TWITTER-MPHON: STUDYING MORPHOPHONOLOGICAL VARIATION WITH TWITTER DATA*

Michael Dow, François Lareau, and Patrick Drouin
Université de Montréal

1. Introduction

While variation has come to play an important role in theoretical linguistics, it brings with it a slew of questions. For instance, on the end of analysis, should the mechanisms which account for variation be considered as linguistic entities, or are they extralinguistic, existing primarily for the researcher describing them (e.g., Coetzee, 2016)? Meanwhile, on the practical end, at what point can we be certain our observations are reflective of reality, or in more basic terms: how big do our data have to be?

It was with these questions in mind that this project set out to pioneer the use of Twitter[1] data for documenting morphophonological variation, seeing as, at the time of conception, we noted a relative gap in the exploitation of this data source, in comparison with other subfields of linguistics. In this paper, we provide a report on the earlier stages of the project, dubbed Twitter-MPhon.

While recent changes to the API plans offered by Twitter threaten the long-term viability of this project as its methodology currently stands, many of the challenges and potential insights provided by Twitter data are not unique to the medium. As such, we present this research with an eye to the future, as a first step towards studying phonological variation in large-scale, online, written corpora. However, the reader should be advised that this report is not a typical one, seeing as access to our data was cut off towards the beginning of the project.

The rest of this paper is structured as follows: In §2, we present certain general principles which informed Twitter-MPhon and provide detail the specific phenomena under study as well as Twitter as a source of data. Section 3 presents our methodology. Results are provided in §4, and §5 closes the paper with discussion, conclusion and reflection for the future.

*Thanks to Chang Chen, Nathan Samson, Youyang Peng, 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.

[1] While the company formally known as Twitter has recently been rebranded to X, we continue to use the name Twitter in order to keep continuity with our previous work.

Actes du congrès annuel de l'Association canadienne de linguistique 2023.
Proceedings of the 2023 annual conference of the Canadian Linguistic Association.

2. Background

In the initial phase of this project—the one discussed here—we focused on phenomena in French and English, being languages spoken natively by the authors. We hypothesized that variation involving morphophonological phenomena would be more abundant in the Twitter data and easier to target than spellings targeting non-standard pronunciations. Finally, we targeted processes which have been studied from a quantitative point of view in more traditional studies (such as laboratory and corpus studies). This was done in order to give us a point of comparison with our results, be it in terms of simple rates of application or with respect to dialect or sociolinguistic variables.

With these criteria in mind, we chose (1) elision in the definite article before *h-aspiré* words in French (hereafter simply *h-aspiré*) and (2) the *a* ~ *an* alternation in the indefinite article in English. In what follows, we describe these phenomena and outline previous studies which have examined their variability.

2.1 Linguistic phenomena

The so-called *h-aspiré* in French refers to a type of word which phonetically begins with a vowel but which exceptionally blocks external sandhi such as liaison or elision with the definite article. The class of words is arbitrary, but one finds within many words of Germanic origin and recent borrowings (Fagyal et al., 2006). The orthographic forms of these words typically begin with the letter <h>. In opposition, another class of words dubbed *h-muet* also begin with <h> in their orthographic form and with a vowel in their phonetic form but do not block these sandhi. For instance, the *h-muet* word *hôtel* patterns with an unambiguously vowel-initial word such as *action*, in that both take the elided article (i.e., *l'hôtel, l'action*). In comparison, an *h-aspiré* word like *hibou* 'owl' takes the full article (i.e., *le hibou*). Note that some but not all semivowel-initial words also behave like *h-aspiré* words, such as *le yoga*.[2]

Experimental and corpus research on *h-aspiré* finds variable rates of blocking of external sandhi. First, Gabriel and Meisenburg (2009) find that four of their twelve speakers realize liaison consonants before an *h-aspiré* word (though their focus is less experimental), and application of liaison occurs at a rate of approximately 54% before *h-aspiré* words in the pilot results of Tessier and Jesney (2021). The latter study speculates that loss of *h-aspiré* may be occurring in the French spoken in Quebec. Finally, Moisset (1996) looked at elision and liaison in various tasks before 20 *h-aspiré* words for 18 participants (mostly from France). With the exception of one speaker who applied sandhi nearly 50% of the time, average rates of application ranged from 0-20% from speaker to speaker. No effect of gender was found, though significantly higher rates of sandhi were found in lower-class participants. Of the various stimuli, only the word *handicapé* stood out with high rates of

[2] Under certain circumstances, vowel-initial words may also block these sandhi. For instance, the word *onze* 'eleven' when referring to a bus line will take the full article, as in *le onze* (Gabriel and Meisenburg, 2009). We do not consider such cases here.

sandhi. Liaison with /z/ drives higher rates of application, versus liaison with /n/ or elision. Interestingly, inter-speaker acceptability judgments of sandhi application did not correlate with rates of production; generally ranging from 35-45% on average.

In comparison with *h-aspiré*, the *a* ∼ *an* alternation in the indefinite article in English is normatively regular. That is, the prescriptive distribution of each allomorph is *a* before consonant-initial words and *an* before vowel-initial words, without lexical exception. Research shows, however, that this is not the case in all forms of English. Common use of *a* + vowel-initial words has long been documented (Wright, 1905), especially in English (UK) dialectology. In general, 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.

2.2 Basics of Twitter

Twitter is a social media website and application touted as a "microblogging" platform. While various types of media and links are supported, we target only text in our project, which has ranged from 140 to 280 characters up until the time of data collection. Various types of posts, hereafter "tweets," exist, ranging from original tweets to replies, retweets (in which a tweet is shared verbatim to one's account without commentary) and quote tweets (a retweet with one's own commentary). Note that information in the metadata of tweets allow us to distinguish these various types, as well as the quote tweet commentary from the retweeted portion.

As of early 2023, the total number of Twitter users surpasses 450 million accounts (though it should be noted that individuals may possess multiple accounts, including automated "bot" accounts) (Hirose, 2022), with an estimated 500 million tweets sent per day (Sayce, 2022). User age is distributed fairly normally, with the age range of 25-34 representing the largest percentage of users, while user sex is overwhelmingly male on average (Shepherd, 2023). With respect to race, users in the United States are represented fairly evenly (Dixon, 2023). Historically speaking, the majority of content published to Twitter from the United States has come from Democrat-leaning accounts (Center, 2020), but recent changes to Twitter's management are influencing a shift towards the political right.

As of January 2021 and up until spring 2023, Twitter provided a free level of access to its public data in the guise of the Academic access API tier. This essentially allowed researchers to programatically download tweets, user information and associated metadata from its full archives without needing to pass by the website or application interface. While certain restrictions existed (e.g., a limit of 10 million tweets per month), it represented a significant improvement on the constraints of the basic tier of the time without the cost of the business tier (Perez, 2021). In §3.2, we detail the various aspects of this tier which we exploited.

3. Methodology

3.1 Queries

The goal of this project has been from the beginning to be as language-non-specific as possible, such that a generalized workflow could be implemented as automatically as possible between languages. To this end, our main sources of lexical data for generating queries come from large, open-access projects representing a minimum of 100 languages each.

The essential process for generating queries for both languages was the same. Details for each language are provided below, but the general process was as follows: First, the Wiktionary lexicon was downloaded from kaikki.org and imported into R. This was then merged with the relative frequency data from the most recently available web corpus for that language from Sketch Engine (Kilgarriff et al., 2014). Language-specific lexical and phonological filters were applied, and additional subclasses were created. Finally, within each subclass, the *n* most frequent words were selected, and for each word, a normative and non-normative variant provided our queries.

French words were limited to singular nouns, and each word's grammatical gender was extracted. Words were subdivided into the following categories based on the initial segment of the IPA transcription (unless otherwise noted): orthographic <h>, vowel (without orthographic <h>), semivowel, and consonant. Semivowel- and <h>-initial words were further divided according to their "aspiration" behaviour. First, the presence of the "aspirated" tag was detected in the `senses` field. Due to some perceived errors and gaps, this information was supplemented for <h>-initial words from a list of *h-aspiré* words was downloaded from Wikipedia. Cases of disagreement between the two sources were manually verified. Concerning semivowel-initial words, a list of aspirated-like words was manually constructed from the Multidictionnaire de la langue française (De Villers, 2009), which we found to be the most comprehensive dictionary available with respect to this information. Finally, we selected the maximum 25 most frequent words within each category, and we combined it with a full definite determiner (*le* or *la*, depending on its gender) and the elided form. For instance, the queries for a *h-aspiré* word like *hibou* were *le hibou* and **l'hibou*, and so on. This process yielded 135 words, or 270 queries.

English words were limited to singular nouns and adjectives and divided according to initial segment in the IPA transcription into categories of vowel, /h/, /j/, and other consonants. Stress of the initial syllable (primary, secondary, or absent) for vowel-initial words was extracted, and initial vowels were further classified according to height (high, mid, and low). Note that schwa was classified as a mid vowel. The maximum 20 most frequent words according to initial segment type, stress pattern and vowel height were selected, yielding 392 words. An expression with each of the two forms of the indefinite article was created for each word (e.g., *a system*, **an system*), yielding 784 queries.

3.2 Data collection & processing

Data collection was performed in R (R Core Team, 2022) using the `academictwitteR` package (Barrie and Chun-ting Ho, 2021). For each query, the relevant language was specified, limiting searches to only those predetermined as that language by Twitter's proprietary language detection algorithm. Each expression was queried as an exact phrase. The time frame specified was April 2006 until November 2022, with an infinite limit set on the number of pages returned by the package. Counts were gathered for every query, and percentages of non-normative to normative variants were performed at various levels, depending on the phenomenon, and plotted as boxplots.

While we looked at variation in non-normative rates over time, we ultimately found nothing of interest there, and so it is not further discussed. Additionally, we did not find an effect of lexical frequency, nor did any preliminary statistical models find significance between word classes in either of the languages.

As a follow-up, we gathered the tweets and metadata for the English-language queries in the United States which contained geographic information as well. This information was stored as bounding boxes in the `geo` field, or pairs of longitude-latitude coordinates. The centroids of these bounding boxes were calculated, and thus each tweet was associated with a single point. We then matched these points with its respective county or parish in the following way. First, the `counties` data from the `urbanmapr` package (Strochak et al., 2021) was imported and matched with 2019 census data. Each tweet's coordinates were then matched to a county via a proprietary function with the help of the `sp` and `sf` packages (Bivand et al., 2013, Pebesma and Bivand, 2023) and using the WGS 84 coordinate system and an EPSG code of 4269.

For mapping purposes, counts of article *a* + vowel-initial words were then performed within each county and divided by its population to provide a density score. Each tweet's coordinates were then plotted as a heatmap with the `stat_density2d` function from the `ggplot2` package (Wickham, 2016), passing the density score to a gradient.

4. Results

Aggregated use of normative forms obtained at near-ceiling rates for both phenomena. Figure 1 shows the results for French by initial segment (in order: *h-aspiré*, consonant, *h-muet* and vowel). Only a handful of plausible *h-aspiré* words such as *harcèlement* 'harassment' and *handicap* 'handicap' showed rates of full article use beneath 95%. (That is, elided article forms represented more than 5% of tokens for only very few words.) Other words with lower rates of normative article use, such as *hélas* 'alas' at 59% use of full article, strain credibility and warrant follow-up. Figure 2 shows the rate of normative article use in English for vowel-initial words only (i.e., *an*). Again, only a handful of plausible items with lower rates of *an* stood out, such as *administrative, edition* and *available*. It should be noted that the possibility of formulaic tweets artificially inflating these counts cannot be entirely ruled out, especially now that academic access has been phased out. Finally, Figure 3 shows a heatmap of *a* + vowel-initial word usage in the United States. The most striking

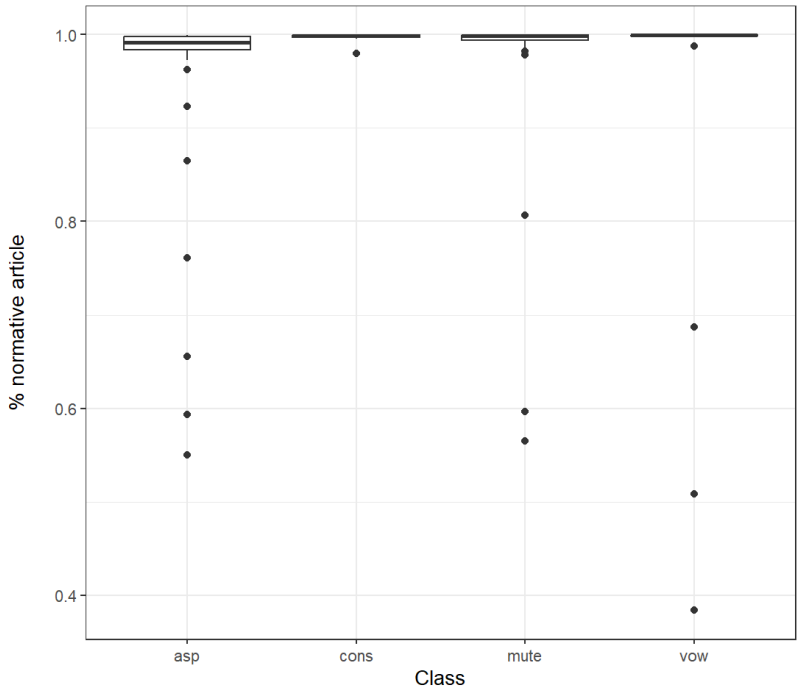

Figure 1. Percent use of normative article in French, by class.

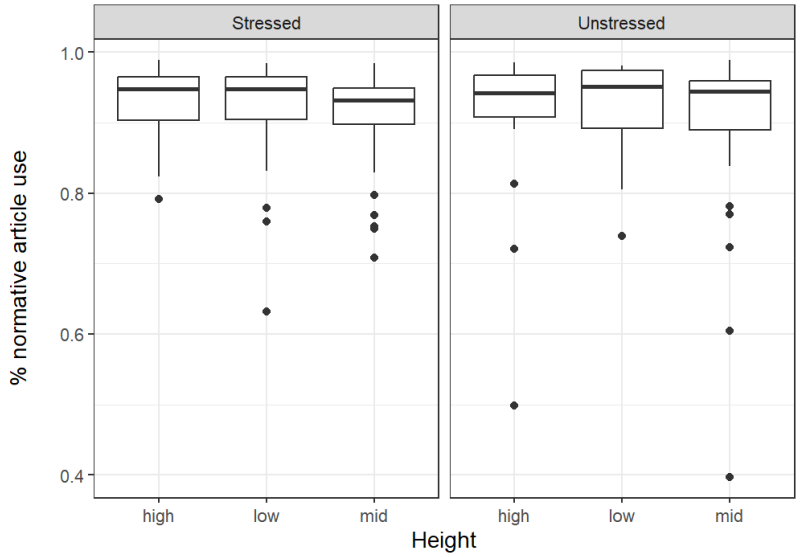

Figure 2. Percent use of *an* in English, by height and stress.

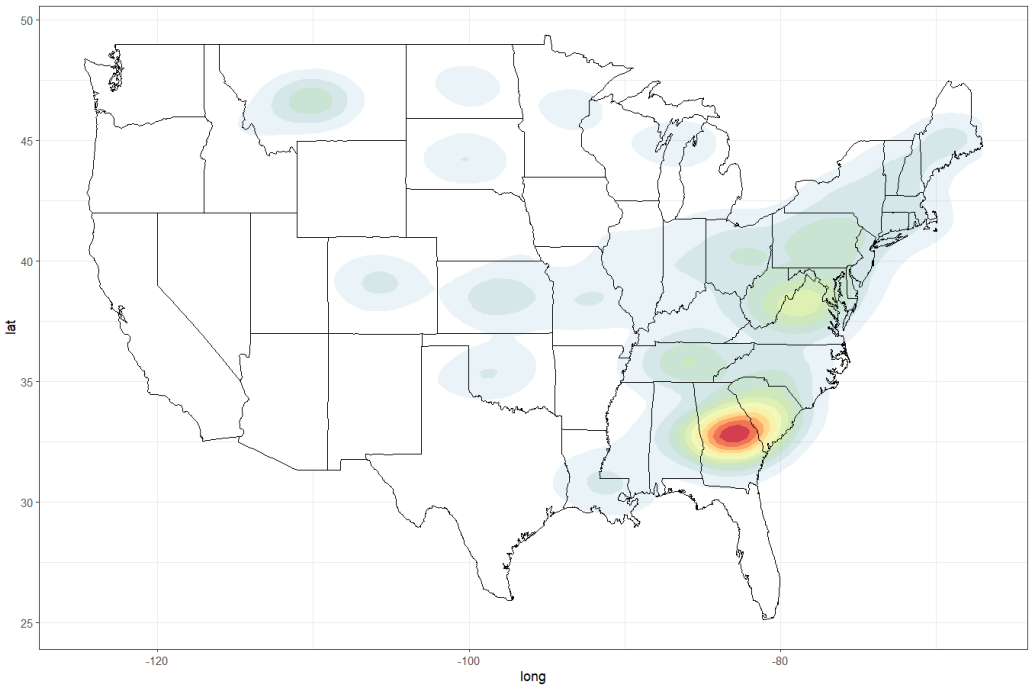

Figure 3. Heatmap of % *a* + vowel English tokens, per capita.

observation in this map is the concentration of *a* + vowel usage in the state of Georgia.

5. Discussion and conclusion

The initial goal of this project was to explore the viability of Twitter data with respect to certain types of phonological phenomena and to develop a long-term, generic workflow. In light of the improbability of the project's future as is, we sought in this report to document the project's philosophy and its initial results. Additional, more promising data have been gathered and are discussed later in this section. First, we would like to contextualize and speculate upon the results presented above.

Aggregated results using the "tweet count" feature yielded little variation, with rates hardly reflective of those reported in the literature. This may not be entirely unexpected in a corpus of such size (nearly 355 million data points for French and more than 888 million for English) and with such an inevitably heterogeneous pool of speakers. (Based on a random sample of the fuller English (geographic) dataset, we estimate an average of only 1.5 tokens per user.) While it would naturally be desirable to look at the effect of known sociolinguistic factors, reliable, automatic methods of identifying certain factors in a Twitter corpus remain elusive (e.g., Golder et al., 2022, Morgan-Lopez et al., 2017). With that said, the factor of geography appears adequately present in the metadata of large corpora and yields promising results.

The comparatively low rates of variation should not be taken to suggest that data

sources such as Twitter are intractable for studying any type of phonological phenomena, as evidenced by Zuraw (2006) and Lamontagne and McCulloch (2022), for instance. It may be, however, that phenomena such as those examined here are so codified as to underrepresent variation in a written medium (however informal as it may sometimes be). For this reason, our project has branched out and managed to collect data for additional phenomena involving more natural, phonetically-motivated processes.

Our conclusions are the following. First, the "big data" previously available to phonologists through Twitter may likely be too big without the integration of additional factors. Second, perhaps unsurprisingly, processes reflected in orthography may not be viable for studying variation in a written corpus, regardless of (in)formality. Finally, while companies such as Twitter may offer enticing and inspiring products for academic research which appear stable, they remain susceptible to external forces and the whims of individuals. A project's eggs should be placed in multiple baskets.

We feel compelled to end on an editorial note. While the results presented in this report are in and of themselves fairly uninteresting (namely in the aggregated ceiling effects), we believe they show potential which was unfortunately cut short. The persisting lack of alternative API plans which are viable to academic researchers is disappointing. We also personally find the silence of official Twitter channels regarding the future of academic access (after professing in March 2023 to be "looking at new ways to continue serving this community") negligent and—especially in light of the use of Twitter data for studying high-stakes topics such as medical misinformation and hate speech—unethical. We encourage scholars to remain vocal in calling for the reinstitution of an API access tier designed for academic research.

References

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.

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio. 2013. *Applied spatial data analysis with R, second edition*. Springer, NY. URL `https://asdar-book.org/`.

Center, Pew Research. 2020. Differences in how Democrats and Republicans behave on Twitter. URL `https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/`. Accessed on March 1, 2023.

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

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

Dixon, Stacy Jo. 2023. Daily Twitter usage in the United States as of August 2018, by ethnicity. URL `https://www.statista.com/statistics/945945/daily-frequency-usage-twitter-usa-ethnicity/`. Accessed on March 1, 2023.

Fagyal, Zsuzsanna, Douglas Kibbee, and Frederic Jenkins. 2006. *French: A linguistic introduction*. Cambridge University Press.

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.

Golder, Su, Robin Stevens, Karen O'Connor, Richard James, and Graciela Gonzalez-Hernandez. 2022. Methods to establish race or ethnicity of Twitter users: Scoping review. *Journal of medical Internet research* 24(4): e35788.

Hirose, Alyssa. 2022. 24 Twitter demographics that matter to marketers in 2023. URL `https://blog.hootsuite.com/twitter-demographics/`. Accessed on March 1, 2023.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36.

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.

Morgan-Lopez, Antonio A, Annice E Kim, Robert F Chew, and Paul Ruddle. 2017. Predicting age groups of Twitter users based on language and metadata features. *PloS one* 12(8): e0183537.

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.

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. URL `https://r-spatial.org/book/`.

Perez, Sarah. 2021. Twitter's new API platform now opened to academic researchers. URL `https://techcrunch.com/2021/01/26/twitters-new-api-platform-now-opened-to-academic-researchers/`. Accessed on March 1, 2023.

R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

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.

Sayce, David. 2022. The number of tweets per day in 2022. URL `https://www.dsayce.com/social-media/tweets-day/`. Accessed on March 1, 2023.

Shepherd, Jack. 2023. 23 essential twitter statistics you need to know in 2023. URL `https://thesocialshepherd.com/blog/twitter-statistics`. Accessed on June 1, 2023.

Strochak, S, K Ueyama, and A Williams. 2021. urbnmapr: State and county shapefiles in sf and tibble format. R package version 0.0. 0.9002.

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.

Zuraw, Kie. 2006. Using the web as a phonological corpus: A case study from Tagalog. In *Proceedings of the 2nd international workshop on web as corpus*.