) are notoriously hard to detect automatically in these kind of corpora.
- Recent advances with machine learning may help, however—I’ve already had some success in classifying political affiliation like this.
- But such an approach takes lots of extra training data, and the clock has essentially already run out.

“Future” directions

- We’ve downloaded data for more natural processes, namely intervocalic spirantization & voicing in various dialects of Italian and word-final /r/-deletion in Brazilian Portuguese.
- All depend on non-standard orthographic representations. Deletion/epenthesis processes seem much more tractable, but featural processes are still promising.
- Other corpora include consonant mutation in tandem with expletive infixation in Welsh and vowel harmony in Turkish borrowings.
- Venturing outside of Twitter, other social media are developing API tools which may eventually fit our needs, but they will necessarily be only one prong of a multifaceted approach.

Acknowledgements

Thank you!

- Many thanks to Chang Chen, Nathan Samson, Yutaka Suzuki, Ariel Sosic and Georges Awaad for their contributions to this project.
- 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

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

Barrie, Christopher and Justin Chun ting Ho. 2021. *academictwitter: an R package to access the Twitter Academic Research Product Track v2 API endpoint*. doi:10.5281/zenodo.4714637. URL `https://github.com/cjbarrie/academictwitteR`, R package version 0.0.0.9000.

De Villers, Marie-Éva. 2009. *Multidictionnaire de la langue française*. Québec Amérique.

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.

Works Cited II

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

Lamontagne, Jeffrey and Gretchen McCulloch. 2022. Phonological variation on twitter: Evidence from letter repetition in three french dialects. *Journal of French Language Studies* 32(2): 165–196.

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

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

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.

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.

Works Cited III

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.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. URL `https://ggplot2.tidyverse.org`.

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