ON THE DEVELOPMENT OF A SUITE OF TOOLS

FOR THE ANALYSIS OF TWITTER DISCOURSE

By

ROBERT REDMON
Bachelor of Arts in Mass Communication
Midwestern State University
Wichita Falls, Texas
2007

Master of Arts in English
Midwestern State University
Wichita Falls, Texas
2013

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2022

ON THE DEVELOPMENT OF A SUITE OF TOOLS
FOR THE ANALYSIS OF TWITTER DISCOURSE

Dissertation Approved:

Dr. Carol Moder

Dissertation Adviser

Dr. Stephanie Link

Dr. An Cheng

Dr. Shelia Kennison

ACKNOWLEDGEMENTS

This document represents on one hand the conclusion of a journey and on another hand simply a point, albeit a significant point, on the continuation of many other journeys. The tool described below will continue to evolve. My identity as a researcher and an academic will continue to evolve. Getting to this point was, nevertheless, a significant endeavor—one that would not have been possible without a variety of external motivators and supporters. Foremost among these are my parents, whose seemingly-infinite patience, motivation, and love gave me license to be a student for a **very** long time, and whose eccentricities provided a model for accepting and engaging with my own.

More to the project at hand, I am particularly grateful for the support of my committee members, particularly Dr. Carol Moder, my advisor, and Dr. Steph Link, who became what can only be described as a mentor. Dr. Moder provided a model for the kind of researcher I would like to be, nurtured my interests in discourse and cognitive linguistics, and instilled in me the importance of studying language based on real usage, in context. Dr. Link recognized my technical, computational skillset, showed me its value in linguistic work, and pushed me to develop those skills with an eye to not only functionality, but usability and user experience as well. TWIG would not be what it is without the contrasting influences of both of these women.

Finally, I must thank my partner, Clare Paniccia, whose love, support, and encouragement—while she navigated her own journey and her own dissertation—provided the fuel with which I was ultimately able to finish this thing.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: ROBERT REDMON

Date of Degree: MAY, 2022

Title of Study: ON THE DEVELOPMENT OF A SUITE OF TOOLS FOR THE ANALYSIS OF TWITTER DISCOURSE

Major Field: ENGLISH

Abstract: Social media has become a ubiquitous avenue of natural language use that represents an enormous amount of data, which has in the last decade been used in an increasing amount of linguistic research in such topics as regional variation, stylistic variation, and identity and community construction. For a number of technical reasons, however, the collection and analysis of Twitter data for research purposes is currently prohibitively difficult, especially for studies with a particular interest in Twitter as a site of discourse creation. Existing collection tools require almost all require programming skills, and existing analytical tools are not equipped to handle the metadata that Twitter data includes, much less to leverage that metadata to highlight linguistic trends or to make large datasets navigable. In an attempt to address this problem, I have developed a web-based set of tools, called TWIG (TWItter Getter), for collecting and analyzing Twitter data. This document begins with a discussion of the basic mechanics of Twitter and its nature as a discourse context and then details the need for specialized Twitter data collection and analysis tools that focus on the needs of researchers interested in Twitter generally, and those interested in discourse specifically. The remainder of the document describes TWIG and demonstrates its use in two sample studies—one an analysis of a community's construction of identity and values on twitter, and the second a usage-based analysis of semantic change in the construction *really do be.*

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Social media has become a ubiquitous avenue of natural language use. Twitter, the platform of focus for the current project (with aims towards expanding to other platforms in the future) has around 330 million active users worldwide (Twitter, 2022b) generating, in 2014 (the last year for which there is an official report), around 500 million tweets per day (Stricker, 2014) and is used by around 23 percent of American adults (Auxier & Anderson, 2021). This represents an enormous amount of data which has in the last decade been used in an increasing amount of linguistic research in such topics as stylistic variation (e.g. Clarke & Greive, 2019), memes (Dancygier & Vandelanotte, 2017), and identity and community construction (Zappavigna, 2014).

However, for a number of technical reasons, which are detailed in the chapters below, the collection and analysis of Twitter data for research purposes is currently prohibitively difficult, especially for studies with a particular interest in Twitter as a site of discourse creation. Existing collection tools require almost all require programming skills, and existing analytical tools both lack the statistical sophistication to represent corpora at a discourse level are not equipped to handle the metadata that Twitter data includes, much less to leverage that metadata to highlight linguistic trends or to make large datasets navigable.

In an attempt to address this problem, I have developed a web-based set of tools, called TWIG (TWItter Getter), for collecting and analyzing Twitter data. The first chapter of this document begins with a discussion of the basic mechanics of Twitter and its nature as a discourse context and then details the need for specialized Twitter data collection and analysis tools that focus on the needs of researchers interested in Twitter generally, and those interested in discourse specifically. This discussion first addresses the technical need for data collection tools and then addresses both the technical and theoretical needs for specialized tools for analyzing this data. The second chapter of this document describes the tools that comprise TWIG and the decisions made during their development and how they address specific needs described in the second chapter. The final chapters of this document are two sample studies performed using TWIG, aimed at demonstrating its effectiveness in facilitating a range of Twitter-oriented studies. The first of these studies uses a corpus of posts from a specific group of users in an analysis of community and individual identity construction; the second uses a query-based corpus of Tweets in a usage-based analysis regarding an internet meme.

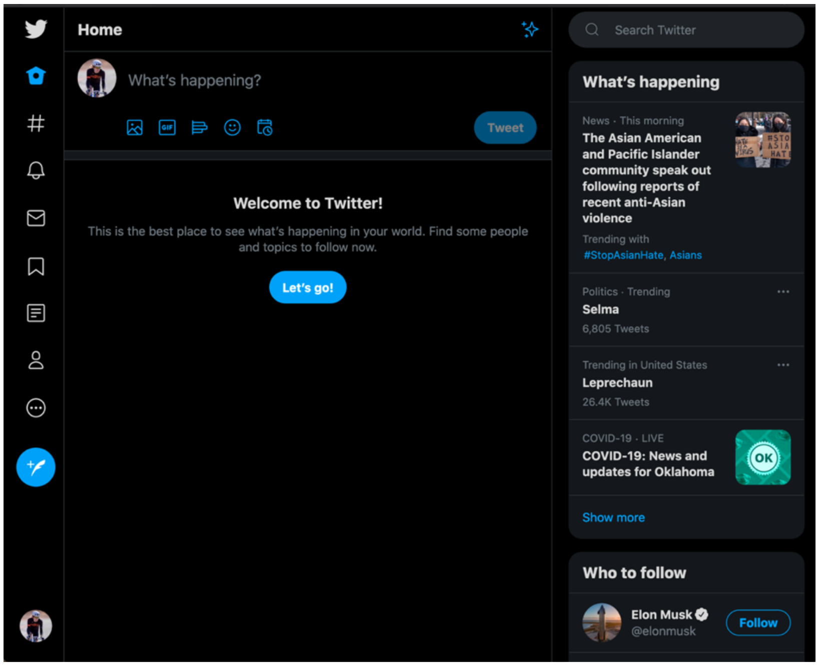
Twitter as a discourse context

Twitter is not a simple source of linguistic data. It must be noted at the outset that active Twitter users are not strictly representative of the broader populations they belong to. They trend younger, more male, more educated, and more politically liberal (Mellon & Prosser, 2017; Mislove et al., 2012). Nevertheless, the tweets they produce can be used as a timely source of data for investigating emerging trends in usage and identity, and the immediacy and scope of Twitter data make it especially valuable in analyses of broad public discourses ranging from official political campaigning (e.g. Engström, 2020; Kreiss, 2016) to grassroots social movements (e.g. Choi & Park, 2013; Crossett et al., 2018). Because such open public discourse is a relatively new phenomenon, however, much of the linguistic research done using Twitter data is necessarily also concerned at least in part with understanding Twitter itself as a context for novel kinds of discourse. Such research is often concerned with questions regarding how these emergent discourse contexts enable new kinds of meaning-making, affiliation, community, and interaction (e.g. Zappavigna, 2011; Zhang et al., 2017; Scott, 2015). To approach such questions, and to understand how a specialized analytical tool like TWIG might help answer them, it is necessary to first understand the mechanics of reading and posting on Twitter.

Twitter is an online social media platform, accessible from both a web browser and a proprietary app for mobile devices, on which users post and read short messages called "tweets." The relationship between users on Twitter is non-reciprocal, or one-to-many. This is to say that, unlike platforms such as Facebook that frame user relationships as two-way, such that users have "friends" who see one-another's posts, Twitter is largely characterized by one-way relationships. Users follow other users who might not follow them back and are in turn followed by users who they might not choose to follow back. This framework, by diminishing personal connections and emphasizing instead the creation of followable content, makes Twitter especially conductive to topical discourse (Zappavigna, 2011), allowing the rapid emergence of discourses surrounding current events and social issues, on both personal and global scales.

How these discourses are created, navigated, and engaged with on Twitter and other non-reciprocal social-media platforms is also notably different both from traditional spoken and written communication and from earlier "web 1.0" forms of online communication, such as web forums and live chatrooms. Unlike web forums and chatrooms, these social-media platforms do not group discourses topically in separate browsable areas of the site or app. Instead, all posts are treated as essentially equal within a single torrent of constantly-emerging content.

By default, for a new user, the flood gates are closed. One way of opening them is by following other users. When a user opens Twitter, they are presented with a an infinitely-scrollable feed of tweets posted by every user they follow, ordered by an algorithm based on chronology and popularity. A second way of accessing tweets is by utilizing Twitter's built-in search functionality. Because Twitter does not itself group or categorize tweets, searching for a discourse means actually searching the text of the tweets themselves. Figure 1.1 and Figure 1.2 below show how Twitter timelines and search results appear on the web app.



Welcome to Twitter!

This is the best place to see what's happening in your world. Find some people and topics to follow now.

What's happening

News · This morning
The Asian American and Pacific Islander community speak out following reports of recent anti-Asian violence
Trending with #StopAsianHate, Asians

Politics · Trending
Selma
6,805 Tweets

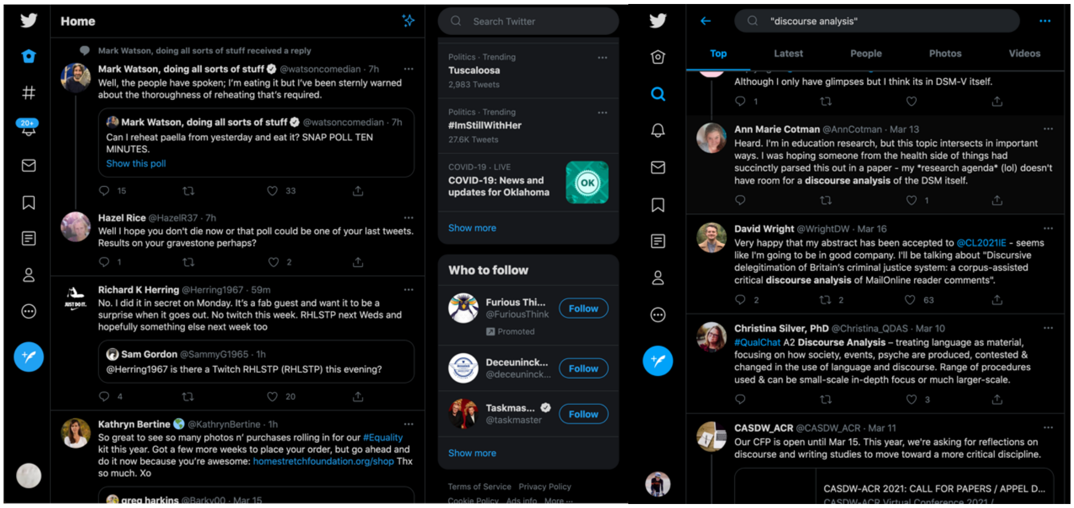
Trending in United States
Leprechaun
26.4K Tweets

COVID-19 · LIVE
COVID-19: News and updates for Oklahoma

Who to follow

Elon Musk @elonmusk

Figure 1.1 - Unpopulated timeline



Mark Watson, doing all sorts of stuff received a reply

Mark Watson, doing all sorts of stuff @watsoncomedian · 7h
Well, the people have spoken; I'm eating it but I've been sternly warned about the thoroughness of reheating that's required.

Mark Watson, doing all sorts of stuff @watsoncomedian · 7h
Can I reheat paella from yesterday and eat it? SNAP POLL TEN MINUTES.
Show this poll

15 33

Hazel Rice @HazelR37 · 7h
Well I hope you don't die now or that poll could be one of your last tweets. Results on your gravestone perhaps?

1 2

Richard K Herring @Herring1967 · 59m
No. I did it in secret on Monday. It's a fab guest and want it to be a surprise when it goes out. No twitch this week. RHLSTP next Weds and hopefully something else next week too

Sam Gordon @SammyG1965 · 1h
@Herring1967 is there a Twitch RHLSTP (RHLSTP) this evening?

4 20

Kathryn Bertine 🌎 @KathrynBertine · 1h
So great to see so many photos n' purchases rolling in for our #Equality kit this year. Got a few more weeks to place your order, but go ahead and do it now because you're awesome: homestretchfoundation.org/shop Thx so much. Xo

greg harkins @Barky00 · Mar 15

Politics · Trending
Tuscaloosa
2,983 Tweets

Politics · Trending
#ImStillWithHer
27.6K Tweets

COVID-19 · LIVE
COVID-19: News and updates for Oklahoma

Who to follow

Furious Thi... @FuriousThink

Deceuninck... @deceuninck...

Taskmas... @taskmaster

Although I only have glimpses but I think its in DSM-V itself.

1

Ann Marie Cotman @AnnCotman · Mar 13
Heard. I'm in education research, but this topic intersects in important ways. I was hoping someone from the health side of things had succinctly parsed this out in a paper - my *research agenda* (lol) doesn't have room for a **discourse analysis** of the DSM itself.

1

David Wright @WrightDW · Mar 16
Very happy that my abstract has been accepted to @CL2021IE - seems like I'm going to be in good company. I'll be talking about "Discursive delegitimation of Britain's criminal justice system: a corpus-assisted critical **discourse analysis** of MailOnline reader comments".

2 2 63

Christina Silver, PhD @Christina_QDAS · Mar 10
#QualChat A2 **Discourse Analysis** – treating language as material, focusing on how society, events, psyche are produced, contested & changed in the use of language and discourse. Range of procedures used & can be small-scale in-depth focus or much larger-scale.

3

CASDW_ACR @CASDW_ACR · Mar 11
Our CFP is open until Mar 15. This year, we're asking for reflections on discourse and writing studies to move toward a more critical discipline.

CASDW-ACR 2021: CALL FOR PAPERS / APPEL D...
CASDW-ACR Virtual Conference 2021 /

Figure 1.2 - Populated timeline (left) and search results (right)

It is worth noting that in these scrollable feeds of tweets, not all tweets are presented equally. Especially in the timeline view on the bottom left, it is evident that there are levels of indentation and lines connecting certain tweets. This is because, while all tweets are in one sense discrete—the most basic elements of a “tweet object,” the data structure Twitter uses to store individual tweets, are a user id, a tweet id, and the text of the tweet—some tweets are stored with references to other tweets that affect how they are ultimately displayed. Tweet objects will be discussed more thoroughly in Chapter II of this document, and the nature of possible connections between tweets will be elaborated on later in this chapter.

Before discussing Twitter’s framework for direct connections between tweets, which are relatively recent additions to the platform, however, it is valuable to understand the framework by which tweets can be indirectly connected, because these are more fundamental to the specific nature of discourse on Twitter. A combined effect of Twitter’s privileging non-reciprocal user relationships and ungrouped, searchable discourses is the emergence of what Zappavigna (2011, 2014) refers to as an “ambient” environment. Zappavigna’s sense of ambience here describes the emergence of discourses on Twitter that often appear not as interactions, but instead as multiple users “talking about the same topic at the same time” (Zappavigna, 2014, p. 211). In ambient discourse, interaction becomes at least somewhat a matter of indirect coordination between parallel streams of tweeting. This means a critical function of language in many tweets is to create connections between themselves and possibly-related tweets, maximizing what Zappavigna calls “findability" (2011, p. 798). As Twitter’s developers have changed how users can interact with one another’s tweets over time, the nature of discourse on Twitter has become less “ambient,” but this notion of findability remains central to both much of the language used on Twitter and to the mechanisms of using the platform.

Twitter has several hard-coded mechanisms for maximizing within the text of a tweet the findability of that tweet, a user, or a specific discourse. These mechanisms have changed over time with the evolution of the Twitter platform. Hashtags, @mentions, and retweets are the most fundamental, having been part of the platform in various forms since its inception. More recent, and so less studied, additions to the platform include quote tweets, replies, and (to a degree) likes. Of these, hashtags have likely received the most academic attention (e.g. Lachlan et al., 2014; Scott, 2015; Zappavigna & Martin, 2018). Hashtags have the form of a # symbol followed by text without spaces. Searching for or clicking a hashtag within a tweet displays a scrollable list of tweets containing that hashtag. This list can be organized to prioritize either popular tweets or most recent tweets. Popularity here is a product of how many users have "liked" a tweet, which is done by clicking or tapping a heart icon that appears beneath the text of the tweet. The effect of Twitter's hashtagging functionality is that using a hashtag allows a user to either find a topical discourse or index their own tweet within a discourse. Users can both create novel hashtags and adopt hashtags that are already in use, though these options carry different functions. It is worth emphasizing here that all hashtags, all searchable discourses and topics, emerge from Twitter's user base, not Twitter itself. Structurally, hashtags often appear at the ends of tweets, outside the tweets' clausal components, as in example 1 below, but also frequently occur within the grammatical structure of the tweet, as in example 2

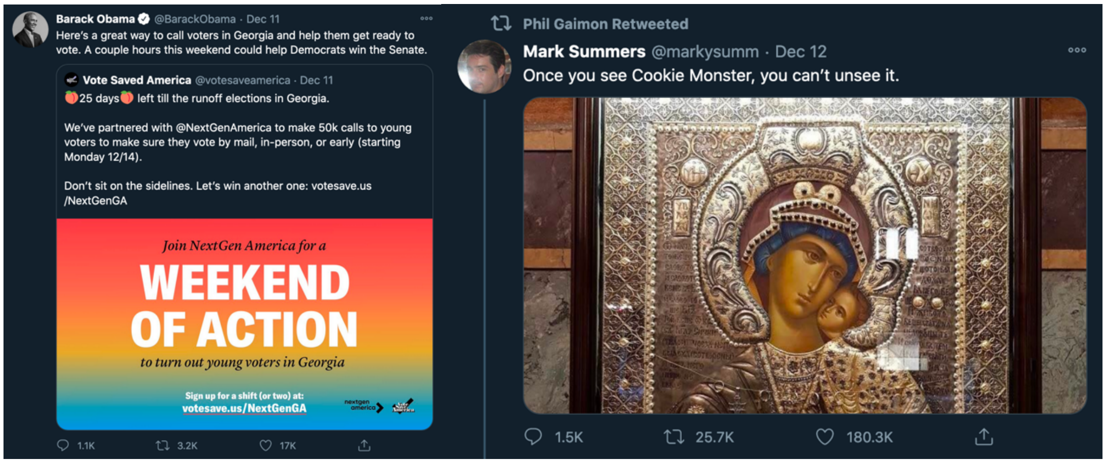
1 Preparing my 14th season as a pro rider. Love my job and hopefully more years to come ✊👌 **#passion #cycling** https://t.co/xrEsacJc9T

2 Final TT prep before the **#nationalchampionships** tomorrow. Let's get that 2017 changed to 2018…

Where hashtags can increase the findability of their containing tweets by linking them to broader discourses, @mentions, retweets, quote tweets, and replies work a bit differently, increasing not the visibility of the tweets themselves, but rather that of @mentioned and retweeted users. (The use of @mention here is adopted from Hemsley et al. (2017) in order to prevent confusion between the Twitter practice and more general meanings of mention.) An @mention is the most basic of these, appearing within the text of a tweet as an @ symbol followed by a username, as shown several times in example 3. Each @mention acts as a hyperlink to that user's profile and facilitates visibility both by notifying the @mentioned user of the tweet and by making that user visible to anyone who sees the tweet.

3 Enjoyed my day in the ochre jersey of the **@tourdownunder** together with my **@Lotto_Soudal**. Chapeau to **@CalebEwan** + **@MitcheltonSCOTT** for that impressive performance in today's stage.

Retweets take this a step farther, making not only the retweeted user more visible to the retweeting user's followers, but also, and especially, the retweeted user's tweet. A retweet appears on a user's timeline or within a list of tweets generated by a search result as the original user's tweet would normally, below slightly lightened text indicating who retweeted it. In Twitter's earlier days, and in many existing studies that use Twitter data, this was different. Retweets would appear as simply the letters "RT" followed by the original user's username, a colon, and the content of their tweet. In this era of Twitter, retweets could appear, for lack of a better term, bare, while others include a brief note from the retweeting poster, which appeared before the RT. The distinction between bare retweets and retweets with notes has since (it is difficult to say when, because Twitter's documentation does not include dates for changes in functionality) become more hard-coded, such that a retweet with a note is now treated as its own tweet type, a "quote tweet." Quote tweets appear on timelines as normal tweets from quoting users with the quoted tweet embedded beneath. Figure 1.3 below illustrates the difference between how quote tweets and retweets appear on Twitter.



Barack Obama @BarackObama · Dec 11
Here's a great way to call voters in Georgia and help them get ready to vote. A couple hours this weekend could help Democrats win the Senate.

Vote Saved America @votesaveamerica · Dec 11
🍑25 days🍑 left till the runoff elections in Georgia.

We've partnered with @NextGenAmerica to make 50k calls to young voters to make sure they vote by mail, in-person, or early (starting Monday 12/14).

Don't sit on the sidelines. Let's win another one: votesave.us /NextGenGA

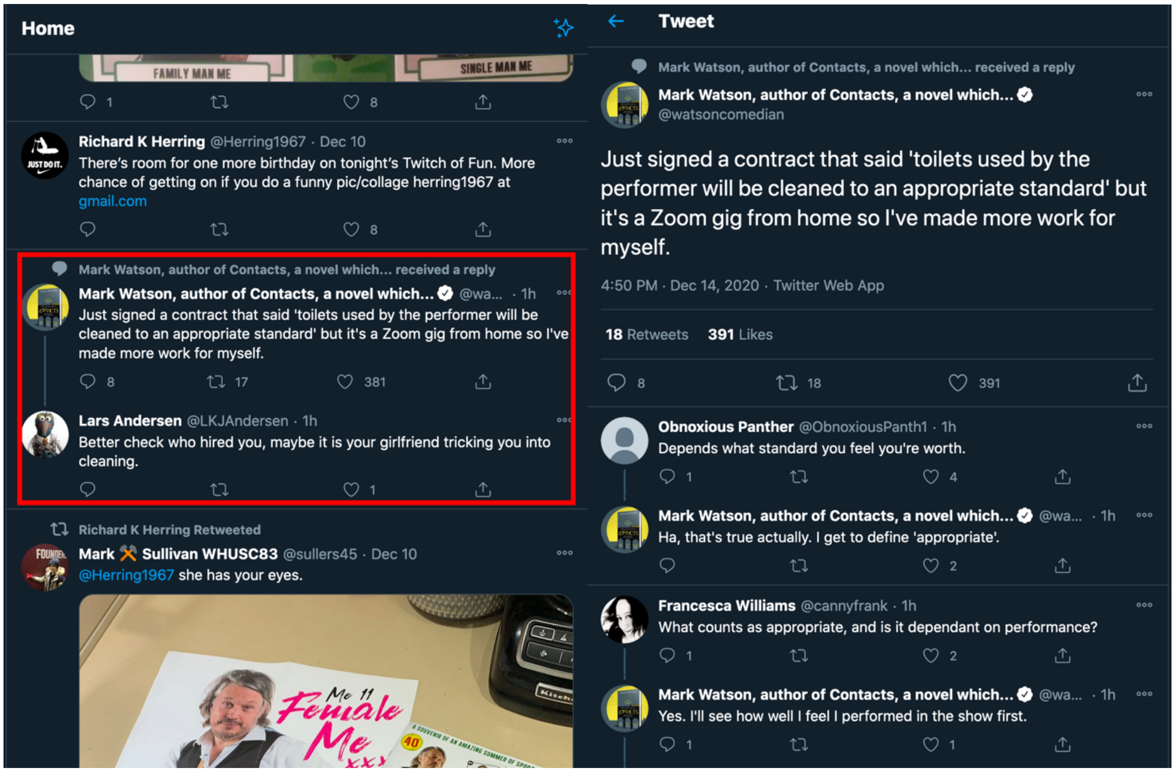
Join NextGen America for a
WEEKEND OF ACTION
to turn out young voters in Georgia
Sign up for a shift (or two) at:
votesave.us/NextGenGA
nextgen america
1.1K
3.2K
17K

Phil Gaimon Retweeted
Mark Summers @markysumm · Dec 12
Once you see Cookie Monster, you can't unsee it.

Figure 1.3 – A quote tweet (left) and a retweet (right)

Note that in the retweet on the right posted by Phil Gaimon, Gaimon is represented in a diminished capacity, with no avatar and a dimmed username above a full original tweet. Conversely, in the quote tweet on the left posted by Barack Obama, Obama is represented as is typical of a top-level tweet, with an avatar and a bright username, and it is the quoted tweet that is slightly diminished.

A final functionality of Twitter is that users can reply to one-another's tweets. Replies are still tweets, but they appear on timelines in a diminished form, paired with, but beneath the tweet that is being replied to, with lightened text above the pair indicating that a user you follow has replied to another user's tweet or received a reply. This can be seen in Figure 1.4 below.



Richard K Herring @Herring1967 · Dec 10
There's room for one more birthday on tonight's Twitch of Fun. More chance of getting on if you do a funny pic/collage herring1967 at gmail.com

Mark Watson, author of Contacts, a novel which... received a reply

Mark Watson, author of Contacts, a novel which... @wa... · 1h
Just signed a contract that said 'toilets used by the performer will be cleaned to an appropriate standard' but it's a Zoom gig from home so I've made more work for myself.

Lars Andersen @LKJAndersen · 1h
Better check who hired you, maybe it is your girlfriend tricking you into cleaning.

Richard K Herring Retweeted
Mark ⚒ Sullivan WHUSC83 @sullers45 · Dec 10
@Herring1967 she has your eyes.

Mark Watson, author of Contacts, a novel which... received a reply

Mark Watson, author of Contacts, a novel which...
@watsoncomedian

Just signed a contract that said 'toilets used by the performer will be cleaned to an appropriate standard' but it's a Zoom gig from home so I've made more work for myself.

4:50 PM · Dec 14, 2020 · Twitter Web App

18 Retweets **391** Likes

Obnoxious Panther @ObnoxiousPanth1 · 1h
Depends what standard you feel you're worth.

Mark Watson, author of Contacts, a novel which... @wa... · 1h
Ha, that's true actually. I get to define 'appropriate'.

Francesca Williams @cannyfrank · 1h
What counts as appropriate, and is it dependant on performance?

Mark Watson, author of Contacts, a novel which... @wa... · 1h
Yes. I'll see how well I feel I performed in the show first.

Figure 1.4 – Replies shown on timeline (left) and in chain beneath isolated tweet (right)

These paired tweets can be clicked to view the original tweet and all of its replies in isolation. Replies are another functionality that was not always part of the Twitter platform, and which seem at first glance to be a dramatic departure from Twitter's ambient roots. However, the fact that reply pairs appear in the timelines of users who follow either the replier or the replied to makes them rather dynamic in terms of the discourses they engage with. As a reply to a specific tweet, a reply is in one sense an entry to a narrow discourse framed within the original tweet, but by appearing on user timelines, along with standard tweets, retweets, and quote tweets, replies also have a similar visibility enhancing function to other features of Twitter, inviting anyone who sees the smaller tweet-reply pair into the narrow discourse surrounding a specific tweet.

Existing linguistic and discourse research

From a linguistic perspective, hashtags, @mentions, retweets, quote tweets, and even to a certain extent replies have no direct spoken-language analogues. They often inhabit the positions and functions of normal parts of speech or kinds of interaction, but they also carry quasi-prescribed technical functions that cause the messages in which they occur to be treated differently by the software that carries the messages from sender to receiver. These differences in turn create the possibility of new kinds of meaning making. As such, it’s not surprising that these features, considered individually or together, have featured heavily in studies of language on Twitter. To an extent, some studies specific features of Twitter in the abstract, such as Boyd et al.’s (2010) investigation of retweeting and Scott’s (2015) look at the pragmatics of hashtags. However, on the whole, these features have been analyzed more within specific discourse contexts, such as Lachlan et al.’s (2014) comparison of local and national hashtags pertaining to weather events, Hemsley et al.’s (2017) look at political candidates’ @mentioning practices.

Working from Zappavigna’s (2014) notion of identity as “patterns of bonds when approached in terms of the social relations that they enact” (p. 223), it seems natural that language that literally links, or creates bonds between, utterances and discourses (hashtags) or other entities (@mentions, retweets) would do identity-constructive work. And indeed a number of studies have looked at such construction of bonds. Hashtags in particular have received a lot of attention for their use in constructing online communities around particular political or activist groups (Blevins et al., 2019; Choi & Park, 2013; Kuo, 2018; Ross & Bhatia, 2019) or in constructing ad-hoc communities around current events (Eriksson, 2018; Kreis, 2017).

Other studies have considered more fixed communities, such as language scientists (Jünger & Fähnrich, 2019) and parody accounts (Highfield, 2015). Perhaps because of its currency, Twitter has also been frequently used in analyses of political discourse, including topical analyses of politicians’ accounts (Graham et al., 2016), live tracking of Twitter activity during debates (Busy et al., 2020), analysis of an individual candidate's stylistic variation (Clarke & Greive, 2019), and the strategic use of Twitter in political campaigns construction of their agenda (Kreiss, 2016).

The studies noted above represent only a fraction of the linguistic work being done with Twitter data, but they all illustrate bare technical necessities of working with Twitter-sourced data. First, researchers must be able to collect tweets, en masse, that include specific features or from specific users or groups of users. Further, researchers need to be able to analyze the data they have collected in ways that differentiate between users, groups of users, communicative contexts, time periods, topics, and so forth. The following sections detail the technical hurdles and theoretical concerns that currently exist for conducting such research, with a particular focus on the use of Twitter data for discourse analysis.

The technical need for data-collection tools

While Twitter is an extremely productive source of timely natural linguistic data, accessing and using that data, for both technical and legal reasons, is not simple. As with other kinds of web media, the data it provides cannot be easily or effectively used for research via direct interaction. While Twitter's basic functionality allows easy access to all of a specific user's tweets or, using the search feature, to an abundance of tweets containing a word or phrase of interest, the kinds of analysis that are possible via such inquiry are prohibitively limited. For research purposes, then, it is advantageous to the point of near necessity that data be collected and stored as deliberately-analyzable datasets or corpora. While differences between datasets and corpora are important for certain research purposes, having largely to do with questions of representativeness, those differences are not particularly salient to the discussion of the technical aspects of collecting Twitter data and, as such, will not be addressed at this point of the current document. However, the value of storing language use taken from the web in structured, research-focused datasets is well explained by the creators of the BYU Wikipedia Corpus:

> [T]his corpus allows you to search Wikipedia in a much more powerful way than is possible with the standard interface. You can search by word, phrase, part of speech, and synonyms. You can also find collocates (nearby words), and see re-sortable concordance lines for any word or phrase.
>
> (Davies, 2015)

With this in mind, the usability of Twitter data in linguistic research is tied meaningfully to the collectability of that data in analyzable forms. Such collectability, however, is complicated by shifting technical and legal boundaries. Historically, there have been three ways to collect Twitter data: by copy and pasting tweets directly from twitter, which is incredibly time consuming; by automating the collection of data using the Twitter API, an Application Programming Interface provided by Twitter itself, which requires computer programming experience; or by utilizing readymade corpora of Twitter data, most prominently the now-defunct HERMES corpus (Zappavigna, 2012).

Readymade corpora were central to an early wave of studies investigating Twitter discourse (Zappavigna, 2012; Page, 2012). However, such studies are no longer possible, because such corpora are no longer possible. As of 2013, Twitter's Developer Agreement and Policy explicitly forbids the redistribution of data collected using the Twitter API. It is worth noting that this policy has parameters. The Developer Agreement permits direct person-to-person sharing of a limited amount of data—50 thousand full Tweet Objects (entries that include the tweet text and associated metadata) or 1.5 million Tweet IDs (unique identifiers with which full Tweet Objects can be retrieved). As of 2020, academic users specifically may share unlimited Tweet IDs—but, again, not complete datasets (Twitter, 2020b). A few corpora do exist that attempt to compensate for these guidelines. For example, the National Institute of Standards and Technology makes available the Tweets2011 dataset, which can only be obtained by emailing the creator directly and requesting what is not the dataset itself, but instead access to several lists of Tweet IDs (each list includes 1.5 million) that can then be used by the researcher to reconstruct the corpus (NIST, 2019)—a process that, because of further Twitter-imposed limits on how many tweets can be collected at once, can take days—and will include only tweets from 2011 and earlier.

It is also worth noting that a very limited number of corpora also exist that are essentially outside of Twitter's policy. A few recent studies (e.g. Clarke & Grieve, 2019) have used readymade corpora of the tweets of Donald Trump, including those which Trump (or a handler) deleted after posting. These corpora likely owe their survival to a 2017 federal court ruling that President Donald Trump's tweets are public statements. Additionally, Twitter itself makes available datasets of Tweets made by accounts that engage in "[p]latform manipulation that [Twitter] can reliably attribute to a government or state-backed actor" (Twitter, 2020c). These datasets have been used in a few linguistic studies (Lundberg & Laitinen, 2020), but, like Trump corpora, they are ultimately of very narrow use for linguistics researchers. In practical terms, then, Twitter's data-sharing policy clearly prevents the meaningful development and publication of general-use academic corpora in the vein of the Corpus of Contemporary American English (Davies, 2008).

The resultant situation is one in which Twitter still represents an enormous amount of data, but it is data which researchers must collect and organize themselves—either by manually copying and pasting tweets from Twitter, which, again, is incredibly time consuming, or by accessing Twitter's API on their own. The Twitter API is a service offered by Twitter that allows developers, who must be approved for API credentials, to create queries in a programming environment that return a structured dataset of "tweet objects" of tweets that match the query criteria. This allows the rapid collection of thousands of tweets, complete with metadata. Despite this, some studies are still conducted with manually entered data (e.g. Graham et al., 2016; Kreis, 2017). A likely reason for this is that many researchers who want to study Twitter data either do not know about API access or they lack the technical skills to collect data in this way.

The technical knowledge barriers to confidently navigating this process are large. At the bare minimum, researchers must have enough of a working understanding of computer programming to acquire existing code, modify it to fit their project, and run it. The need for more accessible tools has been addressed to a degree both within the field and externally in the distribution of extensions for R and Python programming environments (Hemsley et al., 2014, 2019; Tweepy, 2017), and standalone software (Anthony, 2018).

Even resources that exist to streamline this process—such as Tweepy, a Python library aimed at simplifying collecting Twitter data and STACKS (Social Media Tracker, Analyzer, and Collector Toolkit at Syracuse), a more robust toolkit that allows for the creation of savable projects and includes analysis features—can only be used within a programming environment. Of these, simple collection libraries like Tweepy are most frequently used in real-world scholarship. Toolkits that aim to provide more complete research solutions tend to be more complex to set up and to use, and as a result they tend to be used mostly by their initial developers. For example, in searching for studies that have used STACKS to collect or analyze data, I was able to find only one in which one of the STACKS developers was not a listed author. And that was a dissertation from a PhD student at Syracuse. This is not surprising given the following from the STACKS documentation:

> **This documentation assumes the following:**
> You know how to use ssh.
> Your server has MongoDB already installed.
> You understand how to edit files using vim (“vi”) or nano.
> You have rights and know how to install Python libraries.
>
> (Hemsley et al., 2019)

This is a lot to ask for the simple collection of text data from the web and, along with the broader picture described by this section, is illustrative of the need for a more broadly accessible system for collection of Twitter and other social-media data for linguistic research. Lawrence Anthony began to address this need with the development of FireAnt (2018), a standalone application, available for multiple operating systems, which provides tools for the collection, analysis, and visualization of social media data. The current project, TWIG, and FireAnt differ both in terms of the guiding philosophy behind the provided analytical tools, which will be discussed in detail below, and in terms of their approach to making a more accessible tool for collecting social media data.

The primary difference in approaches that TWIG and FireAnt take to accessibility is that TWIG was developed as a web application, to be run in web browsers, where FireAnt is a downloadable standalone application. Standalone applications allow users to work offline and to store data locally, where web applications, inversely, require an internet connection and necessitate that the data being analyzed be stored on a server beyond the researcher's control. Given that an internet connection is already a requirement for collecting social media data, however, the most pressing concern of web-based research tools is maintaining researchers' access to their own data, which is important. Data stored on servers can become inaccessible, or even permanently lost, if the tool is taken down or if there is some problem with the server. Furthermore, researchers may wish to have their data locally in order to perform specialized analyses that are not possible within the web application. For such concerns, there is no foolproof simple solution, but they can be mitigated by simply allowing users to download their data and notifying them of the value of a secure local backup.

I believe the advantages of a web-based application outweigh the potential harm of this disadvantage. The primary advantages are sustained compatibility and more flexible user-experience design. Compatibility is a complex feature. Standalone applications must contend with compatibility issues that arise as operating systems update over time. This is a particularly salient problem for niche academic software, which, as Anthony (2018) notes, can be difficult to find the resources to maintain after the initial release. This problem is compounded by having to separately maintain software for multiple operating systems. As an illustration of this, I was unable to run FireAnt's data collection component on a 2019 MacBook Pro with MacOS Catalina, which was released only a year after FireAnt.

The most glaring compatibility issue faced by applications that collect Twitter data, however, is that Twitter occasionally changes its API in ways that require connecting differently or constructing queries differently and that return data in a slightly different structure. This is particularly salient now, because in January 2021 Twitter announced v2 of their API, with many significant departures from its predecessor v1.1, and among these departures was drastically enhanced access for academic users. Academic API v2 users now have access to Twitter's full history, where v1.1 users are limited to tweets from the last six months. Further, monthly collection caps are twenty times higher and Twitter has implemented additional metadata and filtering options (Torness & Tujillo, 2021). For users of a standalone tweet collection application to get access to these improvements or any other updates, the software would have to be updated, recompiled for different operating systems, and published—which, as noted above, is time-consuming and infrequent with academic software—and, crucially, the users would have to be aware of the update and manually download and install it. For a web application like TWIG, however, updates are simpler to implement and are dispersed automatically when the code is updated on the server. This is to say that users are inherently always using the most up-to-date version of the software.

The technical need for specialized analysis tools

An additional challenge to conducting research that involves Twitter data is the lack of analytical tools that can effectively parse such data, because existing tools are not designed with social media data in mind. As such, tools designed to facilitate the analysis of specialized researcher-generated corpora, the norm in Twitter-based research, are not equipped to use as part of the analysis either lexical items that hold special functionality in social-media discourse (such as hashtags, @mentions, retweets), other metadata collected with the tweets (usernames, timestamps, geolocation information, and the number of likes, replies, and retweets each Tweet receives), or metadata (coding) added by the researcher (demographic or other grouping information, evaluative categorizations of the tweet's content, etc.). As a result, while a corpus of tweets might contain all of the necessary information for analyses that track use of a construction over time or compare language use between groups of users, actually performing those analyses could require complex spreadsheet functions, the development of project-specific code in Python or R, or even manual modification of the corpus. For example, to compare language use between two groups of users in AntConc, a popular open-source corpus analysis tool, once could split the corpus into two subcorpora, one for each group, and look at each separately. This adds steps to the process of analysis that might discourage researchers from performing certain kinds of analysis.

It is worth noting at this stage that, while Twitter can provide a great deal of metadata for each tweet, much of which is analytically valuable, there are limitations that affect the kinds of study that can be easily performed with only Twitter-provided metadata. For example, the provision of geolocation information must be enabled by the individual user, and most users elect to not enable it. Furthermore, demographic information such as age, gender identity, and ethnicity are not provided. As such, conducting studies in which these attributes are variables requires hand-coding of tweets by the researcher, who must have an external source for this information. To this end, user-bios, which Twitter does provide with each tweet, can serve as a valuable reference, though, of course, the researcher in this case must decide whether to trust the user. Nevertheless, once all necessary supplemental annotation is added to a dataset, it is the same as any other metadata and so is only as analytically valuable as the analytical tool allows.

The theoretical need for such tools

The layer of complexity that is added to the process of performing comparative analyses represents not just a technical concern with existing tools, but the beginnings of a set of theoretical concerns. As discussed by Vorobel and Smith (2020), in their case referring to CALL tools, software developed for and used in research should emerge from a theoretical need that can be demonstrated by existing work in the field. This is to say that software is not theoretically neutral. The kinds of analysis a tool enables and prioritizes are a direct reflection of the values of its developer, with due affordances, of course, for what is technically possible. In this light, TWIG was developed to address not only a technical need for more accessible research tools, but also a theory-driven need for tools that address longstanding concerns regarding the value of using corpus methods in the analysis of discourse.

Discourse analysis

Before addressing the role of corpus methods, however, it is important to clarify the notion of discourse analysis used in this document. This notion begins with the concept of discourse itself, which has various realizations in linguistics, sociology, political science, and elsewhere. Definitions range from a general idea of “language in use” (Brown and Yule, 1983) to the more specific idea of discourse as “a set of meanings, metaphors, representations, images, stories, statements and so on that in some way together produce a particular version of events” (Burr, 1995, p. 48). The latter definition is more representative of what is meant by discourse in research, because it creates a framework for discussing specific topical discourses. Both definitions, however, suggest specific analytical concerns. The most fundamental of these is the centrality of “the text,” in which a text can be loosely thought of as any completed written or spoken communicative scenario. For analysis of discourse on Twitter, each tweet is a text. From a discourse perspective, then, the meanings of words, phrases, and sentences are poorly determined in isolation but rather emerge from their specific use in the broader text in which they appear, and furthermore from the broader discourse in which the individual text appears. The meaning of any utterance emerges not in isolation, but as a product of how who spoke or wrote it, to whom, in reply to what, when, where, and so on. It is with this with this notion of discourse in mind that the technical need for inclusion of metadata described above becomes also a theoretical need, because the metadata is what provides this layer of information about the tweet. As such, analytical tools that lack or poorly implement access to available metadata fundamentally do not allow analysis of the full text.

Inclusion of corpus methods

Definitions like Burr's above also suggest the importance of considering how meanings emerge not just within individual texts, but from patterns of use in other texts of the same discourse. From this perspective, to analyze a discourse is to understand its frequent subjects and objects, the values and social relationships of its participants, its situation in time and space. Given this interest in patterns spanning multiple texts, it is not surprising that many discourse analysts have integrated computational corpus-based methods into their methodologies as technologies for doing so have emerged. However, the merits and shortcomings of combining corpus and discourse methodologies have long been topics of critical discussion. A dominant criticism of corpus methods is that they rely too heavily on quantitative measures, particularly of raw frequencies, abstracting possible inferences by rendering researchers "out of touch with the texts" (Egburt & Schur, 2018), stripping context from the kinds of language use being observed (Widdowson, 2000; Baker et al., 2008; Marchi and Taylor, 2018). Context here applies in multiple senses. One kind of context that is lost is information regarding things like the linguistic environment of words in frequency distributions. Baker et al. (2008) note that for this reason discourse studies often prefer use of concordances and n-gram/cluster detection, but such approaches on their own still ultimately disregard both potentially-salient context in the broader text and broader situational context about the language use being observed—information about the speaker, the purpose of the language use, and so on.

There is an extent to which dealing with smaller, specialized corpora of Twitter data inherently mitigates some of these discourse-contextual concerns, however. As opposed to large, general corpora, social-media-based discourse studies tend to be relatively contextually bound, focusing either on discourses surrounding specific topics or hashtags or on language use of specific users or groups of users. However, this still does not address the anomalies that can arise relying too heavily on quantitative measures. Egburt and Schur (2018) illustrate such concerns with the example of keyword analyses, which use frequency distributions from a general reference corpus to identify statistically more frequent "keywords" in a specialized corpus. The problem they note is that such analyses can identify keywords as statistically frequent in a corpus even if they only appear, with extreme frequency, in a few texts in the corpus. They further note that such analytical shortcomings can be addressed using more complex methodologies in which researchers take analytical steps beyond the measurement of raw frequencies. This sentiment is frequently mirrored in other discussions of the shortcomings of corpus methods (Baker et al., 2008; Gries, 2022; Middowson, 2000).

These shortcomings viewed from slightly a different perspective, then, become less a problem with corpus methodology and researcher intent, and actually more a problem of current research tools not providing adequate or accessible avenues for these next steps. Though, as Egburt and Schur (2018) note, there is also a case to be made that existing corpus tools fundamentally frame the data in a manner that is less than ideal for discourse analysis, where individual, complete texts, rather than aggregate measures, are primary units of analysis. Egburt and Schur's essential claim is as follows:

> While computer programs offer many benefits, most existing corpus software programs (e.g. concordancers, web-based corpus interfaces) analyse and present linguistic patterns at the level of the corpus, rather than the level of the text. This has created a situation in which many discourse analysts investigate and report discourse patterns in terms of results from an entire corpus rather than individual texts. (Egburt & Schur, 2018, pp. 159-160)

While existing corpus software might be lacking in terms of statistical complexity, statistically-minded corpus linguists employ a variety of methods for more robust discussion of linguistic patterns within a corpus. Egburt and Schur (2018) propose that instead of raw frequencies alone, frequency and keyword distributions might instead show mean instances per text, standard deviation, and high and low extremes. Similarly, Gries (2022) argues for the value of measures of dispersion and variability using a form of bootstrapping that uses "linguistically meaningful sampling unit[s]" (p. 8). Bootstrapping essentially involves measuring frequency not once over the entire corpus, but instead sampling frequency at intervals throughout the corpus. Sampling from "linguistically meaningful" units, rather than randomly, or even simply text-by-text, is valuable because it allows for trends to emerge within whatever units are used—such as register, author, time, or place.

With this in mind, measures of frequency and dispersion are particularly interesting in the case of Twitter data, where each text is quite short. Frequencies, on one hand, might be better represented by displaying simply the number of tweets (i.e. texts) a pattern appears in. Beyond this, though, the metadata collected with each tweet can be leveraged to create "linguistically meaningful" units for a bootstrapping approach to analyzing dispersion of linguistic patterns within a corpus of tweets. To extend this reframing of frequency for discourse purposes, a tool that effectively leverages metadata stored with the collected tweets can further break down frequencies and show differences over time, between users, groups of users, topics of discourse, and so on. Such affordances would allow for more precise discussion of frequent items and would prevent individual users or widelydiscussed but temporally-limited topics from being treated as representative of the broader corpus or dataset.

However, discussion of frequencies at all, regardless of how they're calculated, is still discussion of the text that is inherently removed from the text. For the text to be the primary unit of analysis, measurements should be supplementary, should be a starting point rather than a conclusion. This specific need is the primary motivation for the development of TWIG's analytical component, or, more precisely, the motivation for a predecessor to TWIG that I developed for a study of identity construction of a particular sub-categorizable group of Twitter users. The fundamental concept of the tool is relatively simple: provide linguistically-meaningful statistics to put full texts (in this case, tweets) in front of the researcher. This is achieved by presenting a feed of tweets that can be be narrowed, in any combination, by purposeful quantification of both the text—such as n-grams, frequent hashtags, and @mentions—and associated metadata—such as author, group affiliation, time, geolocation, tweet topic, and so on. This framework and the workflow it enables are elaborated on in the following chapter.

CHAPTER II

DESCRIPTION OF TWIG AND THE DEVELOPMENT PROCESS

The aim of this project was to develop an application that both simplifies the conducting of (many of) the kinds of linguistic studies that are currently done using Twitter data and enables the conducting of more exploratory, quantitatively-driven discourse studies of said data. TWIG was developed as a web application using web languages—HTML, CSS, PHP, and JavaScript. The decision to develop TWIG for the web, rather than as a standalone desktop application, was made for three main reasons. The primary reasons for developing TWIG as a web application are ease of access and ease of use. A tertiary reason is the ease of rapid development, especially of the user interface. In many ways, these reasons go together and define the fundamental difference in approach between TWIG and other tools, such as Tweepy, STACKS, and FireAnt. Web applications can be run on any operating system on nearly any device with a modern web browser. Users do not have to worry about installation procedures, updates, or programming environments. As a further convenience, their data is accessible at any time from any computer with an internet connection. A drawback of this is that some researchers may prefer to have their data stored locally, rather than on a server. For this reason, and others discussed below, TWIG allows researchers to download the raw data it collects for each project.

An added level of difficulty that emerges in developing a tool intended for researchers is that, because (as is described in the preceding chapter) Twitter data is used in a wide variety of studies, the design of the application must be extremely flexible, both in terms of its technical underpinnings and its user experience. From a technical perspective, maintaining flexibility means making sure the main functional components—a query design component, a data collection component, an underlying data storage component, and a set of analytical tools—are each developed with the others in mind. Data must be collected in ways that allow researchers to develop studies around individual or grouped social media users, specific linguistic constructions, collected metadata, researcher-provided metadata, or combinations of these. This data must then be stored in a way that is, in a sense, methodologically and theoretically neutral. For example, storing data hierarchically so all posts from a given user are in one place makes sense for studies that are interested in language use of specific users, but would only complicate an analysis of a specific linguistic construction, which might include posts from thousands of users who don't matter much individually. Similarly, the analytical tools must be designed to work equally well with user-centric studies as with construction-centric studies, group-centric studies, and so on. To this end, a guiding principle in the development of this application is that everything is metadata apart from the language itself, and all metadata must be treated equally by the application so that researchers can prioritize it as they choose.

Maintaining flexibility along with a degree of user-friendliness is a similar puzzle with a different set of challenges. Foremost among these challenges is developing an interface that allows the collection of tweets based on based on a variety of different parameters—and that further allows these tweets to be pre-categorized to represent, for example, different user groups or construction types between which comparisons might be made. A subsequent challenge is to develop an analysis environment in which these different types of dataset can be processed with at least a degree of intuitiveness. The following sections serve as a sort of hybrid technical guide and user manual, addressing how the needs mentioned above were considered and fulfilled in the development of the application's data collection, storage, and analysis components.

Getting started

To use TWIG, researchers must first register an account. This allows them to create and save unique, protected projects within the web app, which can then be accessed by any computer or tablet with an internet connection. Registering for TWIG requires a .edu email address, a password, and an access token connected to a Twitter developer account. Access tokens will be discussed in the data collection section below. Once logged in, users are taken to a dashboard, shown in Figure 2.1 below, from which they can create, open, and manage projects. A project is a dataset or corpus of tweets collected for a particular study and any saved analytical parameters or annotations.

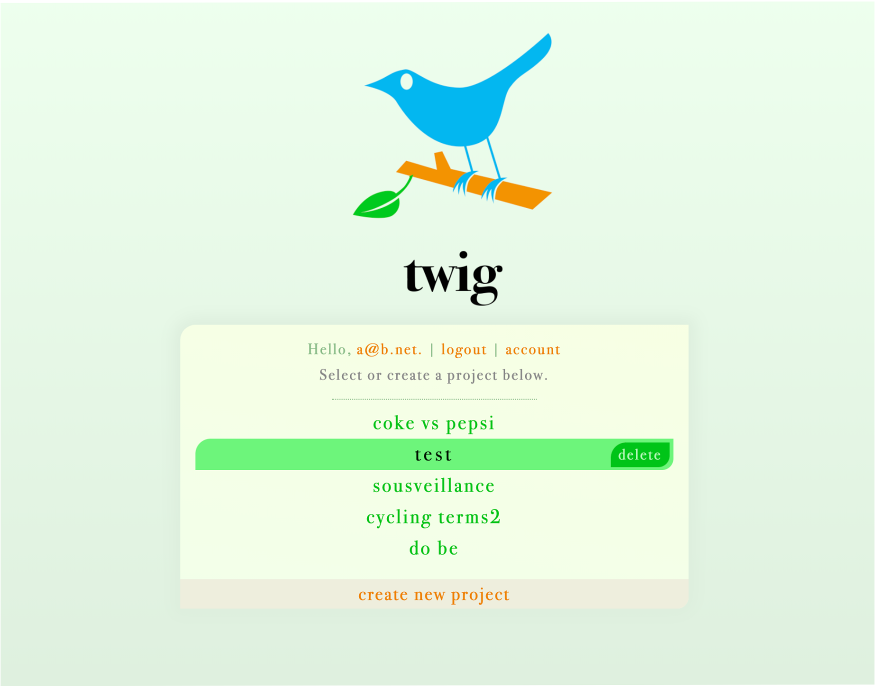
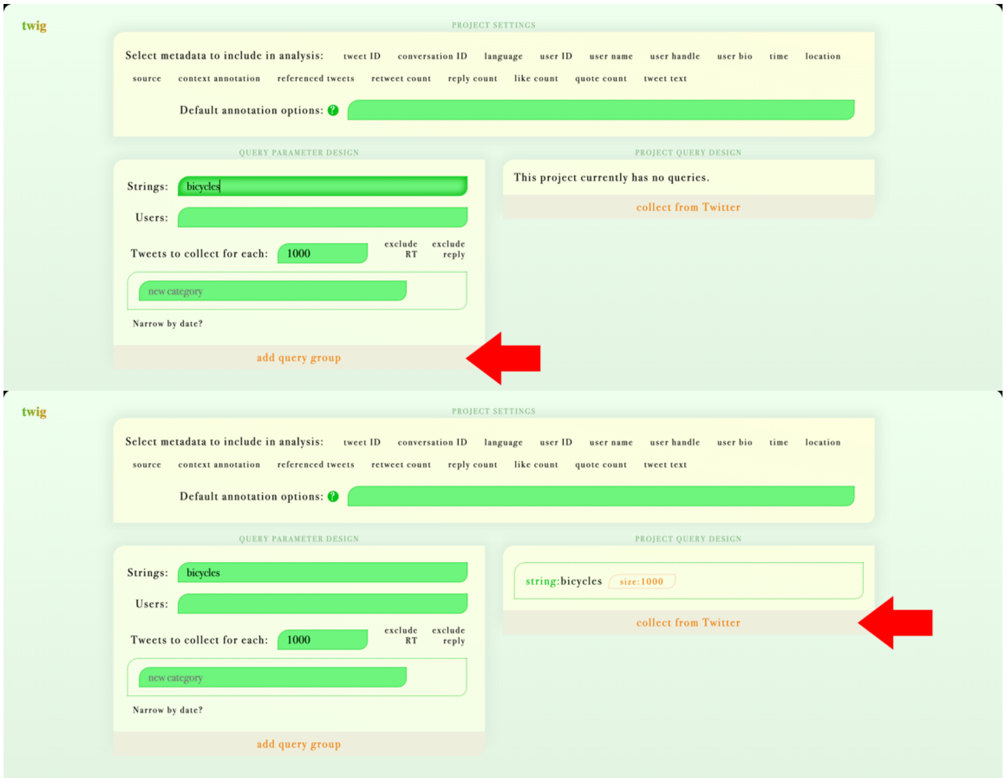


Figure 2.1 - The Twig Dashboard

When a project is opened, users will see one of two environments: the Project Design Environment or the Analysis Environment. The Project Design Environment, shown in Figure 2.2 below, appears by default in new projects and projects for which data has not yet been collected and can be accessed for modification of existing projects. This is where parameters are configured to describe both the kinds of tweets to be collected and the kinds of metadata that will be used in the analysis environment. This figure, with the overlaid red arrows, illustrates the basic process of defining tweets to be collected. Detailed discussion of each this environment's panes—Project Settings, Query Parameter Design, and Project Query Design—will be provided below, but at this stage it is useful to illustrate the fundamental process of creating projects in TWIG and how settings in the query design environment affect the behavior of the analysis environment.



Select metadata to include in analysis: tweet ID conversation ID language user ID user name user handle user bio time location source context annotation referenced tweets retweet count reply count like count quote count tweet text

Default annotation options:

Strings: bicycles

Users:

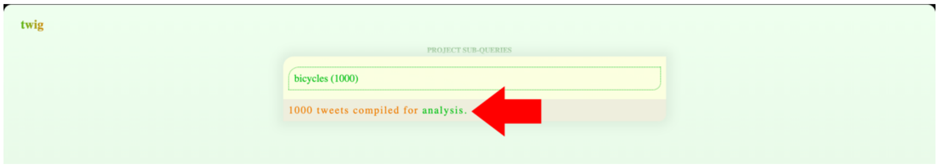
Tweets to collect for each: 1000

This project currently has no queries.

string:bicycles size:1000

Figure 2.2 - The Project Design Environment

In this example, a simple string query is defined in the Query Parameter Design pane. The only defined parameters are a string to search for, "bicycles," and the desired number of tweets, 1000. When a user clicks "add query group" at the bottom of this pane, this query appears in the Project Query Design pane as what will going forward be called a sub-query. TWIG projects can have one or more sub-queries, the applications of which will be discussed in sections below. Once a project has at least one sub-query, the user can click "collect from Twitter," which takes them to the Collection Environment, shown in Figure 2.3 below:

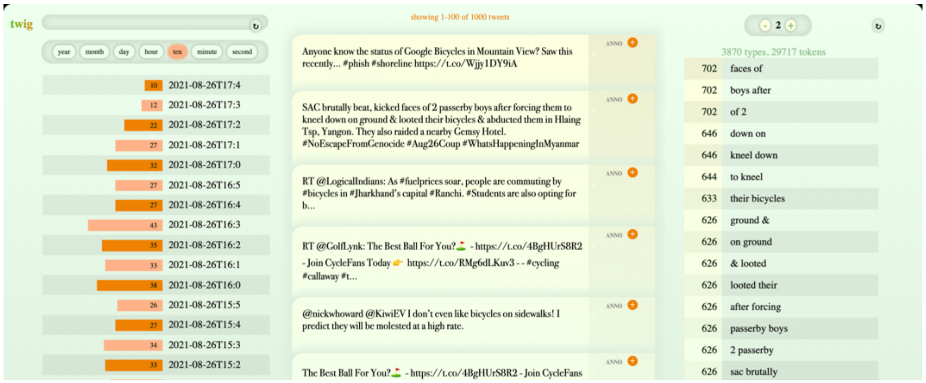


bicycles (1000)

1000 tweets compiled for analysis.

Figure 2.3 - The Collection Environment

The Collection Environment is where data is managed in TWIG. When this page is loaded, TWIG checks each sub-query to see whether tweets have been collected for it, collects tweets if needed, and compiles and formats all collected tweets into a single dataset for analysis. This environment also allows users to manually manage certain aspects of the data and collection, which will be elaborated on in the next section. Once a dataset has been compiled, the analysis environment, shown in Figure 2.4 below, is made available.



10	2021-08-26T17:4
12	2021-08-26T17:3
22	2021-08-26T17:2
27	2021-08-26T17:1
32	2021-08-26T17:0
27	2021-08-26T16:5
27	2021-08-26T16:4
43	2021-08-26T16:3
35	2021-08-26T16:2
33	2021-08-26T16:1
38	2021-08-26T16:0
26	2021-08-26T15:5
27	2021-08-26T15:4
34	2021-08-26T15:3
33	2021-08-26T15:2

showing 1-100 of 1000 tweets

Anyone know the status of Google Bicycles in Mountain View? Saw this recently… #phish #shoreline https://t.co/Wjjy1DY9iA

SAC brutally beat, kicked faces of 2 passerby boys after forcing them to kneel down on ground & looted their bicycles & abducted them in Hlaing Tsp, Yangon. They also raided a nearby Gemsy Hotel. #NoEscapeFromGenocide #Aug26Coup #WhatsHappeningInMyanmar

RT @LogicalIndians: As #fuelprices soar, people are commuting by #bicycles in #Jharkhand's capital #Ranchi. #Students are also opting for b…

RT @GolfLynk: The Best Ball For You?⛳ - https://t.co/4BgHUrS8R2 - Join CycleFans Today 👉 https://t.co/RMg6dLKuv3 - - #cycling #callaway #t…

@nickwhoward @KiwiEV I don't even like bicycles on sidewalks! I predict they will be molested at a high rate.

The Best Ball For You?⛳ - https://t.co/4BgHUrS8R2 - Join CycleFans

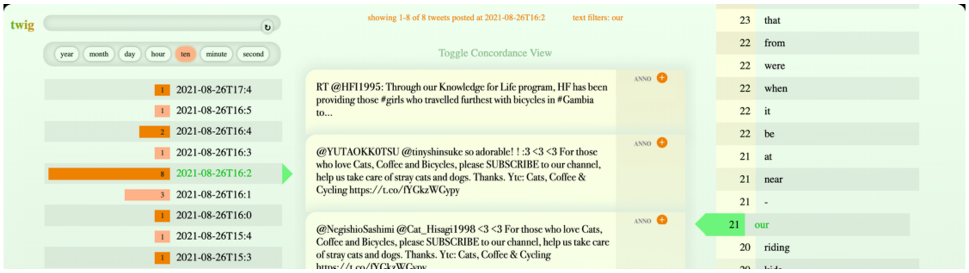
3870 types, 29717 tokens

702	faces of
702	boys after
702	of 2
646	down on
646	kneel down
644	to kneel
633	their bicycles
626	ground &
626	on ground
626	& looted
626	looted their
626	after forcing
626	passerby boys
626	2 passerby
626	sac brutally

Figure 2.4 - The Analyze Environment

The Analysis Environment fundamentally consists of three components: a central column of full-text tweets flanked by panes containing descriptive statistics at the corpus level pertaining, on the left, to metadata collected with the tweets and, on the right, to various lexical frequencies from the tweets' text. It is worth noting that the latter quantifies the number of tweets containing each lexical item, not the number of times each occurs in the corpus overall. This is one area in which TWIG attempts to address criticism of using existing corpus tools in discourse analysis (e.g. Egburt & Schur, 2018; Gries, 2022) by representing data at the scope of the individual text. This is admittedly a minor concern, considering that tweets are very short individual texts, but TWIG's analysis environment is designed to eventually scale to other kinds of text.

TWIG's Analysis Environment is designed entirely around the concept that all of these descriptive statistics are interactive. Clicking an element immediately filters the tweets that are shown to include only tweets containing the selected term. Alternatively, right-clicking excludes the selected term. The statistics on either column can then be updated to describe only this narrowed "active" dataset by clicking the refresh button in the top right of either pane. These filters can also be stacked, as shown in Figure 2.5 below and discussed later in this chapter:



showing 1-8 of 8 tweets posted at 2021-08-26T16:2   text filters: our

1	2021-08-26T17:4
1	2021-08-26T16:5
2	2021-08-26T16:4
1	2021-08-26T16:3
8	2021-08-26T16:2
3	2021-08-26T16:1
1	2021-08-26T16:0
1	2021-08-26T15:4
1	2021-08-26T15:3

RT @HFI1995: Through our Knowledge for Life program, HF has been providing those #girls who travelled furthest with bicycles in #Gambia to...

@YUTAOKK0TSU @tinyshinsuke so adorable! ! :3 <3 <3 For those who love Cats, Coffee and Bicycles, please SUBSCRIBE to our channel, help us take care of stray cats and dogs. Thanks. Ytc: Cats, Coffee & Cycling https://t.co/fYGkzWGypy

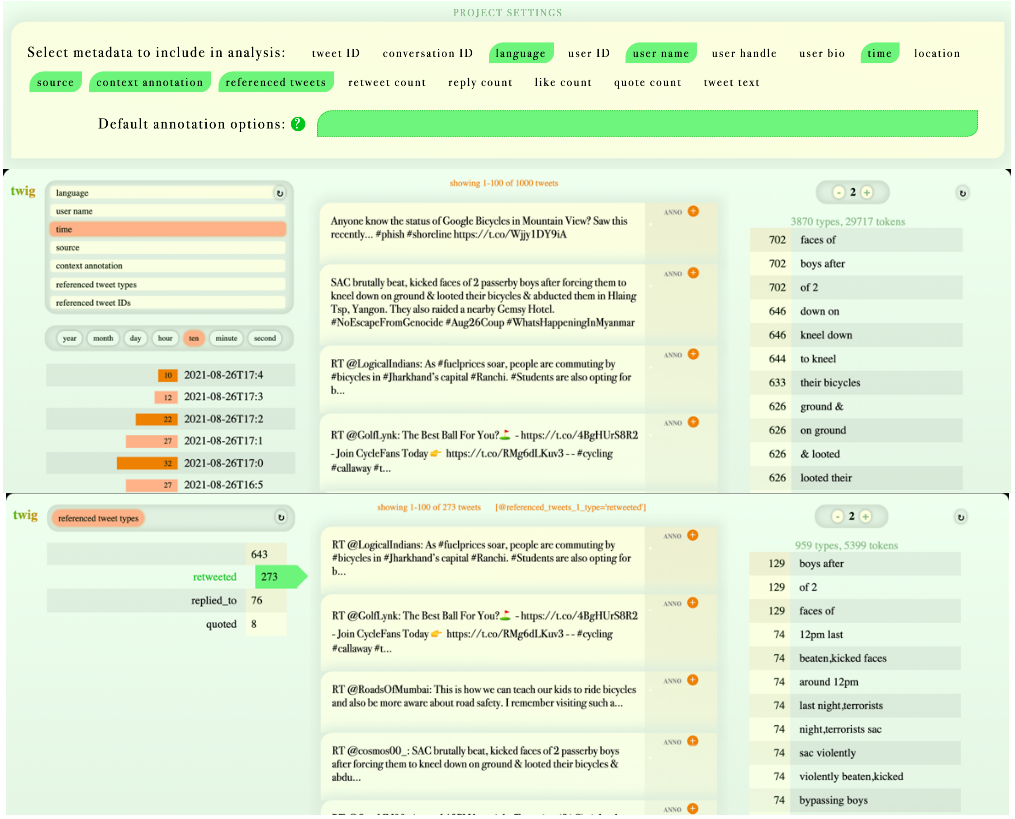
@NegishioSashimi @Cat_Hisagi1998 <3 <3 For those who love Cats, Coffee and Bicycles, please SUBSCRIBE to our channel, help us take care of stray cats and dogs. Thanks. Ytc: Cats, Coffee & Cycling https://t.co/fYGkzWGypy

23	that
22	from
22	were
22	when
22	it
22	be
21	at
21	near
21	-
21	our
20	riding

Figure 2.5 - Realtime narrowed datasets

In Figure 2.5, "our," which the dataset contains 21 instances of, was selected from the text-level column, then the counts in the metadata column were updated to show the ten-minute span with the most instances of "our" and the span with the most instances was selected to further filter the displayed tweets to only those that both contain "our" and were posted on August 26, 2021 between 16:20 and 16:29 (24-hour time). This style of filtered dataset will henceforth be referred to as the "active" dataset.

On the above example project, the only available statistics and filters in the metadata column pertain to the default temporal metadata. Additional metadata can be enabled by returning to the query design environment and configuring options within the Project Settings pane. This pane allows for the configuration of both the twitter-sourced metadata and any researcher-defined annotation options that may be desired for a given project, as shown in Figure 2.6 below.



Select metadata to include in analysis: tweet ID conversation ID language user ID user name user handle user bio time location source context annotation referenced tweets retweet count reply count like count quote count tweet text

Default annotation options:

10	2021-08-26T17:4
12	2021-08-26T17:3
22	2021-08-26T17:2
27	2021-08-26T17:1
32	2021-08-26T17:0
27	2021-08-26T16:5

showing 1-100 of 1000 tweets

Anyone know the status of Google Bicycles in Mountain View? Saw this recently... #phish #shoreline https://t.co/Wjjy1DY9iA

SAC brutally beat, kicked faces of 2 passerby boys after forcing them to kneel down on ground & looted their bicycles & abducted them in Hlaing Tsp, Yangon. They also raided a nearby Gemsy Hotel. #NoEscapeFromGenocide #Aug26Coup #WhatsHappeningInMyanmar

RT @LogicalIndians: As #fuelprices soar, people are commuting by #bicycles in #Jharkhand's capital #Ranchi. #Students are also opting for b...

RT @GolfLynk: The Best Ball For You?⛳ - https://t.co/4BgHUrS8R2 - Join CycleFans Today 👉 https://t.co/RMg6dLKuv3 - - #cycling #callaway #t...

3870 types, 29717 tokens

702	faces of
702	boys after
702	of 2
646	down on
646	kneel down
644	to kneel
633	their bicycles
626	ground &
626	on ground
626	& looted
626	looted their

	643
retweeted	273
replied_to	76
quoted	8

showing 1-100 of 273 tweets [@referenced_tweets_1_type='retweeted']

RT @LogicalIndians: As #fuelprices soar, people are commuting by #bicycles in #Jharkhand's capital #Ranchi. #Students are also opting for b...

RT @GolfLynk: The Best Ball For You?⛳ - https://t.co/4BgHUrS8R2 - Join CycleFans Today 👉 https://t.co/RMg6dLKuv3 - - #cycling #callaway #t...

RT @RoadsOfMumbai: This is how we can teach our kids to ride bicycles and also be more aware about road safety. I remember visiting such a...

RT @cosmos00_: SAC brutally beat, kicked faces of 2 passerby boys after forcing them to kneel down on ground & looted their bicycles & abdu...

959 types, 5399 tokens

129	boys after
129	of 2
129	faces of
74	12pm last
74	beaten,kicked faces
74	around 12pm
74	last night,terrorists
74	night,terrorists sac
74	sac violently
74	violently beaten,kicked
74	bypassing boys

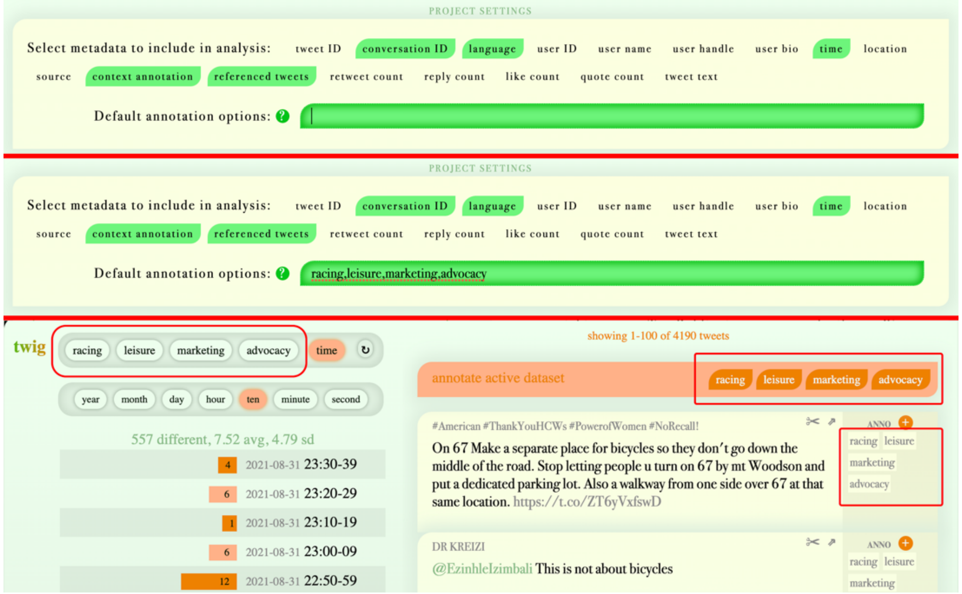
Figure 2.6 - Enabling metadata filters

Figure 2.6 shows, first, the updated configuration of the Project Settings, in which several metadata options are enabled and, in the middle and bottom images, how these updates are reflected in the analysis environment. Note that in the middle image the dropdown of options above the metadata column match what is enabled above. In the bottom image, the metadata column is set to show the project's distribution of tweet types, and the tweet feed and text-level column have been narrowed to show only retweets. Metadata is enabled by selecting or deselecting items from a pre-defined list of available information that TWIG collects about individual tweets—such as temporal and geographical information, how many times the tweet has been liked, retweeted, etc.—and information about the users who posted them. This list is derived from what is available from the Twitter API (Twitter, 2021a, 2021b). A description of each of these options and some possible analytical applications are presented in Table 2.1 below.

Metadata Name	Description	Application
user ID	A unique numeric identifier referring to the specific user who created the tweet.	Because lower user IDs indicate older accounts, may be useful in analyzing Twitter norms over time.
user handle	A unique text identifier chosen by the user who created the tweet. Notably used in @mentions.	In user-centric studies, enables rapid comparison of different users.
user name	A not-necessarily unique display name, chosen by the user. Often but not always a real name.	As above. Also, more likely to be a recognizable name than the handle.
user bio	A short self-description users can define in their account setting, visible when viewing their timeline.	Can be considered in isolation or in conjunction with tweet text in identity construction studies.
tweet ID	A unique numeric identifier referring to the current tweet.	Used by TWIG to generate hyperlinks to tweet in context.
conversation ID	A numeric identifier assigned to a tweet and to any tweets in the associated reply chain.	Useful for finding interactions among tweets in a dataset.
language	The language of the tweet, determined by Twitter.	Can be used to narrow scope of dataset or
time	When the tweet was posted, including year, month, day, hour, minute, second.	Useful in diachronic work when frequency or rate of tweeting are salient factors.
location	The geographical location, if available, from which the tweet was posted.	Rarely available, but incredibly useful for analyses of regional variation or localized discourse.
source	The type of device from which the tweet was posted.	For analyses of public figures who may not always do their own tweeting.
context annotation	Topical annotation provided automatically by the Twitter API.	Can be a quick way of identifying broad topical patterns among subject(s) or in usage of a particular construction.
referenced tweets	Includes tweet ID and relationship for retweets, quote tweets, and replies	Useful in narrowing dataset to show only certain kinds of interaction (or to exclude interaction)
retweet count, quote count, reply count, like count	The number of times the tweet was retweeted, quoted, replied to, or liked.	Useful in identifying interaction or tweet reach. Reply count can be a useful way of finding long reply chains for analyses of conversation on Twitter.
tweet text	The full text of the tweet.	TWIG allows this as a form of "metadata" because it can be a useful way of identifying tweets that originate from automated accounts.

Table 2.1 - Twitter-sourced metadata

Below the metadata selection options in the Project Settings pane is the annotation options field. This field allows users to streamline the analysis process by defining study-specific annotations, which in TWIG refer to user-defined true/false metadata options for which the values are assigned manually by the researcher as part of a post-data-collection coding process. An example of this process is shown in Figure 2.7 below.


PROJECT SETTINGS

Select metadata to include in analysis: tweet ID, conversation ID, language, user ID, user name, user handle, user bio, time, location, source, context annotation, referenced tweets, retweet count, reply count, like count, quote count, tweet text

Default annotation options:

PROJECT SETTINGS

Select metadata to include in analysis: tweet ID, conversation ID, language, user ID, user name, user handle, user bio, time, location, source, context annotation, referenced tweets, retweet count, reply count, like count, quote count, tweet text

Default annotation options: racing,leisure,marketing,advocacy

showing 1-100 of 4190 tweets

557 different, 7.52 avg, 4.79 sd

4	2021-08-31	23:30-39
6	2021-08-31	23:20-29
1	2021-08-31	23:10-19
6	2021-08-31	23:00-09
12	2021-08-31	22:50-59

#American #ThankYouHCWs #PowerofWomen #NoRecall!

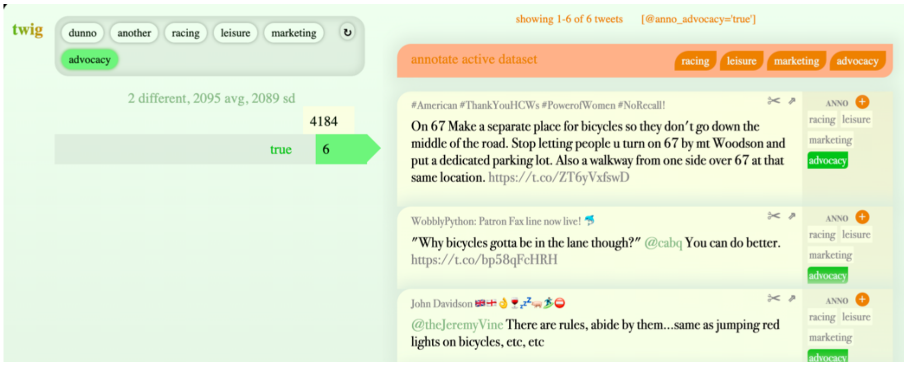
On 67 Make a separate place for bicycles so they don’t go down the middle of the road. Stop letting people u turn on 67 by mt Woodson and put a dedicated parking lot. Also a walkway from one side over 67 at that same location. https://t.co/ZT6yVxfswD

DR KREIZI

@EzinhleIzimbali This is not about bicycles

Figure 2.7 - Procedure for updating project annotation options

Annotation options here should be divided by commas, as shown in the middle image above. Changes to annotation options are saved by pressing return, and the successful save is visually confirmed by the Project Settings pane’s shadow pulsing green. Unlike changes made to metadata configuration, which immediately take effect in the analyze environment, changes here do not take effect until the project data has been recompiled, which is done by returning to the Collect Environment and selecting the “recompile,” an option that appears when the mouse is hovered over the bottom of the Project Sub-Queries pane. At this stage, the Analysis Environment will be updated in several ways, which are indicated by red boxes. Primarily, each tweet in the central column will be presented with available annotations so that a researcher can rapidly annotate or code tweets as they scroll through. Additionally, all tweets in an active (narrowed) dataset can be annotated at once using the “annotate active dataset” bar above the tweets. Annotated tweets can then be isolated using the Metadata Analysis Pane, as shown in Figure 2.8.



#American #ThankYouHCWs #PowerofWomen #NoRecall!

On 67 Make a separate place for bicycles so they don't go down the middle of the road. Stop letting people u turn on 67 by mt Woodson and put a dedicated parking lot. Also a walkway from one side over 67 at that same location. https://t.co/ZT6yVxfswD

WobblyPython: Patron Fax line now live! 🐬

"Why bicycles gotta be in the lane though?" @cabq You can do better. https://t.co/bp58qFcHRH

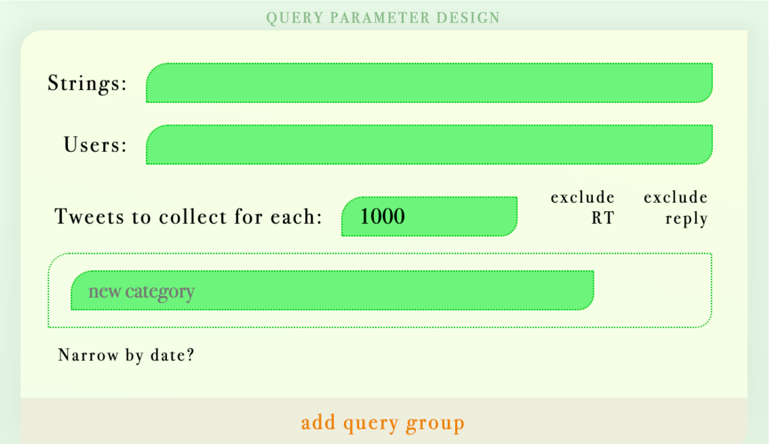
John Davidson 🇬🇧🏴󠁧󠁢󠁥󠁮󠁧󠁿👌🍷💤🐖🏃⛔

@theJeremyVine There are rules, abide by them...same as jumping red lights on bicycles, etc, etc

Figure 2.8 – Active dataset narrowed by annotation value

Advanced project and query design

The above examples demonstrate the basic use of TWIG for analysis of a simple single-query project. However, many linguistic studies of Twitter discourse are much more specific about the kinds of tweets they collect, who they’re from, and how they can be categorized for analysis. To these ends, the query design environment enables a number of advanced options for project design. In more advanced projects, it becomes more important to understand how the Query Parameter Design pane and the Project Design pane function together to define all of the tweets to be collected for a given project. The Query Parameter Design pane is where all of the user input happens. The input fields on this pane—shown in Figure 2.9 below and detailed in the following paragraphs—provide TWIG with information needed to both construct Twitter API calls and format the final dataset used by the analysis environment.

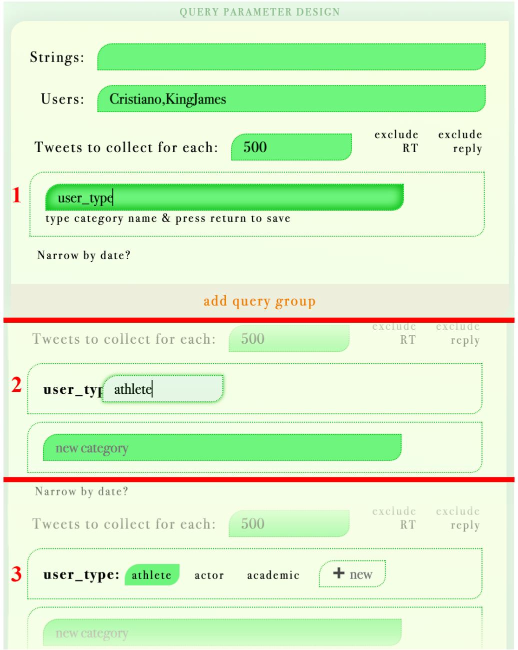


Tweets to collect for each: 1000

Figure 2.9 - The Query Parameter Design pane

Within the Query Parameter Design pane, the main parameters are “strings” and “users.” String queries tell TWIG to collect tweets based on textual search criteria, such as hashtags or key phrases. User queries tell TWIG to collect tweets from one or more specific users, separated by commas. Either, but not both, of these can be left blank. Combined, they can be used to collect only tweets from specific users that contain specific language. Beneath these fields are a field for specifying the number of tweets to collect and options for excluding the collection of retweets and reply tweets. It’s worth noting that the number entered into the quantity field does not specify a total for the project or even for the current sub-query, but rather a total to collect for each API call generated by the sub-query. So for example, for a sub-query with the string parameter “bicycles,skateboards” and the quantity field set to 500, TWIG would collect 500 tweets containing “bicycles” and 500 containing “skateboards,” for a total of 1000 tweets. If two users were added to this query, TWIG would collect the same for each user (provided they each had posted sufficient matching tweets), for a total of 2000 tweets. Skipping the “new category” field for a moment in the movement down the Query Parameter Design pane, at the bottom of the pane is a field in which users can setting start and end dates for their queries.

The parameters described above are all of the controls over what tweets will be collected. The remaining "new category" field beneath the quantity field marks a return to controls that affect how tweets are displayed in the analysis environment. This field opens a panel of tools for configuring researcher-provided metadata, referred to in TWIG as "categories," for the users or language features being queried. Like the researcher-provided annotation options set in the Project Design field, this configuration does not affect the tweets to be collected, but rather how they can be viewed in the analysis environment. The crucial difference between this metadata and the annotation options is that category values are assigned automatically during the data compilation process. This is where researchers can, for example, group users of interest by demographic or social information. Figure 2.10 below illustrates the process of adding a "user type" category and assigning the options "athlete," "actor," and "academic."



QUERY PARAMETER DESIGN

Strings:

Users: Cristiano,KingJames

Tweets to collect for each: 500

1 user_type

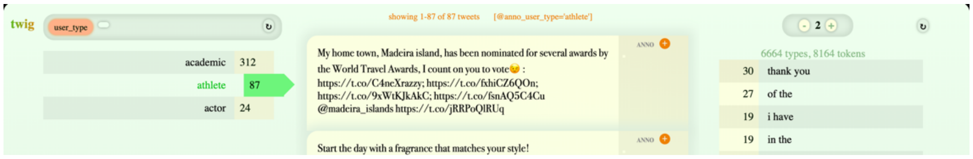
type category name & press return to save

2 user_typ athlete

3 user_type: athlete actor academic

Figure 2.10 - Defining and setting sub-query category parameters

This sub-query begins with the usernames of Cristiano Renaldo and LeBron James, who are to be categorized in this sample project as athletes. To do this, a researcher must first define a broad category by entering a value in the “new category” field and pressing return. This creates a new area in the Query Parameter Design pane just above the “new category” field, and automatically moves the keyboard cursor into a new text field in which possible values for the just-created category can be defined, by again pressing return, which adds the keyboard cursor to a new field for the definition of subsequent values for the current category. This process is shown in B, above, in which “athlete” has been typed, and in C, where the additional options “actor” and “academic” have been defined and the “athlete” option has been selected for the current sub-query. When tweets and data are collected for this project, the category will appear as an option in the analysis environment’s metadata column and the values become options for narrowing the active dataset, as shown below in Figure 2.11.



academic	312
athlete	87
actor	24

showing 1-87 of 87 tweets [@anno_user_type='athlete']

My home town, Madeira island, has been nominated for several awards by the World Travel Awards, I count on you to vote😉 : https://t.co/C4neXrazzy; https://t.co/fxhiCZ6QOn; https://t.co/9xWtKJkAkC; https://t.co/fsnAQ5C4Cu @madeira_islands https://t.co/jRRPoQlRUq

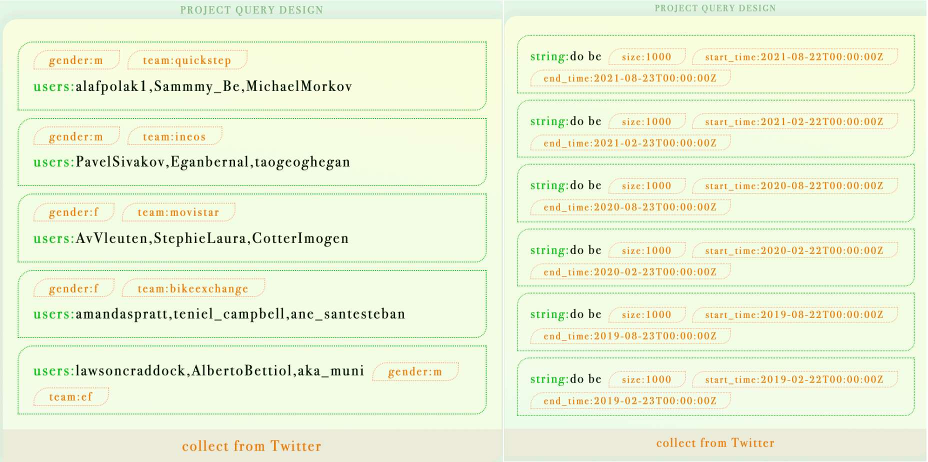
Start the day with a fragrance that matches your style!

6664 types, 8164 tokens

30	thank you
27	of the
19	i have
19	in the

Figure 2.11 - Filtering and quantifying by collection category

Completed sub-queries appear as a list on the Project Query Design pane. Each sub-query corresponds to one or more calls to the Twitter API. This method of defining project queries with multiple sub-queries can be used to collect complex datasets for a wide variety of studies. Figure 2.12 below shows query design schema for two specialized datasets, one for a study that involves groups of users and another for a diachronic investigation of a particular construction.



gender:m team:quickstep
users:alafpolak1,Sammmy_Be,MichaelMorkov

gender:m team:ineos
users:PavelSivakov,Eganbernal,taogeoghegan

gender:f team:movistar
users:AvVleuten,StephieLaura,CotterImogen

gender:f team:bikeexchange
users:amandaspratt,teniel_campbell,ane_santesteban

users:lawsoncraddock,AlbertoBettiol,aka_muni gender:m
team:ef

string:do be size:1000 start_time:2021-08-22T00:00:00Z
end_time:2021-08-23T00:00:00Z

string:do be size:1000 start_time:2021-02-22T00:00:00Z
end_time:2021-02-23T00:00:00Z

string:do be size:1000 start_time:2020-08-22T00:00:00Z
end_time:2020-08-23T00:00:00Z

string:do be size:1000 start_time:2020-02-22T00:00:00Z
end_time:2020-02-23T00:00:00Z

string:do be size:1000 start_time:2019-08-22T00:00:00Z
end_time:2019-08-23T00:00:00Z

string:do be size:1000 start_time:2019-02-22T00:00:00Z
end_time:2019-02-23T00:00:00Z

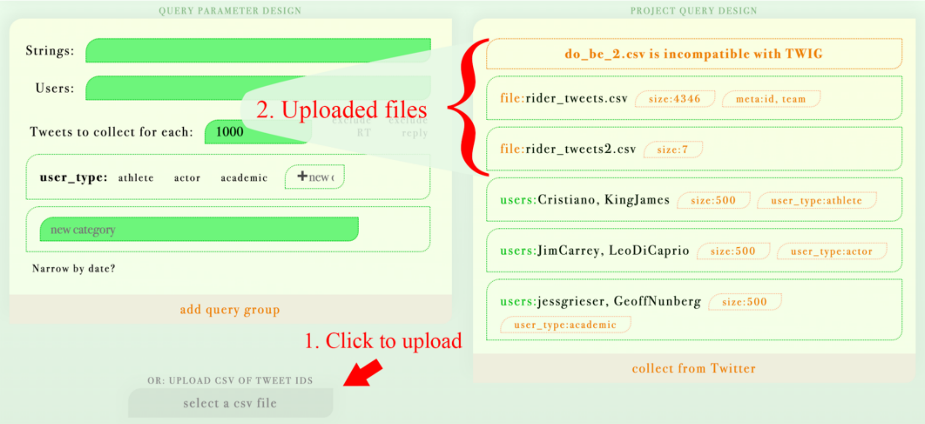
Figure 2.12 - Complex project query designs

The project design on the left is an example of a query-design schema for a study interested in comparing tweets made by different groups of users, in this case tweets from professional cyclists categorized by both team affiliation and sex. To achieve this, categories and values are created for both gender and team affiliation, and each group of subjects, in this case members of a particular gender and a particular cycling team, is assigned to a particular sub-query with the appropriate category values selected.

The project design on the right demonstrates a query-design schema that could be used to track diachronic change of a single construction, "do be." This is achieved by creating several sub-queries that request the same string query but with a range of different start and end dates. Such an approach can be preferable to a single wide-date-range query, because the frequency with which tweets are posted worldwide can make it difficult to get to older tweets. For example, a dataset of 5,000 tweets containing the string "do be" represents, at time of writing, a time span of about 12 hours. As such, sampling at intervals is a much more practical solution.

Re-collecting existing datasets

In addition to the query design procedures described above, TWIG includes a mechanism, shown in Figure 2.13 below, for collecting datasets from CSV (comma-separated values) files that include tweet IDs. This enables researchers to re-create existing datasets within TWIG or to share datasets in a manner specifically permitted by Twitter's developer agreement (Twitter, 2020b). Compatible CSV files take the form, demonstrated in Figure 2.14 below, of either a simple list of Tweet IDs or a multi-column table that includes Tweet IDs and metadata groups and values.



Tweets to collect for each: 1000

user_type: athlete actor academic

do_be_2.csv is incompatible with TWIG

file:rider_tweets.csv size:4346 meta:id, team

file:rider_tweets2.csv size:7

users:Cristiano, KingJames size:500 user_type:athlete

users:JimCarrey, LeoDiCaprio size:500 user_type:actor

users:jessgrieser, GeoffNunberg size:500 user_type:academic

Figure 2.13 – Uploading a CSV

rider_tweets.csv

```
ID, team
968516743965691904,bmc
968051352860585984,bmc
959155452532506625,bmc
955031435630452737,bmc
954478110854012928,boh
954355993886056449,boh
954331156757000192,boh
954307387292311552,boh
953966001405689856,boh
936507572517064705,una
935393956573061121,una
932793577939152896,una
929651906057609217,una
926917284802723841,una
```

rider_tweets2.csv

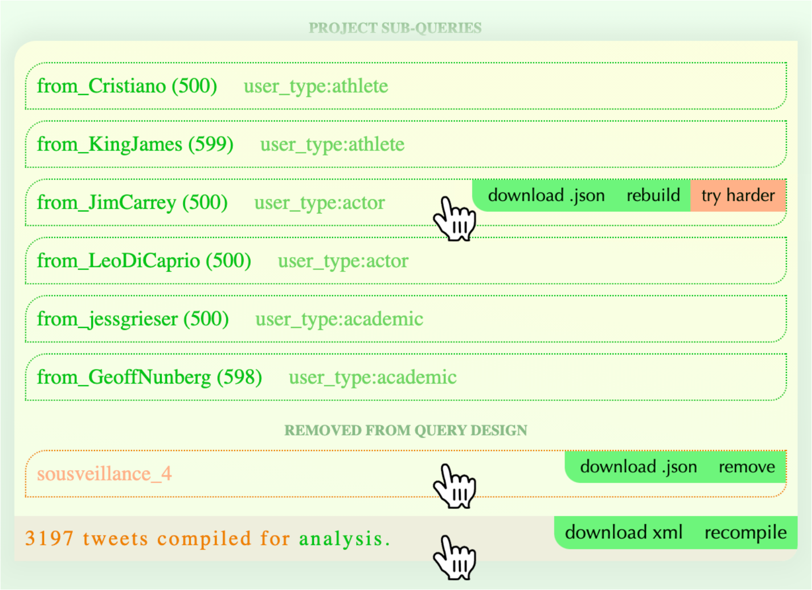
```
968516743965691904,968051352860585984,
959155452532506625,955031435630452737,
954478110854012928,954355993886056449,
954331156757000192,954307387292311552,
953966001405689856,936507572517064705,
935393956573061121,932793577939152896,
929651906057609217,926917284802723841
```

Figure 2.14 – Compatible CSV files

CSVs that include metadata should include a header row, which should include a column labeled “ID,” “Tweet ID,” or “Tweet_ID” and additional columns labeled for the metadata they represent. For example, rider_tweets.csv in Figure 2.14 includes a “team” column, which categorizes each tweet according to the bicycle racing team for which the tweet’s author races. As shown in Figure 2.13, CSV files that are not compatible with TWIG are immediately flagged as incompatible. Incompatible files are automatically deleted. Compatible CSV files are displayed among the project’s other query parameters along with information about how many tweets will be collected with each file and what, if any, metadata will be automatically assigned to the collected tweets.

Data collection and management

For TWIG users, after the query design process is complete, collecting data is typically an automatic process. However, the collect environment is not an entirely static one. It includes functionalities for managing the dataset after changes have been made to the query design, for managing unsuccessful collection from the Twitter API, and for downloading data for analysis outside of TWIG. These functions are controlled by a variety of buttons, shown in Figure 2.15, which appear when the user’s mouse hovers over certain elements of the collect environment.



PROJECT SUB-QUERIES

from_Cristiano (500) user_type:athlete

from_KingJames (599) user_type:athlete

from_JimCarrey (500) user_type:actor

from_LeoDiCaprio (500) user_type:actor

from_jessgrieser (500) user_type:academic

from_GeoffNunberg (598) user_type:academic

REMOVED FROM QUERY DESIGN

sousveillance_4

3197 tweets compiled for analysis.

Figure 2.15 - TWIG's collection environment, showing mouseover options

The image in Figure 2.15 has been edited to show the Collect Environment's different control menus in one image instead of three. At the top, illustrated by the JimCarrey sub-query, are the controls made available for individual sub-queries: "download.json," "rebuild," and "try harder." The "download.json" button allows users to download a file containing the raw JSON returned by the Twitter API for that particular sub-query. The nature of this data will be discussed below. The option to download enables users to store local backups of the tweets they collect or to analyze collected tweets outside of TWIG. The "rebuild" and "try harder" buttons in the sub-query menu are both used to re-collect tweets for that specific sub-query, which is occasionally necessary for cases in which initial collection attempts fail or collect fewer than the desired number of tweets. The "rebuild" button is best used in the case of a failed collection attempt, which is often simply the result of a server timeout or an unstable moment in the user's internet connection. The "try harder" button, on the other hand, is best used when the initial collection gets too few tweets. This repeats the query using a slightly different collection algorithm, which is discussed in detail below.

In projects where sub-queries have been removed within the query design environment, previously-collected data remains accessible within the collect environment under the heading “removed from query design.” As shown in Figure 2.15, the user controls for managing these elements include only the option to download raw JSON data and the option to completely remove this data from the project.

Below the list of sub-queries and removed elements in the collect environment is a pane representing the compilation of all of a project’s collected tweets into a single dataset formatted, as an XML file, specifically for TWIG’s analysis environment. Management controls for this pane allow users to download this XML dataset, again for backup or external analysis, and to recompile this dataset. Recompiling is not automatic and is necessary for changes made in the query design environment to be reflected in the analysis environment.

Development of the data collection and management environment

All datasets collected by TWIG are created by accessing the Twitter API via cURL (Client URL) requests. A cURL request is a method of retrieving structured data using the syntax of web urls. The Twitter API uses a number of URL “endpoints” (Twitter, 2021c, 2021d) for different kinds of requests and returns data in the form of a JSON (JavaScript Object Notation) string, which is a common form of structured data. For example, Figure 2.16 below shows first the URL and then the returned JSON data for a simple request for tweets containing “bicycle.”

```
URL: https://api.twitter.com/2/tweets/search/all?query=bicycles

RESPONSE:
{"data":[
{"id":"1432853169097347076",
"text":"RT @UntitledHK: This is Gilgil Town. The boy was going home with his
sister after school. There are so many school children using bicycles..."},
{"id":"1432852621321179141","text":"@dens_club @melalienetwork Melalie is adding
more and more electric vehicles to its platform and will also give priority to
older bikes. Currently, what percentage of electric vehicles and bicycles are
available on the platform in relation to gasoline vehicles? Are you looking for
ecology on your platform?"},{"id":"1432852611082997764","text":"@theJeremyVine In
what way are they safer, greener or healthier than bicycles?"}, [...]
],"meta":{"newest_id":"1432853169097347076","oldest_id":"1432847769228697602","re
sult_count":10,"next_token":"b26v89c19zqg8o3fpdp73a4uyzbd7ihvo3tudgd7mtl6l"}}
```

Figure 2.16 - Basic Twitter API request URL and JSON response

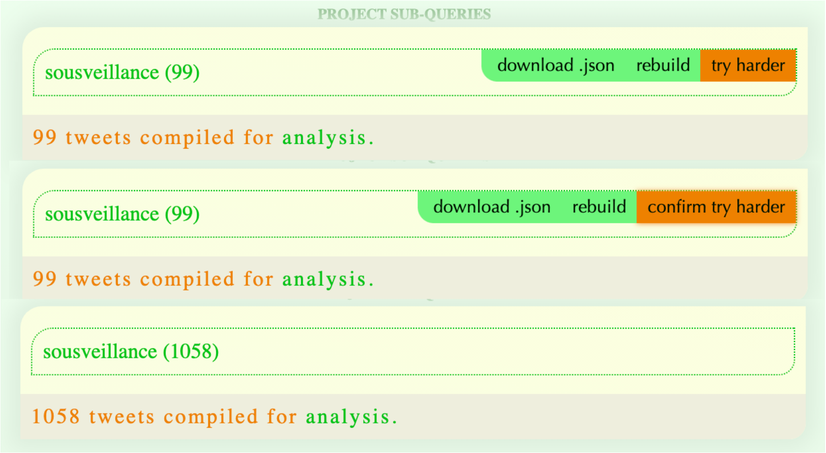
This response has been truncated to show only three of the ten tweets retrieved, in the “data” branch of the JSON response. Note beneath the “data” branch of the response is a “meta” branch, which contains information about the tweets collected, such as newest and oldest IDs in the set. The first tweet “object” here has been bolded for ease of reading. Note that the only information provided is a tweet ID and the text of the tweet. Additional URL parameters are required to collect the metadata that TWIG uses. This and the full structure of the Twitter API’s raw JSON response will be discussed below in the section data structure.

Before discussing the data it provides, however, it is worth discussing the API itself in more detail, because TWIG is specifically optimized to leverage the Twitter API to provide data as simply and effectively as possible for academic users. This is achieved in part by using only endpoints associated with Twitter’s relatively new v2 API. This version of the API was introduced in early 2021 and grants academic researchers much more robust access than older endpoints. To get this access, academic users must first register a developer account and request access to the “academic track,” which is a process that requires researchers to describe their research and their data needs and to prove their university affiliation (Twitter, 2022a). The process is, however, worth the hassle for the benefits it provides. The example in Figure 2.16 uses the Twitter v2 API tweet lookup endpoint (/2/tweets/search/all), which is a full-archive search that can return tweets dating back to Twitter’s 2007 inception and is available only to academic users. Other users have access only to the recent search (/2/tweets/search/recent), which provides “filtered” results from only the last week (Twitter, 2021c). Another advantage of the academic track is throughput. Where academic users can collect up to ten million tweets per month, basic users are limited to five hundred thousand. At time of writing, v2 endpoints and academic-track access are not supported by STACKS, or FireAnt—but TWIG is designed for them.

In addition to its implementation of academic-track API access, TWIG uses methods not described in Twitter's API documentation to better collect results. These were arrived at by experimenting with different endpoints and query constructers to ensure that TWIG delivers a number of tweets as close as possible to what users request. The need for such experimentation became apparent during testing, when certain kinds of queries returned only a fraction of the requested treats, despite their being constructed in line with the Twitter API documentation. For example, while the tweet lookup endpoint accepts an instruction to collect tweets from a specific user, it is very inefficient at actually doing so. In testing, the tweet lookup endpoint was only able to collect three tweets from Cristiano Renaldo, a famous soccer player who is quite active on Twitter. The story was much the same with several other user queries. However, requesting the same users' tweets via the user timeline endpoint reliably returned the requested number of tweets, which in testing was 500. As such, TWIG automatically uses this endpoint for user queries, with no additional steps for the researcher.

During testing, it was also found that string queries involving uncommon words would often return less than the requested number of tweets. For example, a request for 1000 tweets containing "sousveillance" would return only 99 tweets. After examination of the raw JSON response, it became clear that this was because of a problem with the Twitter API's pagination feature. Pagination is the documented way of retrieving more than 100 tweets from the Twitter API. Pagination, in some form, is required because each individual request can collect a maximum of 100 tweets, so to collect larger datasets the same base request must be repeated, with an additional parameter to prevent the collection of duplicate tweets. The API documentation's prescribed parameter is a "next_token," which is returned in the "meta" section of each JSON response (see the bottom of Figure 2.16 above). The basic procedure is to use this token in constructing the next request, which returns a new token for yet another request, and so on. Unfortunately, for unclear reasons, not all queries produce responses that include next_tokens. This is not because more tweets do not exist.

To circumnavigate the unreliability of the documented pagination method, TWIG implements an optional second method of pagination that instead of using a next_token uses the oldest tweet ID in each API response to perform a similar function. However, unlike the method for maximizing the results of user queries, this method is not the default. The oldest ID technique is slightly less stable than standard pagination, such that an error could occur if more tweets are requested than exist for a particular query. As such, this kind of pagination is available only as an option within the Collect Environment. If a researcher does not receive as many Tweets as they would like for a particular query, clicking the "try harder" button and its confirmation results in re-collection using the oldest ID technique, as shown in Figure 2.17 below:



sousveillance (99)

99 tweets compiled for analysis.

sousveillance (99)

99 tweets compiled for analysis.

sousveillance (1058)

1058 tweets compiled for analysis.

Figure 2.17 - The "try harder" function

Data structures

Once TWIG has collected all of the tweets for a project, that data is then compiled and restructured into a single dataset for analysis. To best understand the nature of and need for restructuring the data, it is worth looking more closely at the Twitter v2 API’s default JSON structure. As noted above, in order to retrieve all of the metadata that TWIG collects, parameters are required that were not shown in Figure 2.16. A more robust request URL, equivalent to what TWIG would use for a simple string query for “bicycles”, is shown in Figure 2.18, followed by the JSON response.

```
URL:
https://api.twitter.com/2/tweets/search/all?query=bicycles%20is:retweet&max_resul
ts=2&expansions=author_id&tweet.fields=context_annotations,author_id,created_at,g
eo,conversation_id,lang,public_metrics,referenced_tweets,source&user.fields=id,na
me,username,description

RESPONSE:
{"data":[
{ "source":"Twitter Web App",
  "referenced_tweets":
    [{"type":"retweeted","id":"1432832663182774280"}],
  "id":"1432849952846274563",
  "author_id":"3154858995",
  "text":"RT @RubyPerry11: https://t.co/QOX5ozZ7PP FREE SHIPPING #lamps #lights
#lighting #nightlights #freeshipping #etsy #gifts #giftideas #Retro #...",
  "conversation_id":"1432849952846274563",
  "created_at":"2021-08-31T23:37:04.000Z",
  "context_annotations":
    [{"domain":{"id":"45","name":"Brand Vertical", "description":"Top level
entities that describe a Brands
industry"},"entity":{"id":"781974596706635776","name":"Retail"}},
     {"domain":{"id":"46","name":"Brand Category","description":"Categories
within Brand Verticals that narrow down the scope of
Brands"},"entity":{"id":"783335558466506752","name":"Online"}},
     {"domain":{"id":"47","name":"Brand","description":"Brands and Companies"},
"entity":{"id":"10051086127","name":"Etsy", "description":"Etsy"}}],
  "public_metrics"{ "retweet_count":2, "reply_count":0,
"like_count":0,"quote_count":0},
  "lang":"en"},[...]
"includes":{"users":[
{ "description":"https://t.co/fA52BpKRpV...\n#bowlcovers #kitchendecor \nLove to
make home decor and Jewelry find my shop Crystalscraftycorner @ Etsy and also @
Amazon Handmade",
  "id":"3154858995", "name":"Crystal Willis", "username":"ckwlpn"}],
"meta":{"newest_id":"1432855374999212035","oldest_id":"1432843276554407937",
"result_count":10,"next_token":"b26v89c19zqg8o3fpdp73a4usvrcavk83eq1dzawoilx9"}}}
```

Figure 2.18 - Extended Twitter API request URL and JSON response

This response has been formatted for readability and truncated to show only one tweet, because, clearly, tweets collected with this method include much more metadata than the basic request shown in Figure 2.16. Note that the additional data collected corresponds almost perfectly with the metadata described in Table 2.1 above. Note also that this metadata as provided by Twitter is not flat, meaning there are attributes stored within hierarchies. For example, the “referenced_tweets” element, which has been colored orange above for visibility, has two child elements that separately provide the kind of reference performed by the tweet, in this case a retweet, and the tweet ID of the referenced tweet. Metadata about the user who posted the tweet, shown in blue, is even less directly accessible, given that it is stored in a “user object” on a completely different branch of the JSON tree from that where tweets are stored. The only connection between this user object and the corresponding tweet object is a matching user id, called “author_id” in the tweet and simply “id” in the user object. These aspects of how the v2 API structures the data it returns is unique, differing even from that of earlier versions of the Twitter API—and this uniqueness complicates the possible practicality of directly using the raw data for analysis procedures.

Decisions made in the structure of datasets created by queries

A key aspect of any research corpus or dataset is the manner in which the data is stored. The fundamental requirement of any method for storing research data is that it should maximize analyzability. In the current case, that means first maximizing the ease first with which software can differentiate between individual tweets and then maximizing the ease of accessing each tweet’s associated metadata. Several options exist, each with advantages and disadvantages. A simple option would be to retain the JSON objects returned by the Twitter API. However, because these objects, as discussed above, include nested and remote metadata, analytical software would need separate code to specifically access each kind of metadata. In addition to being time-consuming and inefficient, this would require major updates to the analysis software every time Twitter changes something about the structure of the JSON response, which, again, happens from time to time. Furthermore, because TWIG also allows the inclusion of researcher-added metadata, that metadata would have to be either inserted into Twitter’s JSON or accessed via reference tables—both of which add unnecessary complexity. For these reasons of maintaining compatibility and simplicity, it makes sense to convert Twitter’s JSON data into a more targeted analyzable form.

Such conversions are common in existing Twitter-based research. Several of the studies looked at in the development of TWIG structured their data using database software such as

Mongo DB (e.g. Hemsley et al., 2015) or SQL (e.g. Yaqub et al., 2017). Many others used CSV documents (e.g. Larsson & Moe, 2012), which are essentially plain-text spreadsheets. TWIG converts Twitter's JSON into XML. XML (Extensible Markup Language) is a text-based format that can store relatively large amounts of data in a highly-searchable manner (Quin, 2016). XML was chosen over database solutions as the format for TWIG's analysis datasets because it is more flexible and portable. The flexibility XML provides over databases has to do with the number of kinds of metadata that are stored. Databases are designed to store data in "tables" with a set configuration of "columns," or slots for information such as "username" and "tweet text." Because TWIG allows the inclusion of user-defined metadata, each project would require a separate table design, which would have to be updated every time changes were made to the user-defined metadata options. This is feasible, but more complex than using XML. Using a database would also complicate allowing users to download the final compiled dataset.

The XML produced by TWIG is structurally very simple. The data from Twitter's API JSON is extracted and flattened such that each tweet is represented as a single "entry" element containing the full text of the tweet, the metadata associated with that tweet, the metadata associated with the author of the tweet, and any researcher-provided metadata—all stored in simple name/value pairs, which take the form name="value". For example, Figure 2.19 shows the result of converting the tweet presented in Figure 2.18 above from Twitter's JSON to TWIG's XML.

```
<entry tweet_retweet_count="2" tweet_reply_count="0"
tweet_like_count="0" tweet_quote_count="0"
context_annotations="Brand Vertical:Retail,Brand
Category:Online,Brand:Etsy" tweet_text="RT @RubyPerry11:
https://t.co/QOX5ozZ7PP FREE SHIPPING #lamps #lights #lighting
#nightlights #freeshipping #etsy #gifts #giftideas #Retro #..."
referenced_tweets_1_type="retweeted"
referenced_tweets_1_id="1432832663182774280"
tweet_id="1432849952846274563"
tweet_conversation_id="1432849952846274563"
tweet_author_id="3154858995" tweet_source="Twitter Web App"
tweet_created_at="2021-08-31T23:37:04.000Z" tweet_lang="en"
user_description="https://t.co/fA52BpKRpV...\n#bowlcovers
#kitchendecor \nLove to make home decor and Jewelry find my shop
Crystalscraftycorner @ Etsy and also @ Amazon Handmade"
user_id="3154858995" user_username="ckwlpn" user_name="Crystal
Willis" anno_gender="cis_F"/>
```

Figure 2.19 – TWIG’s XML tweet object structure

This tweet was chosen as an example because it is a retweet, which provides a useful illustration of how TWIG flattens hierarchical data from the Twitter JSON. Where in the JSON, the referenced_tweets name was attached to a list with type and id child elements, in TWIG XML, the names of the child elements were appended to the referenced_tweets name to create two distinct name/value pairs at the same hierarchical “ground level” as all of the other data about the tweet. These have been colored red for visibility. Because it is possible for tweets to reference multiple tweets, for example in the case of retweeting a quote tweet, each type and id is numbered.

This flattening of the data is central to TWIG’s analytical philosophy because it disassociates the analytical environment and codebase from the logic of Twitter’s data structure and allows all metadata to be treated equally and parsed by the same functions. What this means is that when a project is opened in TWIG’s analyze environment, there are no specific names of metadata it expects to find. Instead, TWIG inspects the XML, creates a list of the names in each of the name/value pairs, and uses those to construct the metadata column of the analysis environment. A simple illustration of the value of this is that any data structured to fit TWIG’S

XML format would be equally analyzable in TWIG, such that the fact that the data came from Twitter becomes itself a kind of metadata. In addition to being extremely flexible in terms of the kinds of Twitter research that can be done within TWIG, this creates significant room for future expansion that can include data sources other than Twitter. And because the data source is only another kind of metadata, this would allow the easy creation of cross-platform studies to, for example, compare language use and user behavior between Twitter, Instagram, and so on.

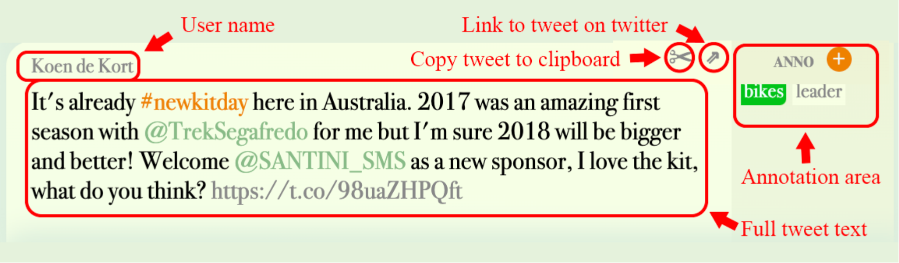
Data analysis tools

Despite the depth of the tools TWIG provides for collecting structured datasets of Twitter data, data collection was not the primary motivation for developing TWIG. As discussed in Chapter 2, TWIG was developed foremost as an analysis tool to address shortcomings of existing text-analysis tools with regards both to the inability of most discourse tools to meaningfully parse metadata and to the tendency of corpus tools to not provide the kinds of text-level information that is necessary for thoughtful discourse analysis. In this section I will provide a detailed overview of the aims and functions of the analysis environment and the various tools it provides.

As touched on in this chapter's overview of the basic TWIG workflow, the analysis environment is designed with three panes: a central column that displays full-text tweets from the current project's dataset flanked on either side by controls for on-the-fly filtering of this dataset into subsets based on descriptive statistics regarding, on the left, metadata associated with each tweet and, on the right, the texts of the tweets themselves. The column of full-text tweets is both centrally-positioned and larger than the statistical columns in order to emphasize that this environment is meant to facilitate analysis, not to complete it. This is to reiterate the point from the final section of Chapter I that TWIG is developed from a discourse-centric perspective that holds that the text itself is central, that meaning is a product of textual and contextual interactions, and that statistical representations of a dataset are thus valuable primarily not as findings, but as intermediary steps that precede continued analysis of the texts themselves. Put simply, TWIG's analysis environment uses a selection of standard and novel corpus tools to find patterns in the text so that it can put tweets in front of researchers in a structured way.

The tweets

The central tweets column shows the full text of tweets from the active dataset, divided into 100-tweet "pages." The active dataset refers to those tweets from the working project's base dataset that remain after the researcher has applied filters in the Text- or Metadata Analysis panes. These filters and their uses will be discussed in the following sections. Within the central tweets column, each tweet is presented in full along with the display name of the user who posted it, a button to copy the text of the tweet to the researcher's computer's clipboard, a link to the original tweet on Twitter, and a small panel on the right in which the researcher can apply a coding schema. A diagram is provided in Figure 2.20 below.



Koen de Kort

It's already #newkitday here in Australia. 2017 was an amazing first season with @TrekSegafredo for me but I'm sure 2018 will be bigger and better! Welcome @SANTINI_SMS as a new sponsor, I love the kit, what do you think? https://t.co/98uaZHPQft

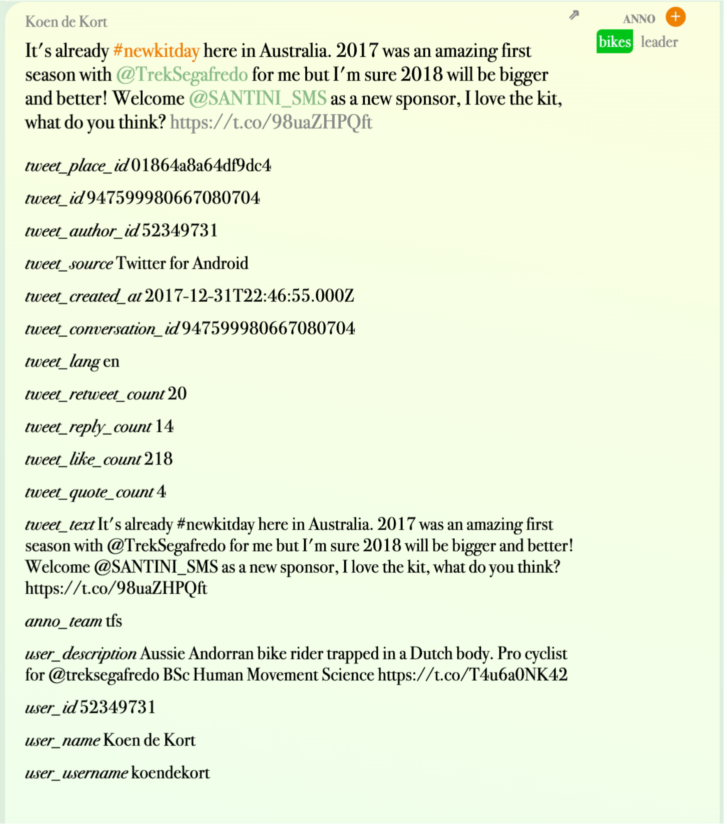
Figure 2.20 - An individual tweet in TWIG

The presentation of individual tweets in TWIG is kept minimal to maximize readability, preserve the centrality of the text, and facilitate efficient coding and analysis processes. Within each tweet's text, colors are used to highlight different interactive elements. Hashtags are made orange, and clicking them updates the active dataset to show only those tweets that contain the same hashtag. @mentions, colored green, function similarly, applying when clicked a filter to show only tweets containing the same @mention. URLs in tweets are colored grey and when clicked open the url in a new window.

To the top right of each tweet's text are two icons, a small pair of scissors and a diagonal arrow (⇗). The scissors icon, when clicked, copies the text of the tweet to the researcher's computer's clipboard. This feature is intended to simplify the collection of examples. When pasted, the tweet is presented in a format intended to further its usefulness as an example, which is to say it copies not only the tweet, but the user handle of the user who posted it, to make the tweet more findable by readers, and the time at which it was posted, as shown in Example 1 below.

1. NathanPeterHaas: In Saitama and have no idea where to ride! Any help from my Japanese cycling friends? We are at the Rafre hotel (2017-11-01T00:42:14.000Z)

The diagonal arrow icon at the top of each tweet is a link that opens the tweet on Twitter itself, in a new window. This can be valuable for tweets that contain media attachments as well as for viewing tweets that are clearly replies in the context of their full reply chains. Another way to view more information about a tweet, as well as the user who posted it, within TWIG is to double-click anywhere in the tweet. This toggles an extended view of the tweet, shown in Figure 2.21, which shows the tweet and all metadata associated with it.



Koen de Kort

It's already #newkitday here in Australia. 2017 was an amazing first season with @TrekSegafredo for me but I'm sure 2018 will be bigger and better! Welcome @SANTINI_SMS as a new sponsor, I love the kit, what do you think? https://t.co/98uaZHPQft

tweet_place_id 01864a8a64df9dc4

tweet_id 947599980667080704

tweet_author_id 52349731

tweet_source Twitter for Android

tweet_created_at 2017-12-31T22:46:55.000Z

tweet_conversation_id 947599980667080704

tweet_lang en

tweet_retweet_count 20

tweet_reply_count 14

tweet_like_count 218

tweet_quote_count 4

tweet_text It's already #newkitday here in Australia. 2017 was an amazing first season with @TrekSegafredo for me but I'm sure 2018 will be bigger and better! Welcome @SANTINI_SMS as a new sponsor, I love the kit, what do you think? https://t.co/98uaZHPQft

anno_team tfs

user_description Aussie Andorran bike rider trapped in a Dutch body. Pro cyclist for @treksegafredo BSc Human Movement Science https://t.co/T4u6a0NK42

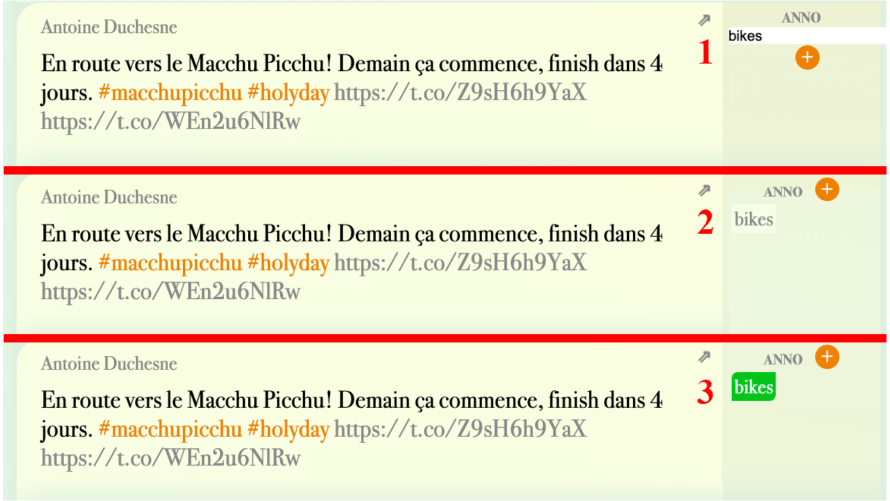
user_id 52349731

user_name Koen de Kort

user_username koendekort

Figure 2.21 - Extended view of an individual tweet

On the right of each tweet is the annotation/coding panel, in which researchers can create and apply coding schema. Annotation options defined in the Query Design environment will appear automatically here, but can also be added by clicking the orange "+" button, typing the name of a new annotation option, and pressing return, as shown in Figure 2.22. Clicking an annotation option toggles the coding of the relevant tweet, such that the tweet in Figure 2.20 is coded with the "bikes" tag but not "leader."



Antoine Duchesne

En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

Antoine Duchesne

En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

Antoine Duchesne

En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

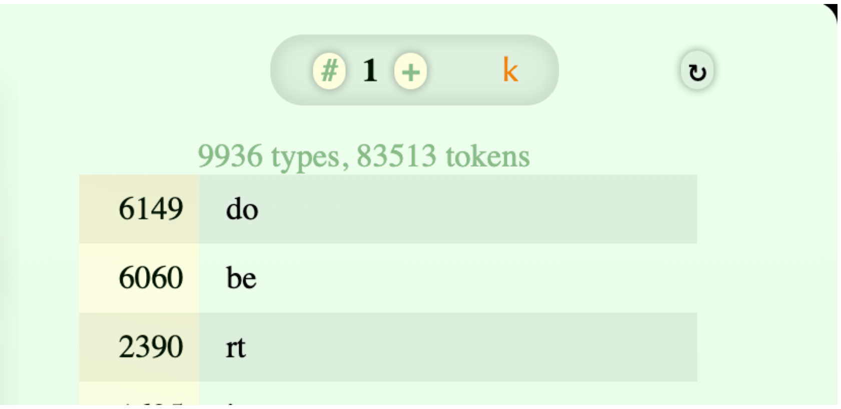
Figure 2.22 - Adding annotation options during analysis

The text analysis pane

The bulk of TWIG's standard corpus text-analysis tools are used via the pane on the right side of the analysis environment. These include tools for displaying frequent words and n-grams, collocates of selected words or n-grams, concordance lines, and ranked keywords. Also available are tools for displaying frequent hashtags and user @mentions, which are lexical items unique, and integral, to social-media discourse. All of the frequency tables TWIG provides serve also as potential filters for the active dataset, which is to say that clicking an n-gram, hashtag, or @mention will narrow the set of tweets that are shown and for which statistics are calculated to only those that include the selected lexical item.

These filters can be stacked, with the effect of displaying tweets that match any, not all, of the selected filters. When a new filter has been applied, clicking the refresh (↻) button on the top right of the Text Analysis Pane refreshes the distribution table to reflect the contents of the filtered dataset. This behavior, in which the tables do not refresh automatically, allows researchers to rapidly change between filters in the central Tweets column without affecting quantitative output.

Navigating between the various frequency tables TWIG can show is done using the small menu at the top of the text-analysis column. The default setting shows frequent two-word n-grams. Other tools are made available by clicking the plus or minus buttons at the top. The plus button increases the n-gram size, up to six words. The minus button reduces the n-gram size. At the single-word level, however, other tools become available. As shown in Figure 2.23, the minus button is replaced with a hashtag, which replaces the word frequency view with hashtag and @mention frequency, and there is a new “k” button to the right, which enables sorting words from the dataset not by frequency but by keyness.

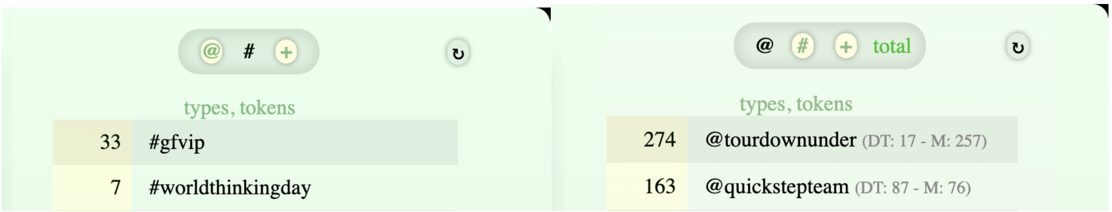


9936 types, 83513 tokens

6149	do
6060	be
2390	rt

Figure 2.23 - Single-word view of text-analysis column}

The hashtag and @mention frequency views are the only social-media-discourse-specific tools in the text-analysis column, and their functionality is relatively simple. As shown in Figure 2.24 the @mention tool is slightly more complex, showing counts for both directed tweets and in-text mentions. These values can be particularly valuable in studies that investigate discourse interaction either within or between communities.



33	#gfvip
7	#worldthinkingday

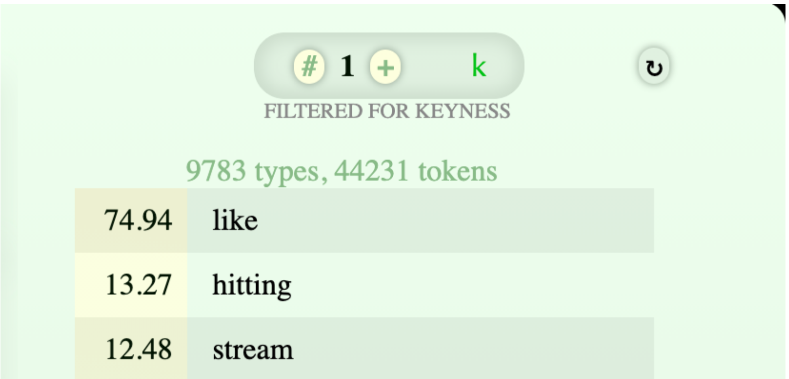
274	@tourdownunder (DT: 17 - M: 257)
163	@quickstepteam (DT: 87 - M: 76)

Figure 2.24 - Hashtag (left) and @mention (right) frequencies

Keyness

As discussed in the previous chapter of this document, corpus methods are often criticized for over-valuing raw frequency data. One option that TWIG offers to compensate for this is the ability to show frequent words ranked not by frequency but by keyness. Keyness values in TWIG represent a simple comparison between the frequency of a word in the active dataset and the frequency of the same word in a reference corpus, normalized to show relative words per million rather than any kind of statistical significance. For example, keyness value of 10 would indicate that the target word occurs 10 times more in the working dataset of tweets than in the reference corpus, and a keyness value of 0.5 would indicate a word that occurs half as frequently in the tweets as in the reference corpus. TWIG's default reference corpus is the British National Corpus, which was used primarily because a word frequency table has been made readily, and freely, available by Leech et. al (2001).

It must be acknowledged, however, that this is not an ideal reference corpus for analysis of online language use, considering that much of its content pre-dates the internet. For example, in Figure 2.25 below, *stream* has a high keyness value, but this likely has more to do with the contemporary prominence of streaming media--which would not have existed at the date of the frequency table's compilation--than with language unique to this dataset. That said, I believe the BNC frequency table still works sufficiently for most exploratory analysis. As TWIG develops, the reliability of keyness results will be expanded by the addition of newer, more appropriate reference corpora.



FILTERED FOR KEYNESS

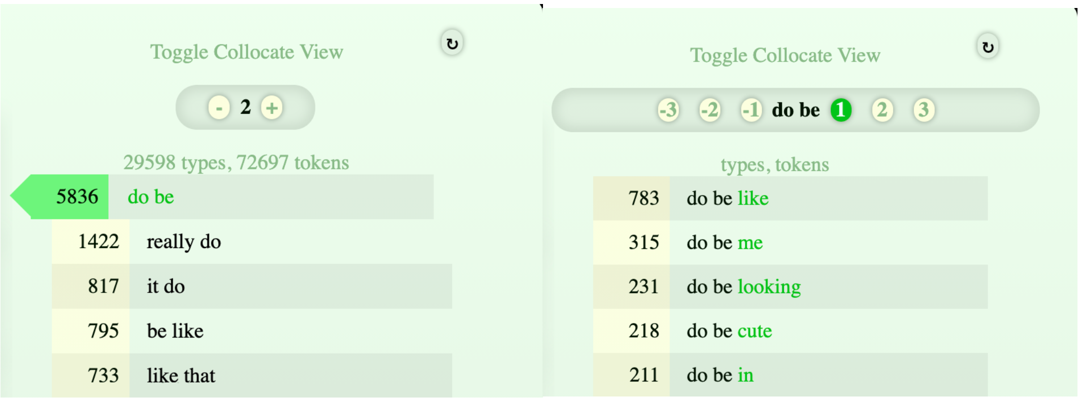
9783 types, 44231 tokens

74.94	like
13.27	hitting
12.48	stream

Figure 2.25 - Single-word view with keyness enabled

Collocation and concordance

When only one lexical filter is applied, two additional tools become available in TWIG: a collocate frequency table and a concordance view of the active dataset. To access the collocate frequency table, the Text Analysis Pane must be refreshed. This enables a "Toggle Collocate View" option, which, when enabled, displays a frequency table of collocates of the active lexical filter. As shown in Figure 2.26, when the collocate view is enabled, the menu at the top of the Text Analysis Pane is replaced with a menu for setting the side and size of collocates to explore. Contents of collocate tables can also be applied as filters to the active dataset, and these filters act separately from other lexical features. For example, considering Figure 2.26, if "do be like" were applied as a filter and subsequently removed, the initial "do be" filter from which the collocate was identified would remain.



29598 types, 72697 tokens

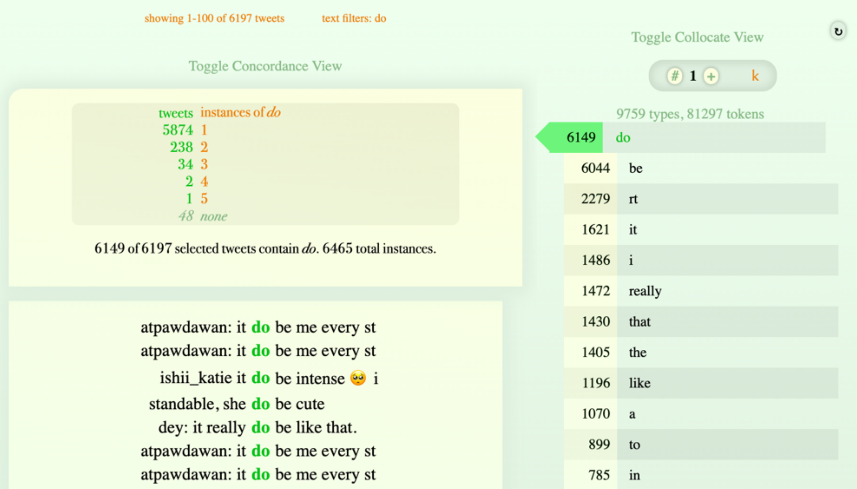
5836	do be
1422	really do
817	it do
795	be like
733	like that

types, tokens

783	do be like
315	do be me
231	do be looking
218	do be cute
211	do be in

Figure 2.26 - Collocate view disabled (left) and enabled (right)

Another tool that becomes accessible when only a single lexical, or collocate, filter is applied is a concordance view of that construction. The concordance view, as shown in Figure 2.27 replaces the standard central full-tweet column with scannable concordance lines to facilitate analysis of the target construction's textual environment. The top of this view also shows additional information on the distribution of the target construction within the corpus in the form of a table that displays intra-text frequencies—i.e. how many tweets contain multiple instances of the target construction.



tweets	instances of *do*
5874	1
238	2
34	3
2	4
1	5
48	*none*

6149 of 6197 selected tweets contain *do*. 6465 total instances.

```
atpawdawan: it do be me every st
atpawdawan: it do be me every st
  ishii_katie it do be intense 🥺 i
 standable, she do be cute
  dey: it really do be like that.
atpawdawan: it do be me every st
atpawdawan: it do be me every st
```

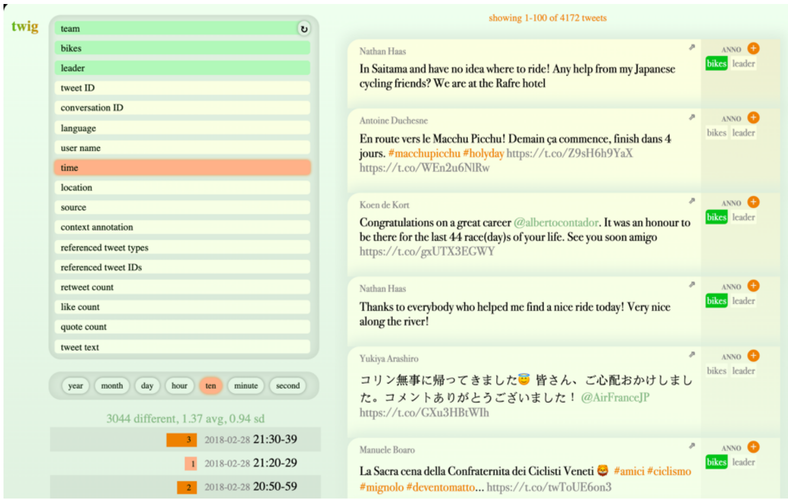
9759 types, 81297 tokens

6149	do
6044	be
2279	rt
1621	it
1486	i
1472	really
1430	that
1405	the
1196	like
1070	a
899	to
785	in

Figure 2.27 - TWIG's concordance view

The metadata analysis pane

What are likely TWIG's most novel capabilities are accessible within the Metadata Analysis Pane, on the left. This pane provides dynamic descriptive statistics regarding the metadata of each tweet in the active dataset, such that frequencies can be found that show tweet distributions by user, time, researcher-defined annotation group, and so on. These are navigated between using a dropdown menu above the pane's distribution table, as shown in Figure 2.28. It should be noted again that only metadata enabled in the Query Design Environment, which can be updated at any time, are available in this menu. Furthermore, to streamline the analysis process, researcher-defined annotation categories are displayed in green in the dropdown and are also made accessible in an abbreviated menu, which does not require a dropdown action to use, as shown in Figure 2.29.



3044 different, 1.37 avg, 0.94 sd

3	2018-02-28 21:30-39
1	2018-02-28 21:20-29
2	2018-02-28 20:50-59

Nathan Haas
In Saitama and have no idea where to ride! Any help from my Japanese cycling friends? We are at the Rafre hotel

Antoine Duchesne
En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

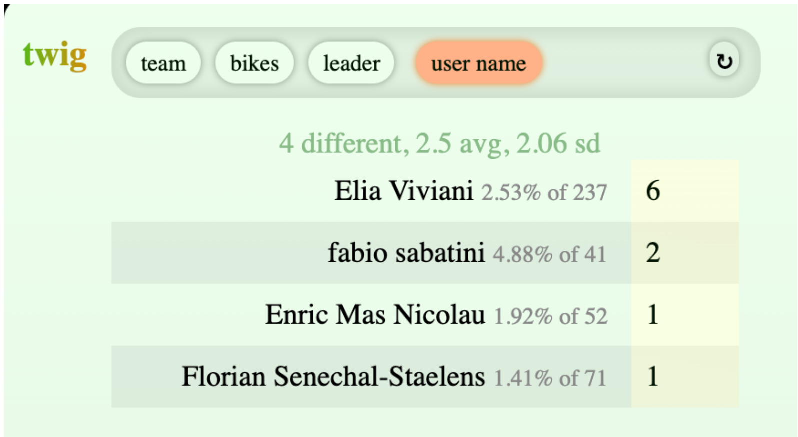
Koen de Kort
Congratulations on a great career @albertocontador. It was an honour to be there for the last 44 race(day)s of your life. See you soon amigo https://t.co/gxUTX3EGWY

Nathan Haas
Thanks to everybody who helped me find a nice ride today! Very nice along the river!

Yukiya Arashiro
コリン無事に帰ってきました😇 皆さん、ご心配おかけしました。コメントありがとうございました！ @AirFranceJP https://t.co/GXu3HBtWIh

Manuele Boaro
La Sacra cena della Confraternita dei Ciclisti Veneti 🤩 #amici #ciclismo #mignolo #deventomatto... https://t.co/twToUE6on3

Figure 2.28 - The metadata selection menu

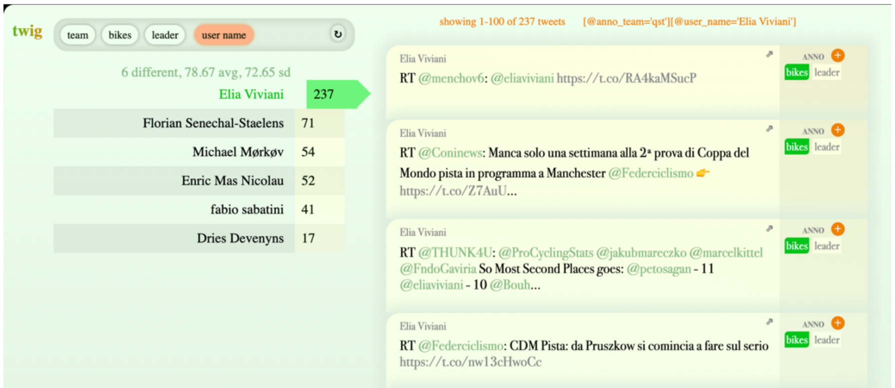


4 different, 2.5 avg, 2.06 sd

Elia Viviani 2.53% of 237	6
fabio sabatini 4.88% of 41	2
Enric Mas Nicolau 1.92% of 52	1
Florian Senechal-Staelens 1.41% of 71	1

Figure 2.29 - The abbreviated metadata selection menu

As with the Text Analysis Pane, the distribution tables provided in this pane also act as navigation menus that can be used to add filters to the active dataset, and these filters can be stacked to show, for example, the distribution of tweets per user within a particular week. When these filters are combined with lexical filters, the active dataset is narrowed to include and calculate statistics using only tweets that match both criteria. This can be applied to quickly find complex measures, such as, for example, the distribution of tweets per user within a particular week considering only those tweets that include a particular word, n-gram, hashtag, or @mention. Furthermore—because, again, each item on a frequency table in TWIG is also a potential filter of the active dataset—it would be very simple, given the current example, to then read and code the relevant tweets for each user. This kind of activity is representative of the kinds of structured exploratory analysis TWIG is designed to facilitate. The combinations are limited only by the available metadata and can be used to approach the data from a variety of theoretical directions, as will be demonstrated in the final two chapters of this document. Figure 2.30 below shows a distribution of tweets posted by riders in a single bicycle racing team, taken from a corpus of tweets from twenty-two teams, in which the tweets of a single rider, Elia Viviani are being shown in the tweets column.



6 different, 78.67 avg, 72.65 sd

Elia Viviani	237
Florian Senechal-Staelens	71
Michael Mørkøv	54
Enric Mas Nicolau	52
fabio sabatini	41
Dries Devenyns	17

showing 1-100 of 237 tweets [@anno_team='qst'][@user_name='Elia Viviani']

Elia Viviani
RT @menchov6: @eliaviviani https://t.co/RA4kaMSucP

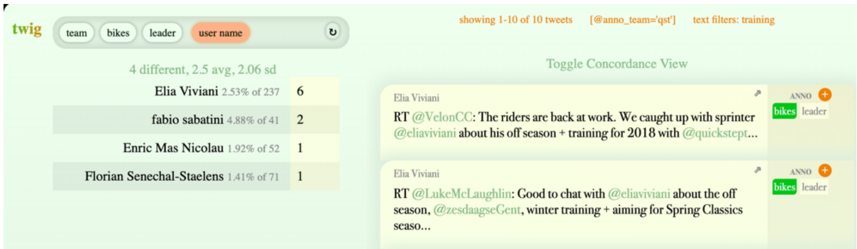
Elia Viviani
RT @Coninews: Manca solo una settimana alla 2ª prova di Coppa del Mondo pista in programma a Manchester @Federciclismo 👉 https://t.co/Z7AuU...

Elia Viviani
RT @THUNK4U: @ProCyclingStats @jakubmareczko @marcelkittel @FndoGaviria So Most Second Places goes: @petosagan - 11 @eliaviviani - 10 @Bouh...

Elia Viviani
RT @Federciclismo: CDM Pista: da Pruszkow si comincia a fare sul serio https://t.co/nw13cHwoCc

Figure 2.30 - Distribution of grouped user tweets

In addition to facilitating structured data exploration, TWIG's Metadata Analysis Pane is designed to address criticisms of existing corpus methods and software, particularly those criticisms regarding uncertainty estimates and dispersion. This is achieved in several ways. Most basically, TWIG's framework of exploratory metadata, as described above, allows dispersion for a target feature, be it lexical or contextual, to be measured across multiple kinds of grouping—such as author, time, and tweet type. The effect of this is similar to the approach of using bootstrapping with "linguistically meaningful sampling unit[s]" (Gries, 2022, p. 8). To bolster the meaningfulness of the distribution shown by the Metadata Statistics pane, at the top of each table is a summary showing the number of matching categories (e.g. days on which tweets matching the filter criteria appear), the average number of matching tweets per category, and the standard deviation. This gives a sense of both dispersion and variability. Furthermore, when a lexical filter is applied to the active dataset, TWIG provides local measures for each item in the metadata distribution table that compare frequency in the active dataset to that of the base dataset. This is illustrated in Figure 2.31 and Figure 2.32 below.



showing 1-10 of 10 tweets [@anno_team='qst'] text filters: training

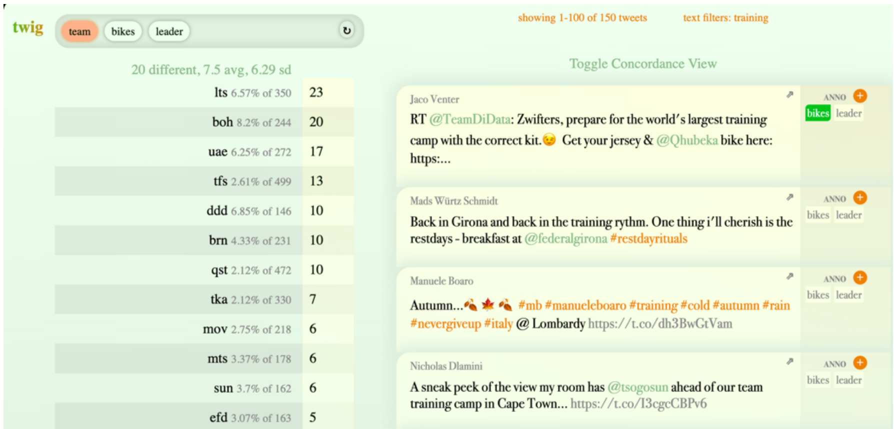
4 different, 2.5 avg, 2.06 sd

Elia Viviani 2.53% of 237	6
fabio sabatini 4.88% of 41	2
Enric Mas Nicolau 1.92% of 52	1
Florian Senechal-Staelens 1.41% of 71	1

Elia Viviani
RT @VelonCC: The riders are back at work. We caught up with sprinter @eliaviviani about his off season + training for 2018 with @quickstept...

Elia Viviani
RT @LukeMcLaughlin: Good to chat with @eliaviviani about the off season, @zesdaagseGent, winter training + aiming for Spring Classics seaso...

Figure 2.31 - Local active user statistics relative to base



showing 1-100 of 150 tweets text filters: training

20 different, 7.5 avg, 6.29 sd

lts 6.57% of 350	23
boh 8.2% of 244	20
uae 6.25% of 272	17
tfs 2.61% of 499	13
ddd 6.85% of 146	10
brn 4.33% of 231	10
qst 2.12% of 472	10
tka 2.12% of 330	7
mov 2.75% of 218	6
mts 3.37% of 178	6
sun 3.7% of 162	6
efd 3.07% of 163	5

Jaco Venter
RT @TeamDiData: Zwifters, prepare for the world's largest training camp with the correct kit.😉 Get your jersey & @Qhubeka bike here: https:...

Mads Würtz Schmidt
Back in Girona and back in the training rythm. One thing i'll cherish is the restdays - breakfast at @federalgirona #restdayrituals

Manuele Boaro
Autumn...🍂 🍁 🍂 #mb #manueleboaro #training #cold #autumn #rain #nevergiveup #italy @ Lombardy https://t.co/dh3BwGtVam

Nicholas Dlamini
A sneak peek of the view my room has @tsogosun ahead of our team training camp in Cape Town... https://t.co/I3cgcCBPv6

Figure 2.32 - Local active group statistics relative to base

Figure 2.31 shows the frequency with which riders for a particular bicycle racing team post tweets containing the word "training." The local measures here show that Elia Viviani has posted six tweets containing "training," which represent 2.53 percent of the 237 Viviani tweets in the full dataset. In Figure 2.32, these measures are shown for each team in the corpus as a whole.

Because one of the stated goals of TWIG is to make complex, exploratory analysis of corpus data accessible to researchers without programming experience or extensive coding skills, it is worth noting at this point, as an illustration of the project's success in that regard, that to reach the analytical points shown in Figure 2.31 and Figure 2.32 required seven and four clicks, respectively, starting from the default settings of TWIG's Analysis Environment. First the n-gram size in the Text Analysis Pane was set to one. Then the keyness view was selected, which showed "training" to be possibly meaningful direction for explanation. After applying "training" as a lexical filter, "team" was selected in the Metadata Analysis Pane, producing Figure 2.32. From this point, Vivianni's team, represented as "qst," was applied to the dataset as a filter, and the Metadata Analysis Pane was set to show statistics by "user name," producing Figure 2.31.

In order to further demonstrate TWIG's functionality and its viability for facilitating multiple kinds of discourse analysis using twitter data, the final two chapters of this document present sample studies conducted within TWIG. The first study concerns the identity and value constructing discourse of a defined group of users, while the second considers diachronic change in use of a particular construction. Both studies make use of temporal query design and metadata as well as descriptive statistics regarding such things as hashtag and n-gram frequency, but otherwise vary greatly in the technical requirements of data collection and analysis and in the theoretical concerns behind the analysis.

CHAPTER III

THE #BOYS OF THE @PELOTON: CONSTRUCTING PROFESSIONAL CYCLISTS

Twitter has become a productive site for the linguistic analysis of identity and value construction (Iveson, 2017; Marwick & Boyd, 2011; Page, 2012; Zappavigna, 2011, 2014, 2018). The current study intends to demonstrate the viability TWIG as a platform for conducting such research. TWIG promises to remove technical barriers to collecting and analyzing Twitter data, allowing for faster and more robust study completion in this rapidly-growing field. To highlight these advantages, this study is a re-creation of a study I previously conducted regarding identity and value construction among professional bicycle racers on Twitter. Re-creating an earlier study allows particular attention to be given to describing the methods of data collection and analysis in TWIG in contrast to what would be involved in performing the same work "by hand." Furthermore, this allows for practical demonstration of TWIG's features for working with pre-existing datasets.

Previous studies of identity and value construction on Twitter have focused their analysis on either broadly-conceived groups of users, such as celebrities (Marwick & boyd, 2011; Page, 2012), or ad-hoc communities that emerge online around specific topics, such as people dealing with depression (Zappavigna, 2018) and Catalonians reacting to a specific political promotional video (Iveson, 2017). This study aims to extend this direction of inquiry by considering online identity and value construction as performed by a community that exists primarily offline.

The community that will be considered is specifically that of professional bicycle racers. This community, unlike ad-hoc communities that exist only on Twitter itself, has both a clear-cut membership and an inherent audience of fans, media, and peers, all of whom can be assumed to have at least a degree of knowledge of the riders, teams, races, and other assorted entities of the sport. This creates a situation in which identity construction for the racers, like with other public figures, is likely fundamentally different from that of "ordinary" users, who must start from a blank slate. Furthermore, certain values and in-group power dynamics of bicycle racing have been identified in sports sociological research (Jutel, 2002; Williams, 1989) that enable a more focused analysis of how those values manifest on Twitter than is possible with such broad categorizations of users as "celebrities," which includes such disparate sub-groups as actors, musicians, athletes, and politicians, all of whom likely have different motivations for using Twitter and different anticipated audiences for their tweets—and as such are likely to construct identity in different ways. Furthermore, bicycle racing exists at a sort of sweet spot in terms of popularity. The racers, being public figures, are almost all on Twitter, but most of them are not high-profile enough to have other people tweeting on their behalf. This results in an abundance of tweets in which identity-constructing posts can be assumed to originate with the racers themselves.

I begin below with a brief discussion of identity construction on Twitter. This is followed by a discussion of the bicycle racing community and the values thereof, specifically regarding aspects of the racing itself and the culture surrounding it that might influence how the racers tweet. Finally, I provide a detailed description of the methods used to collect and analyze the data for the current study and present what I have found regarding the values and identity professional cyclists construct on Twitter and the mechanisms by which they do so.

Identity Construction on Twitter

Zappavigna (2014) gives several examples of identity-constructive work in the use of hashtags to create what she calls "quotidian bonds," of which she describes three kinds in detail: self-deprecation, addiction, and frazzle. These correspond to examples 1, 2, and 3 below.

1. oof. Vi passing out sick on the couch watchign pbs, and I just realized I didn't change her diaper yet this morning #badmama (from Zappavigna 2014, p. 217)
2. I hate waking up at 6am everyday. Ughh #needcoffee (p. 221)
3. I have been installing apps, creating app IDs and profiles for a client all day. Good thing I had the day off! #tired @iphonedeveloper (p. 222)

Zappavigna gives very detailed accounts of the nature of each of these bonds, but because the bonds that occur among professional cyclists are likely different, what's perhaps most currently salient is that bonding can occur in several ways—around explicit, if tongue-in-cheek, identity adoption (#badmama), around "iconization" of objects that carry particular value associations (#needcoffee), and simply around relatable feelings (#tired).

Interestingly, these tweets also show that while hashtags may have been conceived as searchable topic indicators, their uses have grown beyond that. Scott (2015) discusses this in more detail, noting several ways in which tweet-final hashtags can provide information that is fundamental to interpretation of the tweets they follow. Scott isn't explicitly interested in questions of identity, but much of what she describes can have identity-constructive functions. For example, in 4 below, the hashtag #mcfc (Manchester City Football Club) provides the recipient for the tweeter's positive vibes in what would otherwise be an uninterpretable tweet, and this key to interpretation also not only indexes the tweeter as a fan of the club, but emphasizes that indexing by isolating it.

4. Sending positive vibes. Positive vibes. Positive vibes. #mcfc (Scott 2015, p. 16)
5. One week from today I can start throwing again. #finally (Scott 2015, p. 17)

Example 5 is especially interesting, because the hashtag *#finally* seems to serve entirely to guide interpretation—saying something about how the tweeter feels about what's said in the body of the tweet, suggesting an identity-constructive function similar to Zappavigna's "frazzle" bond. What *#finally* does not do is index a broader discourse, considering that a search of tweets containing that hashtag would return incredibly disparate results. This shows that not all of the bonds constructed in hashtag use are about findability or broader discourses. Such hashtags seem to be not to be intended for Twitter's algorithms and a broad topical audience, but instead specifically for the posting user's followers, suggesting that different kinds of hashtag use might in effect construct identity differently for different audiences.

While the above describes only differences in function and use that exist in hashtags, such differences also exist in @mentions and retweets. Specific study of these, however, is limited. Boyd et al. (2010) analyze general retweeting practices, finding that they can be used for such purposes as finding new audiences for a tweet, informing a specific audience, endorsing a tweet's content, or signaling friendship or loyalty (p. 6). One of the more salient kinds of retweet they describe to the current analysis of professional athletes is the "ego retweet," which is when a user retweets another user's mention of them.

Unlike hashtags and retweets, @mentions don't seem to have been the subject of much specific analysis. Hemsley et al. (2017) considers @mentions in the context of politicians' Twitter activity, quantifying their use in Tweets with functions such as calling for action, attacking opponents, and establishing issue positions. These functions are not especially useful to the analysis of how professional cyclists use Twitter, but this fact is itself a useful illustration of the need for linguistic analyses of social-media to consider specific communities of users rather than general populations.

The values of bicycle racing

The values of bicycle racing are largely shaped by the fact that it is a team sport with individual winners. The sport glorifies individual riders and performances, but the races exist at a scale that makes cooperation among riders necessary. Riders race in groups on the road, because, for reasons related to wind-resistance, a group will always go faster than an individual doing the same amount of work (Williams, 1989). The most fundamental groups in bicycle racing are teams. Teams are not only the racers' employers—and so sources of professional identity—they are also the entities around which individual racers orient their racing. Because bicycle racing has individual winners, each team works as a whole for the success of one or two of their riders, who for the purpose of this study will be referred to as "leaders." The other racers on a team ride in a way intended to facilitate the leaders' success.

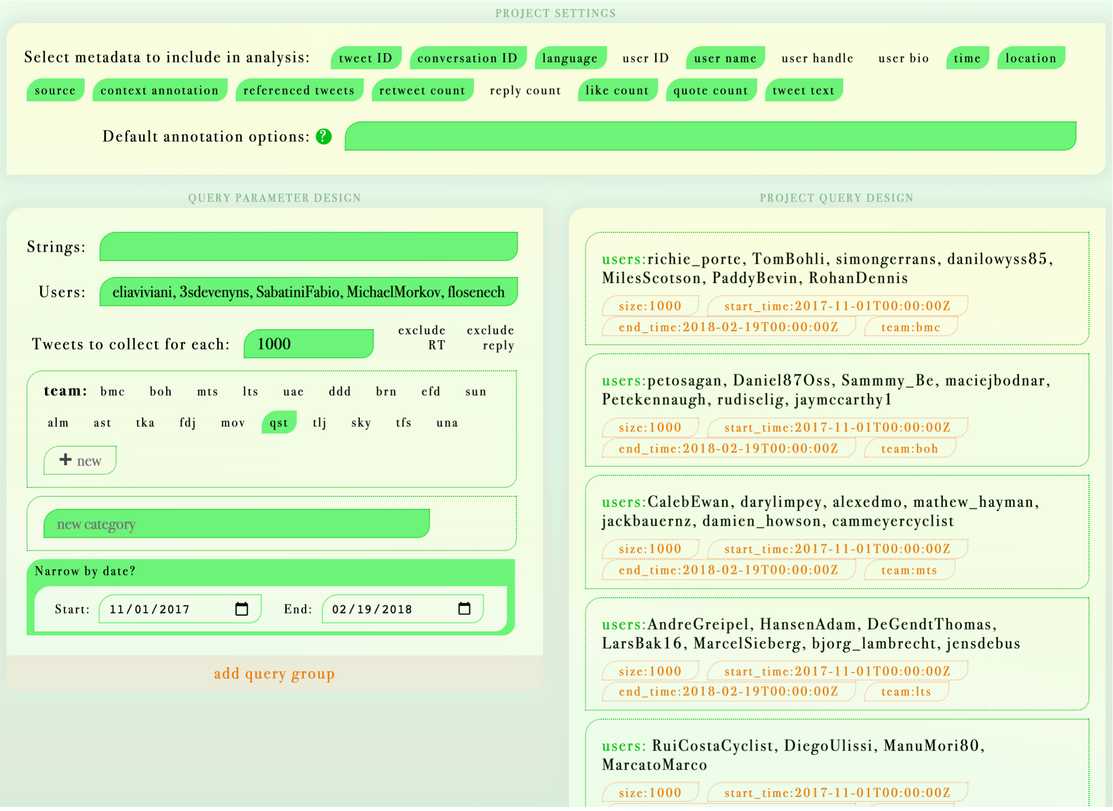
Bicycle racing is unique, however, in that it requires cooperation not just within teams, but between them. Races typically split into a number of mixed-team groups on the road: a main, large group called the *peloton* with one or more smaller groups ahead or behind. These groups rely on the cooperation of riders from multiple teams. One group in which this is especially the case is the *gruppetto,* which emerges at the very back of particularly mountainous stages in multi-day races. The gruppetto is where various teams' sprinters, who are physiologically better-suited to flatter courses, risk disqualification in the mountains if they don't work together to finish before a cutoff time based on that of the first finisher. This overview of bicycle racing is a simplification, but it gives a sense of the conception of cooperation that is among the core values of the sport, and it previews the kinds of identity and values that are likely to emerge in the racers 'tweets.

Data Collection Methods

This study involves a corpus of tweets posted by cyclists representing the 19 teams who competed in the 2018 Tour Down Under (TDU), which was the first major stage race of that year's road racing season, taking place between January 16 and 21 in Adelaide, Australia. The corpus contains all available tweets posted by these riders between the first of November, 2017 and the 28th of February, 2018. This is to say that, of the 130 riders who participated in the race, 121 had Twitter accounts, and 110 of them posted in the target timeframe. This time period allows for both an analyses of how racers construct identity in the off season when they are not racing, and an analysis of patterns of activity surrounding a single race (i.e. discourse regarding preparation, racing in progress, and reflection). For this study, each tweet is also annotated to indicate the team of the rider who posted it. This enables close analysis of how individual racers construct identity within the construct of their teams.

Collecting this corpus "by hand" involved the development of a custom Python script to retrieve tweets from each of the cyclists, within the stated temporal range, to then add a team annotation to each collected tweet "object" (as defined in Chapter II), and to finally convert the collected data to a format suitable for analysis. Using the Tweepy library to simplify the process of connecting to the Twitter API, the completed collection script includes 138 lines of code and collects limited metadata.

No coding was required to re-create this corpus within TWIG. Because this study is based on a previous study, there are actually several approaches to collecting the desired data. To analyze the existing corpus within TWIG's analysis environment, the simplest would have been to simply import the initial study's corpus. However, because the previous corpus included only a fraction of the metadata TWIG collects, it was beneficial to re-collect the data from Twitter. To do this, there are two possible approaches within TWIG's query design module. The first is to create a set of parameters that represent the initial study design. This can be achieved by adding a "team" category in the Query Parameter Design pane, adding options to that category representing each team, and designing subqueries for each team, each of which including usernames for each rider in the team and temporal boundaries set based on the start and end dates of the original corpus. Such a query design is shown below in Figure 3.1.



Select metadata to include in analysis: tweet ID, conversation ID, language, user ID, user name, user handle, user bio, time, location, source, context annotation, referenced tweets, retweet count, reply count, like count, quote count, tweet text

Default annotation options: ?

Strings:

Users: eliaviviani, 3sdevenyns, SabatiniFabio, MichaelMorkov, flosenech

Tweets to collect for each: 1000 exclude RT exclude reply

team: bmc boh mts lts uae ddd brn efd sun alm ast tka fdj mov qst tlj sky tfs una

Narrow by date?

Start: 11/01/2017 End: 02/19/2018

users:richie_porte, TomBohli, simongerrans, danilowyss85, MilesScotson, PaddyBevin, RohanDennis

size:1000 start_time:2017-11-01T00:00:00Z end_time:2018-02-19T00:00:00Z team:bmc

users:petosagan, Daniel87Oss, Sammmy_Be, maciejbodnar, Petekennaugh, rudiselig, jaymccarthy1

size:1000 start_time:2017-11-01T00:00:00Z end_time:2018-02-19T00:00:00Z team:boh

users:CalebEwan, darylimpey, alexedmo, mathew_hayman, jackbauernz, damien_howson, cammeyercyclist

size:1000 start_time:2017-11-01T00:00:00Z end_time:2018-02-19T00:00:00Z team:mts

users:AndreGreipel, HansenAdam, DeGendtThomas, LarsBak16, MarcelSieberg, bjorg_lambrecht, jensdebus

size:1000 start_time:2017-11-01T00:00:00Z end_time:2018-02-19T00:00:00Z team:lts

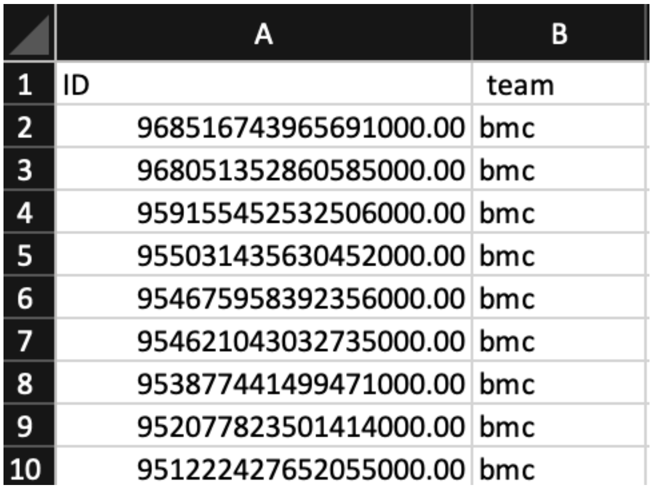
users: RuiCostaCyclist, DiegoUlissi, ManuMori80, MarcatoMarco

size:1000 start_time:2017-11-01T00:00:00Z

Figure 3.1 - Project design for collection of tweets from grouped users

This study clearly requires a rather complex query design that takes a lot of time to configure. For such cases, TWIG's Collect From IDs feature offers a less time-consuming approach to re-collecting an existing dataset. This method requires a degree of work outside of

TWIG that should be comfortable for most researchers who have existing datasets they want to analyze in TWIG. To use this feature, a list of tweet IDs must be created from the original dataset and uploaded as either a plain text document (.txt) of IDs separated by commas or as a CSV with column headers indicating any desired metadata to be added to each Tweet. The first few rows of the CSV file for collecting tweets along with team affiliation are shown below in Figure 3.2.

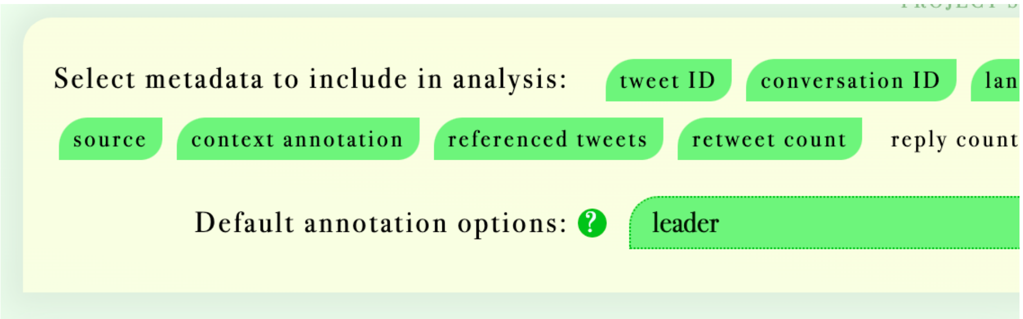


	A	B
1	ID	team
2	968516743965691000.00	bmc
3	968051352860585000.00	bmc
4	959155452532506000.00	bmc
5	955031435630452000.00	bmc
6	954675958392356000.00	bmc
7	954621043032735000.00	bmc
8	953877441499471000.00	bmc
9	952077823501414000.00	bmc
10	951222427652055000.00	bmc

Figure 3.2 - CSV format for collecting tweets from ID with metadata

This generates the same corpus that would be collected by the query design shown in Figure 3.1. Upon completion of the collection process, initiated by clicking “collect from Twitter” in the Project Query Design pane, the corpus becomes viewable in TWIG’s analyses environment. It is worth noting that, upon completion of this collection procedure, the TWIG-generated corpus lost four riders and about two hundred Tweets worth of data relative to the corpus collected initially in 2018. This is a result of time, not technological shortcoming. Tweets that have been deleted from Twitter, even if the Tweet IDs are specifically requested, cannot be retrieved. I believe that, for the purposes of the current study, the data gained by TWIG’s more robust collection procedures makes up for this roughly four percent loss of data and is particularly valuable for the purpose of demonstrating the use of TWIG as an analytical tool.

After the data collection, an annotation category was added post-hoc to differentiate between team leaders and supporting riders. Because racers pursuing results vary between races and are not broadly documented, this study identifies leaders as riders who either won individual stages of the TDU or who finished in the top ten in the final general classification.[1] These riders were assigned the code "leader." The "leader" coding group was initialized by adding "leader" as a "default annotation option" in TWIG's Project Design Environment, as shown in Figure 3.3.



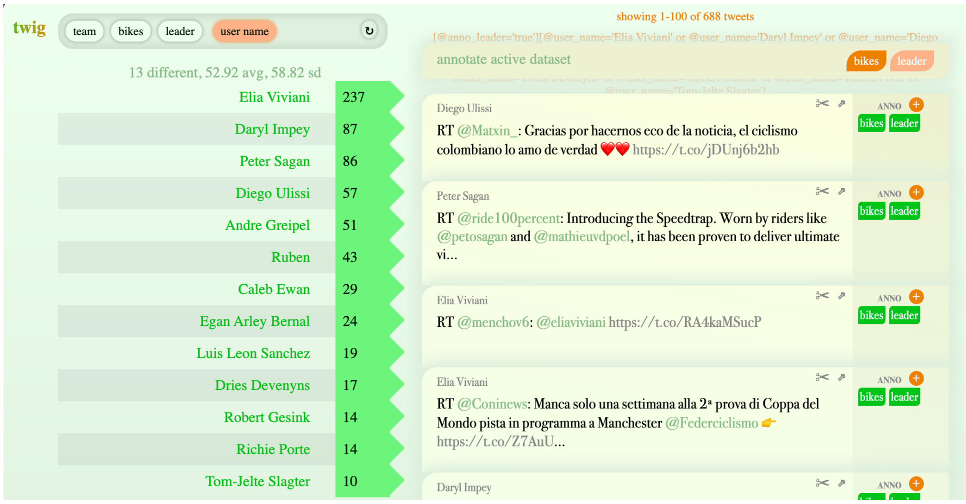
Select metadata to include in analysis: tweet ID conversation ID lan source context annotation referenced tweets retweet count reply count

Default annotation options: ? leader

Figure 3.3 - Addition of "leader" annotation option.

These racers' tweets were then coded en-masse by first selecting the racers individually in the Metadata Analysis pane, resulting in an active dataset containing only their tweets, and then selecting "leader" in the "annotate active dataset" menu at the top of the tweets column, as shown in Figure 3.4 below. This enables broad comparison of the tweets of both team leaders and supporting racers.

[1] Those riders are Daryl Impey, Richie Porte, Tom-Jelte Slaghter, Diego Ulissi, Dries Devenyns, Egan Bernal, Louis Leon Sanchez, Ruben Guerreiro, Robert Gesink, Andre Greipel, Caleb Ewan, Elia Viviani, Peter Sagan, and Gorka Izagirre. Of these, only Izagirre did not use Twitter during this study's collection period.



showing 1-100 of 688 tweets

[@anno_leader='true'][@user_name='Elia Viviani' or @user_name='Daryl Impey' or @user_name='Diego

@user_name='Dries Devenyns' or @user_name='Robert Gesink' or @user_name='Richie Porte' or @user_name='Tom-Jelte Slagter']

13 different, 52.92 avg, 58.82 sd

Elia Viviani	237
Daryl Impey	87
Peter Sagan	86
Diego Ulissi	57
Andre Greipel	51
Ruben	43
Caleb Ewan	29
Egan Arley Bernal	24
Luis Leon Sanchez	19
Dries Devenyns	17
Robert Gesink	14
Richie Porte	14
Tom-Jelte Slagter	10

Diego Ulissi
RT @Matxin_: Gracias por hacernos eco de la noticia, el ciclismo colombiano lo amo de verdad ❤️❤️ https://t.co/jDUnj6b2hb
bikes leader

Peter Sagan
RT @ride100percent: Introducing the Speedtrap. Worn by riders like @petosagan and @mathieuvdpoel, it has been proven to deliver ultimate vi...
bikes leader

Elia Viviani
RT @menchov6: @eliaviviani https://t.co/RA4kaMSucP
bikes leader

Elia Viviani
RT @Coninews: Manca solo una settimana alla 2ª prova di Coppa del Mondo pista in programma a Manchester @Federciclismo 👉 https://t.co/Z7AuU...
bikes leader

Daryl Impey

Figure 3.4 - Defining the “leader” group

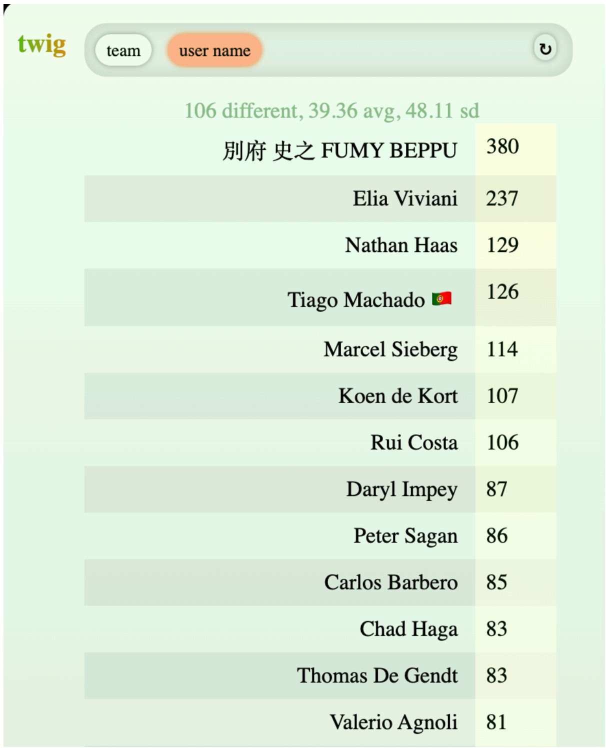
Data Analysis Methods

The analysis of this corpus involves a quantitatively-driven qualitative process using the tools of TWIG’s analysis environment. TWIG’s analysis environment provides quantitative descriptions of frequency patterns within both the text and the metadata of the collected Tweets. The contextual information provided by TWIG’s framework for handling metadata, both that metadata provided by the Twitter API and that added via researcher annotation, was central to this analytical procedure. Salient metadata and the associated statistics are used and combined to explore the corpus, which is to say that the statistics are used to narrow the corpus to show only tweets that fulfill certain criteria. For example, if it is found that a large number of tweets were posted on a particular day, the active dataset can be set to only tweets from that day, and if further it is found that most of those tweets were posted by the same user, only that user’s tweets from that day.

What such exploration looks like in practice has multiple levels. At the broadest level of the current study, such things were measured as topical and relative temporal frequency of what the riders and teams posted at different points of the project’s timeframe and, following the lead of Page (2012), how frequently they used hashtags, @mentions, and retweets. Given, however, that such overall measures are likely to give misleading representations, especially in a corpus of over 100 Twitter users, overall measures, when present, are considered (and presented below) alongside equivalent measures for individual racers represented in the corpora. Inversely, individual measures are, when suitable, discussed along with comparable means and standard deviations. Ultimately, however, these quantitative measures serve largely to frame the qualitative analysis. Navigation of the corpus involved following two kinds of path: a quantitative path facilitated by the software (e.g. looking at frequency over time of a specific hashtag and comparing early tweets to later ones) and a qualitatively-emergent path of patterns noticed in the tweets themselves (e.g. how riders refer to their teammates). The bulk of the analytical work in the study involved reading the collected tweets in TWIG and hand-coding various emergent affiliative patterns in how the racers construct identity on Twitter.

Analysis and Discussion

The analysis below is grouped by the kinds of identity or values that are constructed in the bicycle racing community rather than, for example, by individual discussions of hashtags, @mentions, and retweets. This allows discussion of how the construction of said identity or values might involve combinations of these discourse activities. First, however, it is useful to provide some general statistics regarding the composition of this study's corpus. Of the 130 racers who started the Tour Down Under, 121 have Twitter accounts. Of those, 110 posted during the four-month period considered in this study. As noted above, 106 of these racers are represented in the TWIG dataset, which contains 4,172 individual tweets. The average racer tweeted 39 times—about once every four days. However, the standard deviation is 48, so the notion of an "average" racer is ultimately rather flimsy. Frequencies for individual racers range from a single tweet during the four-month period to 380 tweets. Seven of the 110 racers tweeted more than one hundred times, and 23 tweeted fewer than ten times. This indicates that, while professional bicycle racers are extremely likely to *have* Twitter, very few use it with any regularity. Furthermore, variation among prominent, winning racers suggests that there is no particular correlation between rider status and a desire to maintain an online presence. These basic statistics were all found within TWIG's Analysis environment by selecting the "user name" option in the Metadata Statistics Pane, as shown in Figure 3.5.



106 different, 39.36 avg, 48.11 sd

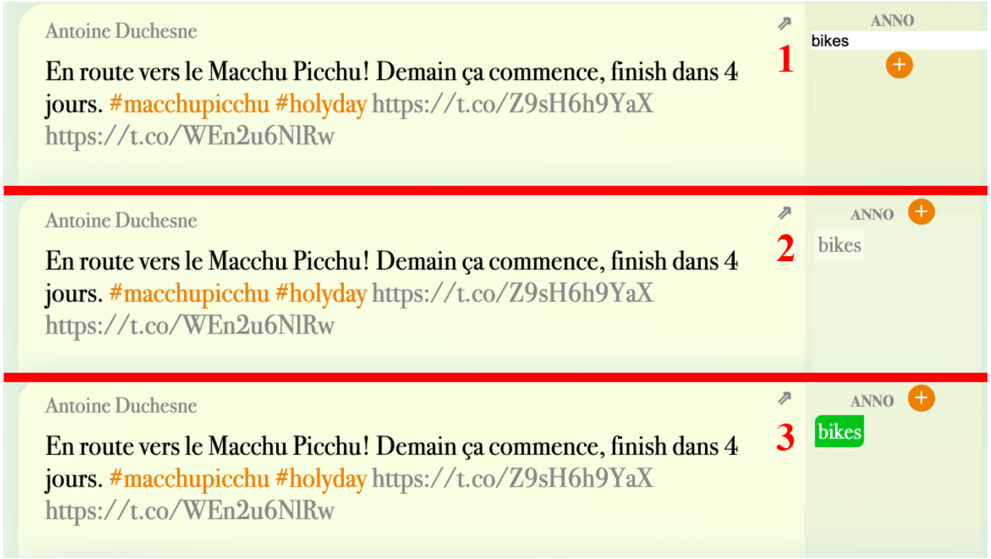
別府 史之 FUMY BEPPU	380
Elia Viviani	237
Nathan Haas	129
Tiago Machado 🇵🇹	126
Marcel Sieberg	114
Koen de Kort	107
Rui Costa	106
Daryl Impey	87
Peter Sagan	86
Carlos Barbero	85
Chad Haga	83
Thomas De Gendt	83
Valerio Agnoli	81

Figure 3.5 - User posting frequency table in TWIG

Baseline Identity

Identity construction among professional cyclists on Twitter operates from a kind of baseline. Racers tweet as professional bicycle racers. This is to say that their tweets almost always pertain somehow to their involvement in cycling and, crucially, emerge from it. Most tweets pertain to cycling explicitly, which for the purposes of this study means the tweet includes cycling imagery or hashtags or refers in some way to a race, a training activity, cycling equipment (bicycles, clothing), a team, or an industry sponsor. Not all 4,172 tweets in this study's corpus were coded for whether they dealt with cycling, but instead four 100-tweet temporal subsets of the corpus were considered.

This was achieved in TWIG by first adding an annotation option, "bikes," to the project. Unlike the addition of the "leader" annotation option, this was added from the Analysis Environment, the process of which is shown in Figure 3.6. To view the desired temporal subsets, the "time" option in the Metadata Statistics pane was selected, the subset interval was set to either "month" or "day," and the desired temporal ranges were selected, as shown in Figure 3.7.



Antoine Duchesne

En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

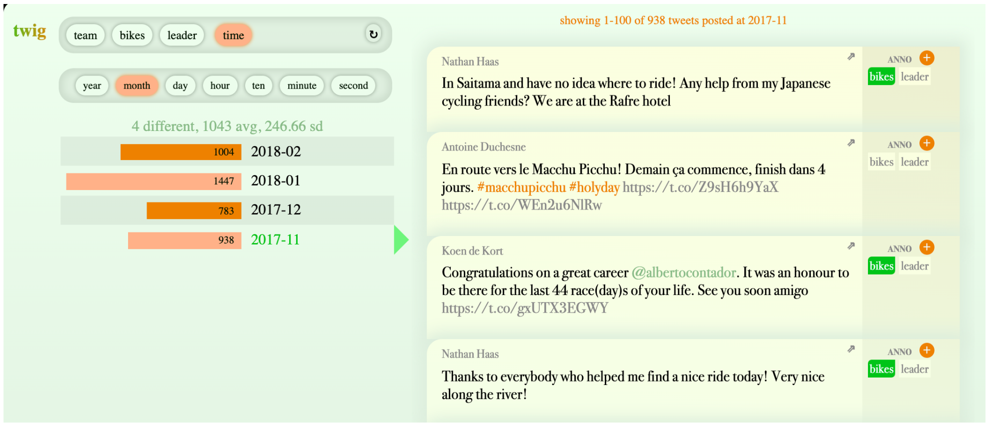
Antoine Duchesne

En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

Antoine Duchesne

En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

Figure 3.6 - "Bikes" annotation option creation



4 different, 1043 avg, 246.66 sd

1004	2018-02
1447	2018-01
783	2017-12
938	2017-11

showing 1-100 of 938 tweets posted at 2017-11

Nathan Haas
In Saitama and have no idea where to ride! Any help from my Japanese cycling friends? We are at the Rafre hotel

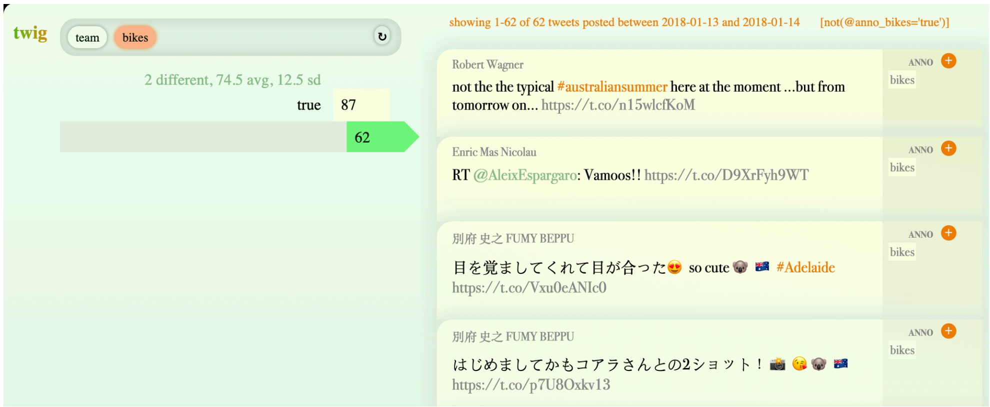
Antoine Duchesne
En route vers le Macchu Picchu! Demain ça commence, finish dans 4 jours. #macchupicchu #holyday https://t.co/Z9sH6h9YaX https://t.co/WEn2u6NlRw

Koen de Kort
Congratulations on a great career @albertocontador. It was an honour to be there for the last 44 race(day)s of your life. See you soon amigo https://t.co/gxUTX3EGWY

Nathan Haas
Thanks to everybody who helped me find a nice ride today! Very nice along the river!

Figure 3.7 - Tweets selected for TDU dates

Because TWIG shows tweets 100 at a time, annotating the first 100 tweets from each temporal subset was achieved by simply clicking the “bikes” annotation option for each tweet that pertains to cycling on the first page of each temporal subset’s data. Totals for each subset were then found by navigating to the “bikes” option in the Metadata Statistics pane, as shown in Figure 3.8 below.



2 different, 74.5 avg, 12.5 sd

true	87
	62

showing 1-62 of 62 tweets posted between 2018-01-13 and 2018-01-14 [not(@anno_bikes='true')]

Robert Wagner
not the the typical #australiansummer here at the moment ...but from tomorrow on... https://t.co/n15wlcfKoM

Enric Mas Nicolau
RT @AleixEspargaro: Vamoos!! https://t.co/D9XrFyh9WT

別府 史之 FUMY BEPPU
目を覚ましてくれて目が合った😍 so cute 🐨 🇦🇺 #Adelaide https://t.co/Vxu0eANIc0

別府 史之 FUMY BEPPU
はじめましてかもコアラさんとの2ショット！ 📸 😘 🐨 🇦🇺 https://t.co/p7U8Oxkv13

Figure 3.8 - Non-bike-related tweets selected for TDU dates

The first two subsets considered tweets posted in the off season, before the TDU. The first consisted of the first 100 tweets in the corpus, spanning from November 1 to November 4, 2017. Of these tweets, eighty were explicitly related to cycling. The second subset consisted of the first 100 tweets posted in December 2017. Of these, 71 were explicitly related to cycling. For the third subset, the temporal interval option in TWIG was switched to “day” to enable the selection of tweets starting on January 13, 2018, the day before the TDU began. Of these, 87 were explicitly related to cycling. For the final subset has a start date of January 18, the most frequently tweeted-on day of the TDU. Ninety of these were explicitly related to cycling. Broadly, this suggests that racers tweet more exclusively about cycling the closer they are to a race, which isn't surprising. What is particularly notable, however, is that nearly three quarters of the measured pre-season activity still pertains to cycling, which emphasizes the centrality of the racers' profession to their online identity.

This is not the whole picture, however, because even many of the tweets that are not explicitly related to cycling retain an implicit connection. In TWIG, tweets not coded as cycling related were isolated by selecting each of the measured temporal subsets, then navigating to the "bikes" metadata and selecting the blank button, which represents all non-true values[2]. For example, Figure 3.4 above shows those tweets from the day before the start of the TDU that were not coded as explicitly cycling related. Several of these, as shown in the examples below, were superficially about travel or tourism, including pictures of the riders posing with either a koala, a kangaroo, or a snake.

6. DiegoUlissi: Oggi incontro ravvicinato con Koala e Canguro!!!🐨🐨🐨 https://t.co/dn2lwt3IhU (2018-01-13T07:20:35.000Z)
7. Fumybeppu: 目を覚ましてくれて目が合った😍 so cute 🐨 🇦🇺 #Adelaide https://t.co/Vxu0eANIc0 (2018-01-13T04:53:27.000Z)
8. cimo89: Selfie con il koalino 🐨 https://t.co/S1ZTxxioX0 (2018-01-13T07:35:55.000Z)
9. koendekort: This snake had me in a chokehold and I could feel it's tongue on my ear. I won't lie... my face might not give it away but i found this a tad scary 🙈 https://t.co/goyPO7MX2X (2018-01-13T12:11:16.000Z)

[2] "Non-true values" accounts for both tweets that were considered and tweets that were not considered because they fell within the active temporal subset but not the 100 tweets from that subset that were coded. For researchers who prefer a more explicit negative coding, an option would be to have two coding values, such as "bikes" and "not bikes."

Considering that these animals—particularly koalas and kangaroos—are conceptually associated with *Australia,* and the racers' reason for being in Australia was to compete in a race, these tweets actually relate to cycling as well. Australia was at the time broadly accessible to both bicycle racing's participants and its fans as the location of a race, so tweets that index Australia also activate racing and, as such, the racers' reason for being there. In a sense, then, the racers' baseline identities imbue not-explicitly-cycling-related tweets with cycling-related meaning. A more extreme example of meaning emerging from a baseline identity rather than directly from the text can be seen in example 10 below, in which a tweet consisting entirely of fire emoji is unambiguously in reference to the heat of the racing on the day it was posted—a day on which several other racers were also tweeting about the heat of the racing conditions.

10. JhonatanRVal: 🔥🔥🔥🔥🔥🔥🔥 https://t.co/3EhXSpM5ON (2018-01-19T08:11:37.000Z)

This notion of a baseline identity that emerges in part from the racers' Twitter audience seems to also come with some constraints on non-cycling discourse. For example, American racer Chad Haga occasionally tweets about American politics, and one such tweet prompted the reply in example 12 below:

11. fletch563: So @ChadHaga is a lemming believing the BS that comes out of DC. smh. **stick to** 🚴
12. ChadHaga: @fletch563 I must have missed the part of my contract that says I'm not allowed to have a political standing . And I suppose you're a politician, seeing as you've got one?

Example 11 is not directly part of this study's corpus. It was found because, in EXAMPLE 6, Haga replied. In TWIG, each tweet is accompanied by a link to its counterpart on Twitter itself, which in this case was used to see Haga's reply in context.

While this exchange is the only example of direct censure identified in this study, it shows that professional cyclists, like other athletes in the public sphere, aren’t on Twitter as “ordinary users.” They are public figures whose tweeting seems to be expected by at least some members of their audience to “stick to” the subjects for which they are known. Interestingly, while Haga rejects this pigeonholing of his identity, specifically as it manifests on Twitter, he does so by noting that political discourse is not restricted by his professional cycling contract. He does’t reject a baseline identity as a cyclist, but instead negotiates what that identity should look like online by referring to its offline source.

Affiliation with events

The offline existence of the bicycle racing community gives more to this community’s Twitter counterpart than a baseline identity for the riders. It also provides the dominant topics of discourse. This is not surprising. Bicycle racing is a community with clearly-defined, scheduled events: races. This manifests in spikes in overall tweeting frequency that occur when races are taking place. These spikes can be visually identified in TWIG by looking at temporal statistics at different interval settings. Figure 3.3 above shows, at the month level, notably increased activity in Jan 2018, which of the represented months is the one with the most racing.

At the day level, spikes in posting frequency tend to correlate with specific races. For example, January 28, which has as many tweets as the days before and after it combined, is the date of the Cadel Evans Road Race. These differences in frequency can also be given some basic statistical context within TWIG. For example, by selecting only the dates of the Tour Down Under and refreshing the Metadata Statistics Pane (by clicking the refresh button in the top right of the pane) statistics for only that day are calculated. As shown in Figure 3.9, compared to the corpus-wide statistics, during the TDU racers posted both much more frequently (an average of around 73.6 tweets per day, as compared to 34.7) and more regularly (with a standard deviation of 9.3, as compared to 26.7).

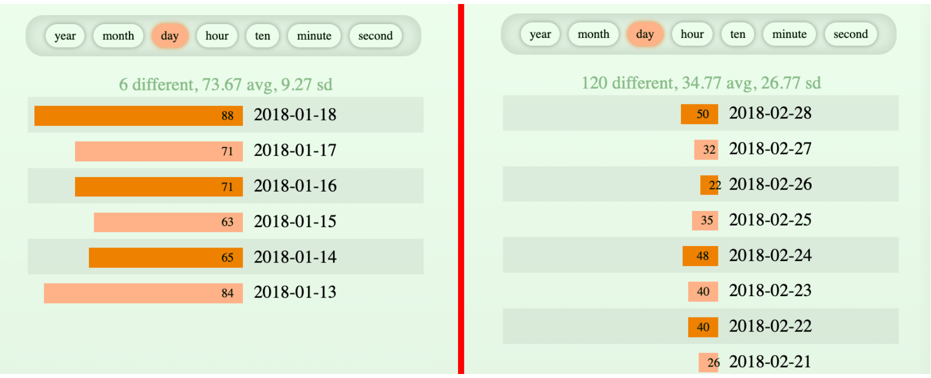


Figure 3.9 - Daily tweet frequency statistics for TDU period (left) and full corpus (right)

In this sense, the racers seem to construct identity via their involvement with the races. This is enacted not just in the text of their tweets but in the riders' affiliative use of hashtags and @mentions. Their particular use of hashtags and @mentions, however, is somewhat surprising. Most racers don't use a lot of hashtags in their own tweets. Instead, hashtags show up mostly in retweets. For example, of the tweets collected for this study that contain the hashtag #tdu (or #TDU), sixty-eight percent are retweets. Only thirty-five percent of tweets in the entire corpus are retweets. This disparity seems to indicate that racers are less likely to use event hashtags than other kinds of hashtags in their own tweets. This suggests that the racers themselves are not especially interested in indexing their tweets in the discourses surrounding the races, perhaps because their affiliation is inferable within the community that would follow such discourses. Furthermore, when racers *do* use event hashtags, those hashtags often serve as interpretive guides (as in examples 13 and 14 below) to or grammatical components (examples 15 and 16) of the main body of a tweet.

13. Fumybeppu: The scorching sun grilled us. Tomorrow even more Hot!!🔥@tourdownunder 🇦🇺 #stage2 #TDU 📸: @keitsuji https://t.co/9CZ8RXIsdt (2018-01-17T09:23:14.000Z)

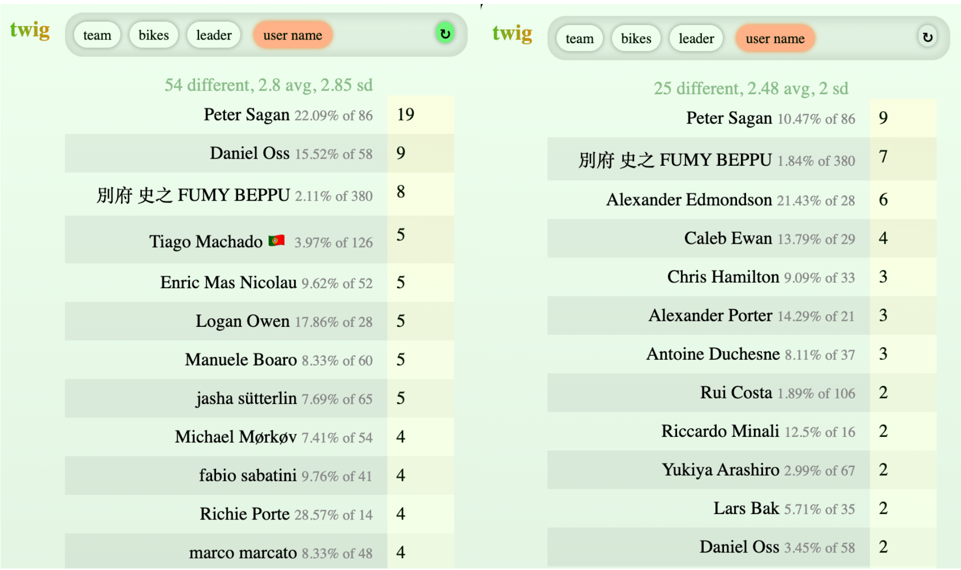
14. CalebEwan: A @MitcheltonSCOTT 1st & 2nd! What a ride by the boys yesterday! #tdu https://t.co/qqB3TrLwEL (2018-01-17T22:01:04.000Z)

15. RohanDennis: Nice cool change just before the hot spell in #Adelaide next week. Can't wait for the #tdu to start tomorrow https://t.co/60FYkzjdC3 (2018-01-13T10:48:31.000Z)

16. eliaviviani: Cannot wait to kick off #TDU and race on my Venge ViAS. Doesn't get more aero than this! #everydayaero @iamspecialized @quickstepteam https://t.co/CXd6Ya0vr7 (2018-01-15T07:34:55.000Z)

Example 13 is an especially strong example of hashtags that guide interpretation, because the main text of the tweet is only about heat. Anyone who had been outside somewhere hot that day could have posted the same. It's only via some combination of Fumy Beppu's baseline identity as a professional cyclist and the context provided by the hashtags that we know he's saying it was a hot day to be out racing bicycles. Using this kind of hashtag has a couple of effects. First, it uses fewer characters than explicitly stating a location and activity during which the heat was experienced, which is salient on Twitter, where character limits apply. But perhaps more importantly, this use of the #tdu hashtag emphasizes the heat, or the experience thereof, as the purpose of the tweet. In identity-constructive terms, then, the #tdu hashtag might actually serve more to allow Beppu to construct himself inline with cycling's value of suffering than it does to affiliate him with online discourse surrounding the race. This is all to say that not only do racers use event hashtags relatively infrequently, when they do use them, they often do so in ways that de-emphasize hashtags' technical function of connecting the tweets to broader discourses.

While racers don't tend to use hashtags to refer to races, they regularly use @mentions. Races often maintain Twitter accounts for promotion and engagement, and racers are actually more than twice as likely to refer to a race with an @mention than with a hashtag. For example, this study's corpus, filtered to consider only original tweets—no retweets, quote tweets, or directed tweets—contains 50 instances of #tdu and 138 of @tourdownunder. Variance between different riders is substantial, as shown in Figure 3.10 below, with only 54 of the 106 riders represented in the corpus ever using @tourdownunder and 25 using #tdu. Nevertheless, these levels of @mention and hashtag use are more-or-less representative of the behaviors of racers who are active on Twitter and, as such, are salient considerations for analysis of how these racers construct identity.



54 different, 2.8 avg, 2.85 sd

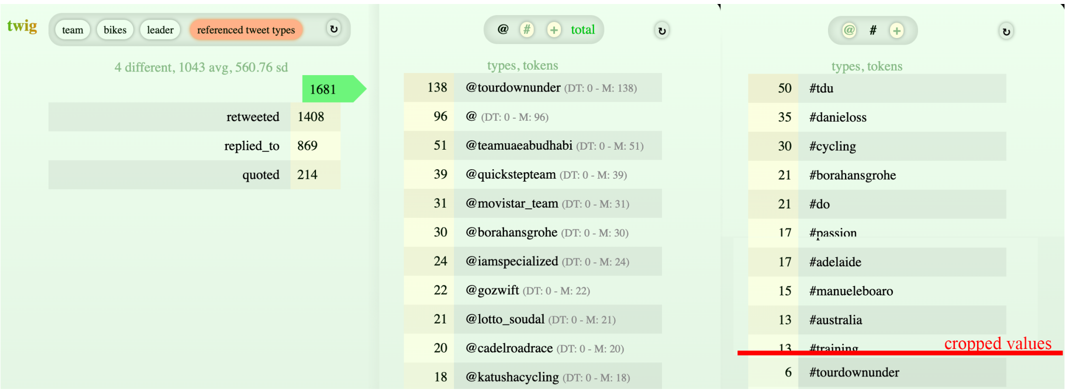
Peter Sagan 22.09% of 86	19
Daniel Oss 15.52% of 58	9
別府 史之 FUMY BEPPU 2.11% of 380	8
Tiago Machado 🇵🇹 3.97% of 126	5
Enric Mas Nicolau 9.62% of 52	5
Logan Owen 17.86% of 28	5
Manuele Boaro 8.33% of 60	5
jasha sütterlin 7.69% of 65	5
Michael Mørkøv 7.41% of 54	4
fabio sabatini 9.76% of 41	4
Richie Porte 28.57% of 14	4
marco marcato 8.33% of 48	4

25 different, 2.48 avg, 2 sd

Peter Sagan 10.47% of 86	9
別府 史之 FUMY BEPPU 1.84% of 380	7
Alexander Edmondson 21.43% of 28	6
Caleb Ewan 13.79% of 29	4
Chris Hamilton 9.09% of 33	3
Alexander Porter 14.29% of 21	3
Antoine Duchesne 8.11% of 37	3
Rui Costa 1.89% of 106	2
Riccardo Minali 12.5% of 16	2
Yukiya Arashiro 2.99% of 67	2
Lars Bak 5.71% of 35	2
Daniel Oss 3.45% of 58	2

Figure 3.10 - @tourdownunder frequency (left) and #TDU frequency (right)

These figures were found in TWIG by first selecting “referenced tweet types” in the Metadata Analysis Pane, as shown in Figure 3.11 below, and then applying the unlabeled item in the frequency table as a filter, effectively telling TWIG to only consider tweets that do not reference other tweets in any way. With this filter applied, the raw frequencies were found in the Text Analysis Pane, and per-rider statistics (as in Figure 3.10) were found by selecting “user name” in the Metadata Analysis pane.



4 different, 1043 avg, 560.76 sd

	1681
retweeted	1408
replied_to	869
quoted	214

types, tokens

138	@tourdownunder (DT: 0 - M: 138)
96	@ (DT: 0 - M: 96)
51	@teamuaeabudhabi (DT: 0 - M: 51)
39	@quickstepteam (DT: 0 - M: 39)
31	@movistar_team (DT: 0 - M: 31)
30	@borahansgrohe (DT: 0 - M: 30)
24	@iamspecialized (DT: 0 - M: 24)
22	@gozwift (DT: 0 - M: 22)
21	@lotto_soudal (DT: 0 - M: 21)
20	@cadelroadrace (DT: 0 - M: 20)
18	@katushacycling (DT: 0 - M: 18)

types, tokens

50	#tdu
35	#danieloss
30	#cycling
21	#borahansgrohe
21	#do
17	#passion
17	#adelaide
15	#manueleboaro
13	#australia
13	#training
6	#tourdownunder

Figure 3.11 – Process for finding filtered @mention and hashtag frequency tables

A key difference between @mentions and hashtags is that @mentions are much more restrictive. When clicked, they direct users not to a broader discourse, but rather to a feed containing only the @mentioned account's activity. In the context of the bicycle racing community, the effect appears to be promotional. By referring to races using @mentions instead of hashtags, racers are choosing not to seek a broader audience by indexing their tweets within a hashtag-level discourse, but are instead assuming already-existing audience of fans who are interested in their activities—and who can then click through to the @mentioned race's feed for more information about the race in which their favored rider is participating. Like hashtags, racers use @mentions both grammatically within the tweet and as associative tags at the ends of tweets, as shown in the examples below.

17. PhilBauhaus: Happy to be back in Australia! 🇦🇺#australia #adelaide **@tourdownunder @ Adelaide**, South Australia https://t.co/QuXVyh4ufc (2018-01-06T11:13:59.000Z)
18. AndreGreipel: Nice to be back **@tourdownunder** #Adelaide for the first km in the heat of 2018. **@hjcsports @rudyproject @lotto_soudal** https://t.co/GGjZYxadH8 (2018-01-06T01:34:14.000Z)
19. Daniel87Oss: Today it was a good first stage at **@tourdownunder**. #KeepFighting #tourdownunder #DO #DanielOss **@BORAhansgrohe @iamspecialized** credits **@bettiniphoto** https://t.co/073ebuNcI7 (2018-01-16T10:09:13.000Z)

These examples show that @mentions are often used, textually, much like hashtags. Example 17 demonstrates that @mentions can provide information that guides interpretation in much the same way Scott (2015) and examples 13 and 14 above have shown occurs with hashtags. Example 18 is also particularly interesting, because it appears to use the @ component of the @mention grammatically, such that it should be read as "nice to be back at tourdownunder," where in example 19 Daniel Oss realizes the "at."

Affiliation with Teams

Racers often construct themselves on Twitter explicitly as members of teams. In doing so, as occurred with races above, the racers seem to prefer to refer to their teams, most of which maintain a specific team account, using @mentions, rather than hashtags. More frequent still are retweets of team accounts. These frequency differences of course vary between individual racers and teams. Table 3.1 below shows these differences at a team level for a randomly-selected seven of the nineteen teams represented in this study.

Team	Total	**Retweets**	**@Mentions**	**Hashtags**
Quickstep (QST)	442	87 (19.7%)	74	11 #thewolfpack 4 #wolfpack
Bora Hansgrohe (BOH)	244	28 (11.4%)	37	22 #borahansgrohe
UAE Team Emirates (UAE)	272	40 (14.7%)	53	14 #uaeteamemirates
Lotto Soudal (LTS)	350	55 (15.7%)	22	4 #lottosoudal2018 2 #lottosoudal
Equipe FDJ (FDJ)	166	48 (28.9%)	11	4 #equipegroupamafdj 2 #groupamafdj 2 #fdj
Katusha Alpecin (TKA)	330	8 (2.4%)	19	8 #teamkatushaalpecin 8 #raceasafamily
Astana (AST)	162	49 (30.2%)	8	37 #astanaproteam 2 #astana

Table 3.1 – Manner of team affiliation in racers' original tweets

This table shows some notable differences between teams, which can likely be attributed to varying levels of organizational interest in social media. For example, the FDJ team account posted 723 times during the study period while the TKA team account posted only 189 times, so it makes sense that FDJ riders frequently retweet their team while TKA riders do not. Despite these differences, however, the overall pattern for constructing team affiliation on Twitter seems to favor retweets over @mentions over hashtags. This ranking of racers' preferred manners of textually representing themselves within their teams has identity and value construction implications that emerge in interaction with the content of the team accounts. Considering, for example, that nearly all of the retweets that include the #tdu hashtag are retweets of teams' corporate Twitter accounts, it seems that team accounts are, unlike the racers themselves, interested in using hashtags to include themselves in Twitter-wide discourses surrounding specific races. An effect of this is that a user following the hashtag for a specific race is most likely to see individual racers from that discourse framed by tweets from their teams. Consider the following examples that contrast two racers' original tweets that @mention their teams with team retweets.

20. MarcelSieberg: So proud of my team @Lotto_Soudal for the first victory @RondeVlaanderen Great ride @petosagan @JasperDeBuyst https://t.co/cDSAzsUKWR (2017-12-20T12:42:44.000Z)
21. MarcelSieberg: RT @Lotto_Soudal: 📷Hugs for #TDU stage winner @AndreGreipel More pictures here: https://t.co/TT9AxZP2YI #flickr https://t.co/b0rZ5k4Yn1 (2018-01-16T21:59:35.000Z)
22. MichaelMorkov: Trainingcamp with #TheWolfpack is done - @NikiTerpstra and I is ready to race against the young guns in Ballerup the 29.dec @quickstepteam 🤝@katushacycling https://t.co/6csre7BHd5 (2017-12-20T15:31:45.000Z)
23. MichaelMorkov: RT @quickstepteam: We're going to a land Down Under, spirited and seven in number: https://t.co/vgNrIHynL8 #TDU https://t.co/x7KgMIvWPR (2018-01-05T10:59:02.000Z)

In these tweets, the racers tweet in such a way that, for their personal followers (i.e. fans), who would see all of the tweets, they act as promoters of the races in which they participate and the sponsors for which they do so, but for more general audiences who follow the hashtag discourse of a particular event, where only the retweets (21 and 23) would appear, the racers are primarily framed by their team affiliation.

Bicycle racers' heavy use of retweets is actually notable relative to other kinds of users that have been analyzed on Twitter. As a point of comparison, racers retweet more than any of the groups considered in Page (2012). Page considered celebrities, corporations, and "ordinary" users. The most analogous of these groups to the cyclists in the current study is celebrities, who Page found retweet about five percent of the time (p. 186). The racers in the current study retweeted about 34 percent of the time (39 percent including quote tweets). Figure 3.12 below shows the full balance of tweet types among various kinds of Twitter users, including accounts managed by bicycle racing teams themselves.

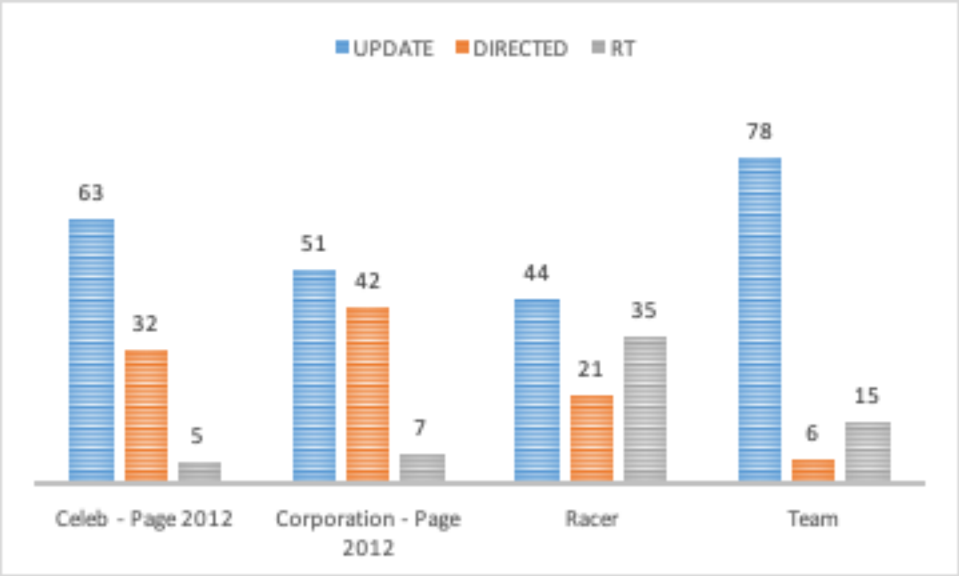


Figure 3.12 - Frequency of different post types

Considering the six years between Page’s (2012) study and the data collected for the current study, the difference in retweet frequency between Page’s groups and the bicycle racers of current study could be a result of general change over time in how people use Twitter. I don’t believe this to be the case, however, because professional cyclists retweet their teams disproportionately. For example, nearly 64 percent of retweets from riders on team FDJ originate from the team’s official Twitter. Retweets (all, not just team) account for about 41 percent of FDJ riders’ tweets. Without team retweets, then, racers’ retweet frequency would be much closer to that Page found for celebrities. Furthermore, the racers’ disproportionate retweeting of team accounts cooccurs with team accounts’ high frequency of original posts, for the racers to retweet. This further suggests that racers’ strong affiliation with their teams is not simply a means of individual identity construction, but rather a means of constructing the values of the sport, with the re-tweeting behavior constructing on Twitter bicycle racing’s offline value of cooperation by framing racers consistently within a cooperative unit, a team.

Constructing team affiliation outside the commercial entity

While racers construct themselves within their teams, they also construct the nature of that relationship with their teams using language that resists its inherent professional or commercial nature in favor of a familial construction. For example, racers often refer to their teammates with some variation of “the boys,” as shown in the examples below.

24. eliaviviani: **The boys are the best**👊🏻 big mistake from myself in the sprint cost me the results but we’re there 💪🏻 and I feel really 🍀 lucky! Thanks @quickstepteam, first race ✅ done https://t.co/BvW3sjCxHc (2018-01-21T05:31:48.000Z)
25. darylimpey: That was some day but **boys** did amazing job. Close 2nd but not quite enough to throw the hands up just yet. Tomorrow will be a mega day for all on Willunga. See you out there (2018-01-19T07:45:29.000Z)

26. simongerrans: Morning media commitments with **the boys**. @richie_porte @rohandennis @tourdownunder @bmcproteam https://t.co/UjtWZ5G9ac (2018-01-11T23:22:22.000Z)
27. tomjelteslagter: Good day for us @tourdownunder the **@teamdidata boys** helped me to finish 3rd on Willunga Hill.… https://t.co/rVUTqSTzAr (2018-01-20T12:51:36.000Z)
28. PhilBauhaus: Happy to finish second after a very hot stage @tourdownunder 🔥A big thanks to **the teamsunweb boys** to make this possible. (2018-01-18T08:01:03.000Z)[Instagram crosspost]

As in examples 27 and 28, "the boys" often occurs with a modifier such as "the [@team] boys," which maintains more corporate affiliation. However, in this study's corpus, there are 27 instances of unmodified use of "the boys," compared to 16 instances of "the @team boys," which suggests it is important for racers to construct (at least the appearance of) bonds with their teammates that exist outside their inherent professional nature.

A more Twitter-specific means by which teammates construct not-explicitly-professional bonds is hashtag use. While it is noted above that racers don't tend to refer to their teams via hashtag, racers from two of the teams use hashtags that refer to their teams without naming them—which is to say they refer to their team*mates* with hashtags. Several Quickstep racers, as shown in the examples below, use either #wolfpack or #thewolfpack, and Katusha Alpecin racers use #raceasafamily.

29. eliaviviani: Winning mood🤪 @quickstepteam #**thewolfpack** 📸 @therussellellis https://t.co/b2TOkF2usU (2018-02-22T21:08:18.000Z)
30. MichaelMorkov: My first raceday with #**TheWolfpack** really enjoy to be part of this strong group 👊 now we are looking forward to get on with @tourdownunder and win some stages @quickstepteam https://t.co/aGX1qqdZL7 (2018-01-14T12:16:58.000Z)
31. SabatiniFabio: Classic way when we win of holm12.16 #**thewolfpack** @quickstepteam #waytoride 📸… https://t.co/IPUVe1123y (2018-02-07T20:28:10.000Z)

While these tweets also @mention the team, use of these hashtags allows racers to construct separate affiliations with their teammates and their team organizations—in a way that further illustrates how racers use hashtags generally. As discussed above, racers don't tend to use hashtags to index themselves in the broader the races they are in or the names of their teams. They enter these discourses via team retweets. However, the use of teammate, but not explicitly team organization, affiliative hashtags has the effect of creating and indexing the racers within searchable discourses on Twitter that include only tweets in which these teammates refer to one another. As in the examples above, the content of these tweets tends to refer to time spent on bikes together, whether in races or training. One way to interpret this is that the racers are constructing their online identities to match their offline position in the community not as commentators, but as participants with co-participant bonds facilitated, but not entirely defined, by commercial entities.

Constructing affiliation between non-teammates: the gruppetto

Because cooperation in bicycle racing happens not just within teams, but also between riders on different teams, one might expect to see such cooperation represented in affiliative behavior on Twitter. It is, but very rarely, probably because membership of intra-team cooperative groups is almost exclusively ad-hoc and fleeting, so the bond is only meaningful during the race in which it occurs. One group that is recurring, though, is the gruppetto, the group of various teams' sprinters that emerges at the back of the race on the mountainous days. This group provided an example of inter-team cooperation being enacted on Twitter. Sam Bennett posted the tweet in example 32 below, which included the image in Figure 3.13. This was was subsequently retweeted by two other regular gruppetto members from other teams, Marcel Seiberg and Robert Wagner.

32. Sammmy_Be: I can't tell you how many times this has happened to me in the gruppetto 😐😅 https://t.co/VKSHVxQcRk (2017-11-02T14:56:57.000Z)

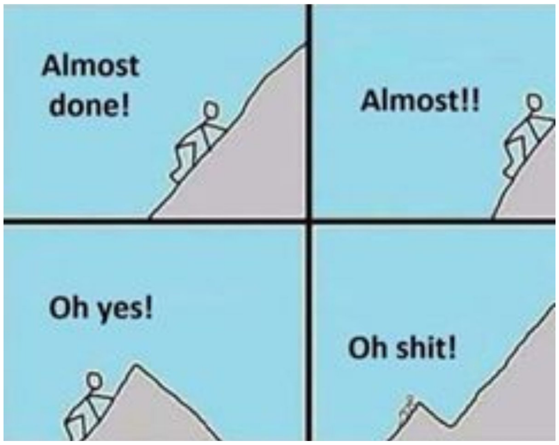
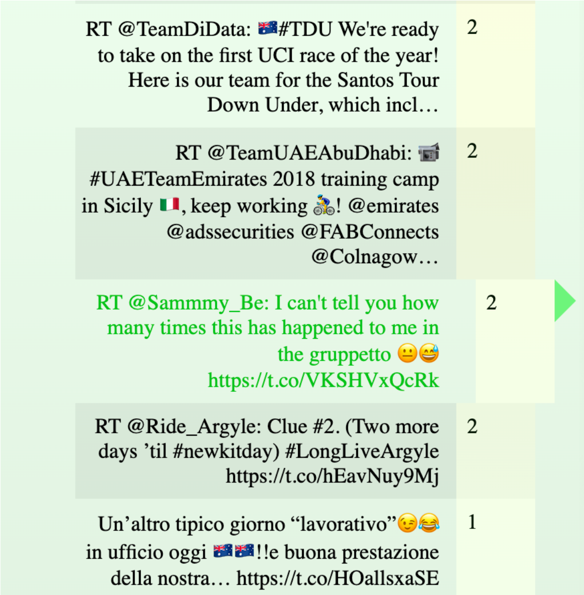


Figure 3.13 - The gruppetto comic

Bennett's tweet both indexes the gruppetto in the text and references the reason the group exists (getting up the hills) in the image, constructing the group as one that exists with a specific purpose. These tweets taken together show that the kinds of affiliation retweets construct can expand beyond professional relationships to include less formal sub-groups within the cycling community.

This tweet was found in TWIG by scanning the "tweet text" frequency table in the Metadata Analysis Pane, as shown in Figure 3.14 below. Because TWIG treats everything about a tweet as metadata, the full text of each tweet can measured for frequency, meaning multiple tweets with exactly matching text can be found very easily. This offered a viable method of finding retweets that proliferated within the corpus (i.e. among the riders themselves) as opposed to retweets that proliferated more generally, which could be found by looking through "retweet count" metadata.



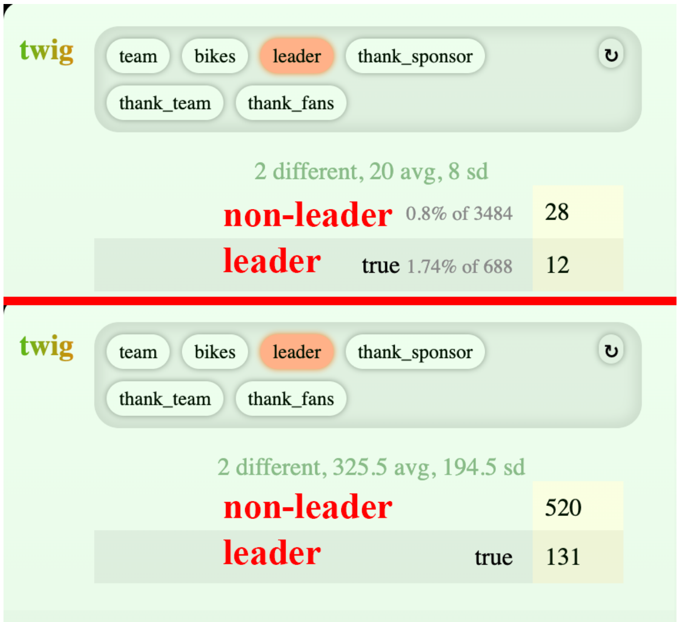
RT @TeamDiData: 🇦🇺#TDU We're ready to take on the first UCI race of the year! Here is our team for the Santos Tour Down Under, which incl…	2
RT @TeamUAEAbuDhabi: 📹 #UAETeamEmirates 2018 training camp in Sicily 🇮🇹, keep working 🚴! @emirates @adssecurities @FABConnects @Colnagow…	2
RT @Sammmy_Be: I can't tell you how many times this has happened to me in the gruppetto 😐😅 https://t.co/VKSHVxQcRk	2
RT @Ride_Argyle: Clue #2. (Two more days ’til #newkitday) #LongLiveArgyle https://t.co/hEavNuy9Mj	2
Un’altro tipico giorno “lavorativo”😉😂 in ufficio oggi 🇦🇺🇦🇺!!e buona prestazione della nostra… https://t.co/HOallsxaSE	1

Figure 3.14 - Sam Bennett’s tweet in “tweet text” frequency table

Constructing the individual result in a team sport

Bicycle racing’s notion of cooperation is more complex than simply belonging to a team or an intra-team group on the road. Fundamentally, racing is a sport, and cooperation on the road is done to facilitate a competition with individual winners. As such, both team leaders (racers pursuing results) and the racers who support them must orient themselves relative to the leader’s results and the rest of the team’s involvement therein. Both in the act of racing and in the act of communicating about the race, success must be shared.

To explore the differences between how team leaders and supporting racers construct success on Twitter, I started by looking broadly at differences between racers annotated as “leader” and those who are not. To maintain the validity of this grouping, because the “leader” code refers only to the racers’ roles in the TDU, the following analysis considers only tweets posted between January 13 and 22, 2018, which represents the TDU race period with a couple days of padding before and after. During this period, about half of leaders’ tweets @mention or retweet team accounts, while only about a third of supporting racers’ tweets do so. This suggests that teams' leaders may feel a greater need to emphasize their team affiliation. Similarly, the keyness tool in TWIG's Text Analysis Pane identified "thanks" as a word of possible interest, with leaders using it in 9.2 percent of their TDU-period non-retweets, compared to 5.4 percent for the supporting racers. This was calculated in TWIG, as shown in Figure 3.11 below, by filtering the dataset to include only "thanks" tweets from the TDU period and dividing them by totals for the TDU period, such that 28 of 520 (5.4 percent) non-leaders' tweets contained thanks.



2 different, 20 avg, 8 sd

non-leader	0.8% of 3484	28
leader	true 1.74% of 688	12

2 different, 325.5 avg, 194.5 sd

non-leader		520
leader	true	131

Figure 3.15 - Frequency of thanks tweets (top) and total tweets (bottom) in TDU period

It is, of course, unsurprising that leaders use "thanks" more frequently, given that leaders are more likely to have a reason to say thanks. Ultimately, then, given the degree of variance between individual racers' use of keywords or @mentions, combined with the small number of tweets that actually pertain to specific race results, quantitative measures cannot meaningfully describe how racers construct their or their teammates' success.

Individual tweets pertaining to race victories were therefore identified in TWIG by filtering the TDU-period dataset by team for each of the five teams that won either a stage or a general classification. This drastically reduced the size of the haystack in which result-oriented tweets were then manually searched for. What was found is that racers share success on Twitter in a variety of ways, both textual and structural, to a variety of effects. As suggested by the quantitative measures described above, contenders often thank or praise their teams, which are often, but not always referred to with a team account @mention. The examples below illustrate the degree of variation in both approaches to sharing victory and the values those approaches reflect.

33. petosagan: First WorldTour victory of the year on stage 4 of @tourdownunder. It's extremely hot in @southaustralia but **thanks to the work of the @BORAhansgrohe squad, I won!** Tomorrow, we have to support @JayMcCarthy1 for the GC. https://t.co/WDJA0PK6gs (2018-01-19T12:34:13.000Z)
34. CalebEwan: **A @MitcheltonSCOTT 1st & 2nd!** What a ride **by the boys** yesterday! #tdu https://t.co/qqB3TrLwEL (2018-01-17T22:01:04.000Z)
35. AndreGreipel: RT @Lotto_Soudal: 📷Hugs for #TDU stage winner @AndreGreipel! More pictures here: https://t.co/TT9AxZP2YI #flickr https://t.co/b0rZ5k4Yn1 (2018-01-16T22:08:22.000Z)

There are several notable differences between these tweets. Sagan (33) is the only of these with a first-person acknowledgement ("I won") of an individual victory. Ewan (34) describes his victory as specifically a team result and Greipel (35) only acknowledges his win within the frame of a team retweet. These differences reflect differences between the racers themselves. Sagan is something of a superstar in bike racing, so it's not surprising that he would take personal credit. Ewan was, at the time, building a reputation as a sprinter, so it's not surprising that he would construct himself primarily as a team player. Greipel's professional racing career started in 2002, so a retweet likely reflects his level of interest in Twitter. A constant among these tweets, however, is framing the teams' involvement in the individual victories. While Sagan does specifically note his victory, the rest of the tweet, in a sense, makes up for it. The "I won" is both qualified by crediting and @mentioning his team's support and followed by a promise to support a teammate on the following day's stage. It is noteworthy that both Sagan and Ewan's tweets refer to their teammates and not just their teams, with Ewan using "the boys" as discussed above and Sagan using an @mentioned team as a modifier in "@BORAhansgrohe squad." This distinction between crediting the team and crediting teammates in tweets about individual results is common. The examples further illustrate how credit is distributed.

As seen in Greipel's tweet in Example 35, retweets are another mechanism by which racers frame their results within a team, but to a different effect than can be achieved in an original tweet. Unlike Greipel, other racers use retweets to supplement their original, team- and teammate-crediting tweets. Consider the following pairs of original tweets and retweets regarding, for Viviani (36 and 37), a win and, for Morkov (38 and 39), his role in a Viviani's win.

36. eliaviviani: **First win** of 2018✅ it's a special feeling **open the year account of @quickstepteam** #thankstoteamandstaff📸 @Dario_beli / @bettiniphoto https://t.co/0TNSPM9Wax (2018-01-18T13:29:05.000Z)
37. eliaviviani: RT @quickstepteam: Yeeeeeeeeeeeeeeeeeeeeees!!!!!! A fantastic **@eliaviviani wins stage 3** of #TDU in spectacular fashion! #WayToRide #No1in20… (2018-01-18T04:41:10.000Z)
38. MichaelMorkov: **Enjoying to be part of this great @quickstepteam first victory** of the year is **in the box** 🏆- 49 to go 👊#TheWolfpack https://t.co/0wy1RXOL3P (2018-01-18T05:04:14.000Z)
39. MichaelMorkov: RT @quickstepteam: .**@MichaelMorkov is doing a great job** in these closing kilometers! Hats off to our Danish rider! #WayToRide #TDU (2018-01-17T06:40:33.000Z)

In the racers' original tweets (36 and 38), both frame the win as a victory specifically for the @mentioned team. Interestingly, they also both do so using metaphors of collection, with Vivianni putting the win in an account and Morkov putting it in a box. Vivianni's use of "first win" even minimizes his agency in the result. That Viviani himself won is only explicitly stated in Example 37, a retweet of a team post. This is what boyd et al. (2010) call an "ego retweet" (p. 9). This retweeting of others aggrandizing them allows the racers to avoid *self*-aggrandizement, which would conflict with the value of honoring the sport’s cooperation. It’s worth noting that beyond these retweets allowing racers to safely broadcast their individual achievements, the fact that the achievements are being presented positively by an outside source gives them actually a little more credence. The most notable quality of the retweets, though is that the original tweeters are the teams, meaning the retweets not only highlight an individual achievement, they do so while framing the victory within the context of the team.

Examples 38 and 39 are particularly interesting because they begin to show how supporting racers construct themselves within their teams’ results. The methods by which they do so, of course, differ at times from those of leaders constructing their own results. Many supporting racers include themselves using retweets, as in 39 above and the further examples below.

40. MichaelMorkov: RT @eliaviviani: Primo sprint dell’anno fatto, grande lavoro dei ragazzi @quickstepteam, dobbiamo solo aggiustare il tiro👌🏻 #TDU [First sprint of the year done, **great work of the boys @quickstepteam**, we just have to adjust the shot] (2018-01-14-22:59:37.000Z)
41. Daniel87Oss: RT @BORAhansgrohe: What a finish to the #TDU! @petosagan finishes third at the end of this very fast-paced stage. - boh 2018 01 21 07:11:22

As with leaders’ tweets, these retweets of either a team leader (40) or a team account (41) frame the racers’ involvement with a victory around the team. This is a frame, however, that anticipates an audience with a baseline knowledge of the racers ’team affiliations. Without this knowledge, these tweets would not work. Example 40 could be taken as retweeted commentary, rather than one of “the boys” endorsing Viviani’s praise, and 41 could be interpreted as a simple congratulations, rather than a subtle “I was there, too” from Oss, one of Sagan’s most important teammates. This is further evidence that suggests racers tweet with an expectation that their audience knows who they are.

In sum, when constructing the results of a race on Twitter, both team leaders and supporting riders seem to use various textual, retweeting, and @mentioning strategies to construct the result within the sport’s value system as primarily the team’s and secondarily the individual racer’s, while emphasizing the value of cooperation among teammates who may be separate from the commercial entity of that team frame. In the interest of a final example that exemplifies these values, consider the following tweet by Daryl Impey.

42. darylimpey: What a day to be remembered on old Stirling. **Finally we bag this stage in @tourdownunder for @MitcheltonSCOTT . Our pocket rocket @CalebEwan winning it for us**. Stoked with 2nd as well (2018-01-17T09:20:51.000Z)

Impey finished this stage in second, which is a strong individual result. However, it appears in this tweet as almost an afterthought. Foremost is the result for the team, which he constructs using @mentions to explicitly position the result within the race, for his team, and relative to the more successful rider.

Conclusions

The aim of this study has been to demonstrate the use of TWIG in an analysis of how Twitter users from a particular offline community, bicycle racing, use Twitter’s native topic- and user-connective language features to construct identity online. This exploration has led to a number of observations regarding the kinds of identities bicycle racers construct and the mechanisms by which they do so. In terms of the behavior of the bicycle racing community itself, bicycle racers do not appear to be on Twitter expecting to be followed by the general public or to behave as commentators. They primarily construct themselves on Twitter as what they are—participants—and, as such, limit their reach for visibility to fans who seek them out and to associations within the bicycle racing community. In doing so, they rely heavily on specific uses of @mentions and retweets to frame themselves and their results within a primary identity as a racer for a specific team.

This analysis also identified some more general means of identity construction on Twitter that are likely unique to users tweeting within communities that are primarily offline. One thing that emerges in this regard is a weakened associative hashtag, because users in such communities are not necessarily interested in broader discourses and so may use fewer topical hashtags. In communities with organizational entities, @mentions seem to even replace hashtags, both in function and form. Retweeting organizational entities or key community members, too, proved to be an especially dynamic means with which users can position themselves and their activities within a community.

Furthermore, this study illustrates the value of several technical and theoretical decisions made in the development of TWIG. For example, TWIG's tools for collecting tweets, both in terms of its ability to reconstruct existing datasets and its framework for constructing queries for specific communities, even communities with subgroups of members. To duplicate this in a programming environment would have required a significant amount of coding to create dozens of unique Twitter API calls, apply post-hoc annotation of riders into teams and statuses within those teams, and to merge the dozens of separate responses into a single analyzable dataset. To recreate the subsequent analysis would have required even more programming work. The analysis above relied upon TWIG's utilization of metadata to locate contextual patterns in the dataset and the data navigation framework that allows frequent ad-hoc creation of narrowly filtered sets of tweets, such as tweets from riders from a specific team that use a specific @mention during a specific period of time. While creating such slices of the base corpus in a coding environment is possible, the fluidity with which TWIG allows them to be modified and navigated between cannot be recreated in a coding environment—or, indeed, in any other graphical analytical environments. This fluidity is particularly valuable in analyses of contexts (rather than specific linguistic forms)—such as the behaviors of communities or individual users, or the discourse surrounding a hashtag—where finding analytically-salient patterns or data points can require significant time learning and exploring the full dataset.

CHAPTER IV

WHAT *IT* IS AND WHAT IT *REALLY DO BE LIKE* (SOMETIMES): CHUNKING MEMES

While the previous chapter demonstrated TWIG's viability in performing a relatively open-ended analysis of how a specific group of users construct identity and community values on Twitter, it did little to demonstrate TWIG's usefulness for performing analyses of specific language features. To which end, this study offers a usage-based analysis of the construction "really do be" after the emergence and popularity of the internet meme expression "it really do be like that sometimes."

Internet memes have become a pervasive feature of online discourse, to the extent that their being called "memes" is a product of the speed at which they seem to propagate, having its origin in Dawkins' (1976) notion of a meme as a sort of culturally self-replicating concept. Acknowledging this broader meaning, all further use of *meme* in this study refers to the internet variety, which typically have the form of a combination of an image and text. Memes have been previously analyzed as multimodal constructions by Dancygier and Vandelanotte (2017), who note that "memes can additionally give rise to new linguistic (monomodal) constructions appearing in more standard usage contexts such as journalism and advertising" (p. 567). The current study aims to expand on this observation by looking at a potentially-novel usage of the construction *really do be* among ordinary Twitter users, which I argue has emerged from the meaning constructed by the “it really do be like that sometimes” meme.

Chunking memes

From a usage-based perspective, linguistic structures are believed to emerge at all levels—from phonological systems to morphosyntax—as a product of repeated exposure to form-meaning pairs in particular contexts (Bybee, 2010; Ellis, 2001). At the level of morphology and syntax, dealing with multi-word expressions, form-meaning pairs tend to be discussed as “constructions” (Bybee, 2010). Central to the process by which constructions attain meaning is the notion of an exemplar, which is a single instance of a particular form with a particular meaning in a particular context to which a particular speaker is exposed. Essentially, meaning of a form emerges from a sort of cognitive probability model that considers meaning and context from a network of previously-observed exemplars of the form (Ellis, 2001). Crucially, this is fundamentally a model of acquisition, so each new exemplar influences future interpretation, and use, of the form (Bybee, 2010). Viewed more broadly than the individual language user, this process, spread among communities and populations, becomes a driver for linguistic change (Bybee, 2010, 2015; Moder, 2016).

Given the significance of exemplars with frequent form-meaning pairs, memes—again, named for the speed with which they spread—seem to represent a valuable avenue of inquiry for usage-based research, offering not only a bulk of readily-available data (especially with the help of tools like TWIG 😉), but also an element of multimodality that gives credence to the idea that usage-based approaches to language study do indeed deal with “domain-general cognitive processes” (Bybee, 2010). While Dancygier and Vandelanotte (2017) approach this multimodality by discussing the cognitive implications and networks of association involved in giving meaning to the memes themselves, within a community of people who produce and share memes, the current study looks instead at how these multimodally-constructed forms and meanings appear in monomodal discourse, used by a more general population.

I do this by observing the existence of a potentially-novel usage of the construction “really do be,” which I propose has risen as a result of a broad (non-AAE-speaking, as discussed in the following section) population chunking the construction’s use from the “it really do be like that sometimes” meme. Chunking is, basically, the process by which constructions are formed. When multiple smaller chunks (i.e. words) frequently co-occur, they can coalesce in the minds of speakers into a larger chunk, which can take on meaning that is not necessarily surmisable by consideration of the individual parts (Bybee, 2010; Ellis, 2001). The extent to which the meaning of chunks can be understood by the sum of its parts is referred to as its *compositionality*, such that a chunk can be said to lose compositionality as its meaning drifts from that of its parts (Bybee & Moder, 2017). This shift in meaning, as noted by Moder (2016), can be explained by a tendency in which “chunks that are used repeatedly within a given context can spread to a larger community of users, leading to wider, more entrenched changes in linguistic form and distribution” (p. 37).

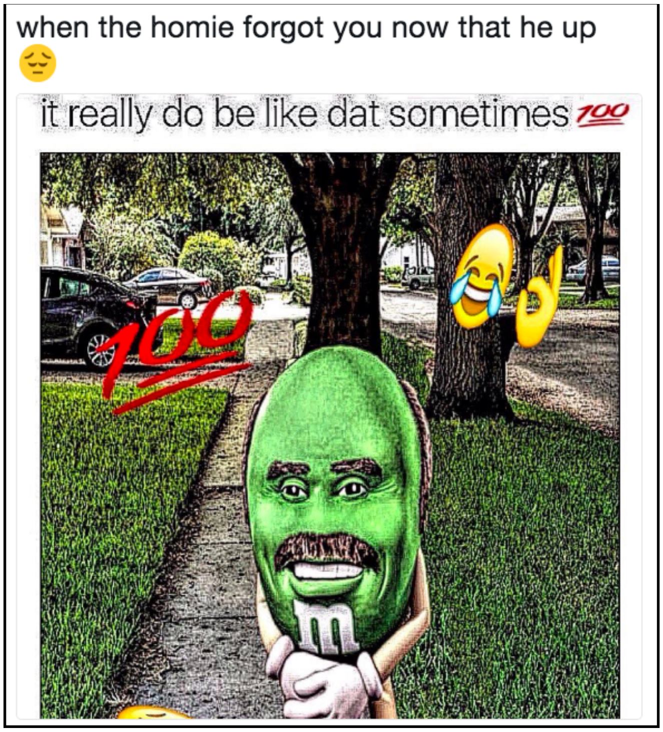
With this in mind, when I refer to the chunking of “really do be” from the meme, I’m referring not to “really do be” as a chunk of “it really do be like that sometimes,” but instead to the coalescence of “really,” “do,” and “be” into a sort of evaluative copula for which the specific evaluation is tied to that of the meme—and is not thus inferable from the meanings of any of the chunk’s parts, which on their own seem to suggest simply an emphatic copula. Broadly, the hope of this study is that looking at chunking effects in the language of memes will contribute to a usage-based understanding of language change by illustrating a case in which chunking a frequent multimodal construction seems to happen very rapidly. In the following sections, I will elaborate on the meaning constructed by the meme, describe the procedures that were taken in

TWIG to collect a corpus of relevant tweets, and, finally, present an analysis of current usage of “really do be” on Twitter.

It really do be like that sometimes

Beyond the scope of meme use, the expression “it really do be like that sometimes” is a grammatical AAE sentence that uses something like a do-strengthened habitual *be,* possibly with the “intensive quality” Labov (1998, p. 12) describes the construction as occasionally carrying. The expression’s meme use likely represents a kind of appropriated “mock” AAE used by non-authentic AAE speakers to construct a persona deemed suitable to the intended meaning (Ilbury, 2019; Smokoski, 2016). However, while the use of AAE features in this and other memes is itself worthy of analysis, the study is primarily concerned with the meaning of the meme form and how the chunking of that form, while arguably retaining its meaning, is evident in language use on Twitter even among users with no other AAE features.

The meme-use of “it really do be like that sometimes” has been traced to July 2017 by Knowyourmeme, a website that provides community-sourced usage histories of memes, which cites the meme in Figure 4.1 as the original meme use of the expression (Knowyourmeme a).

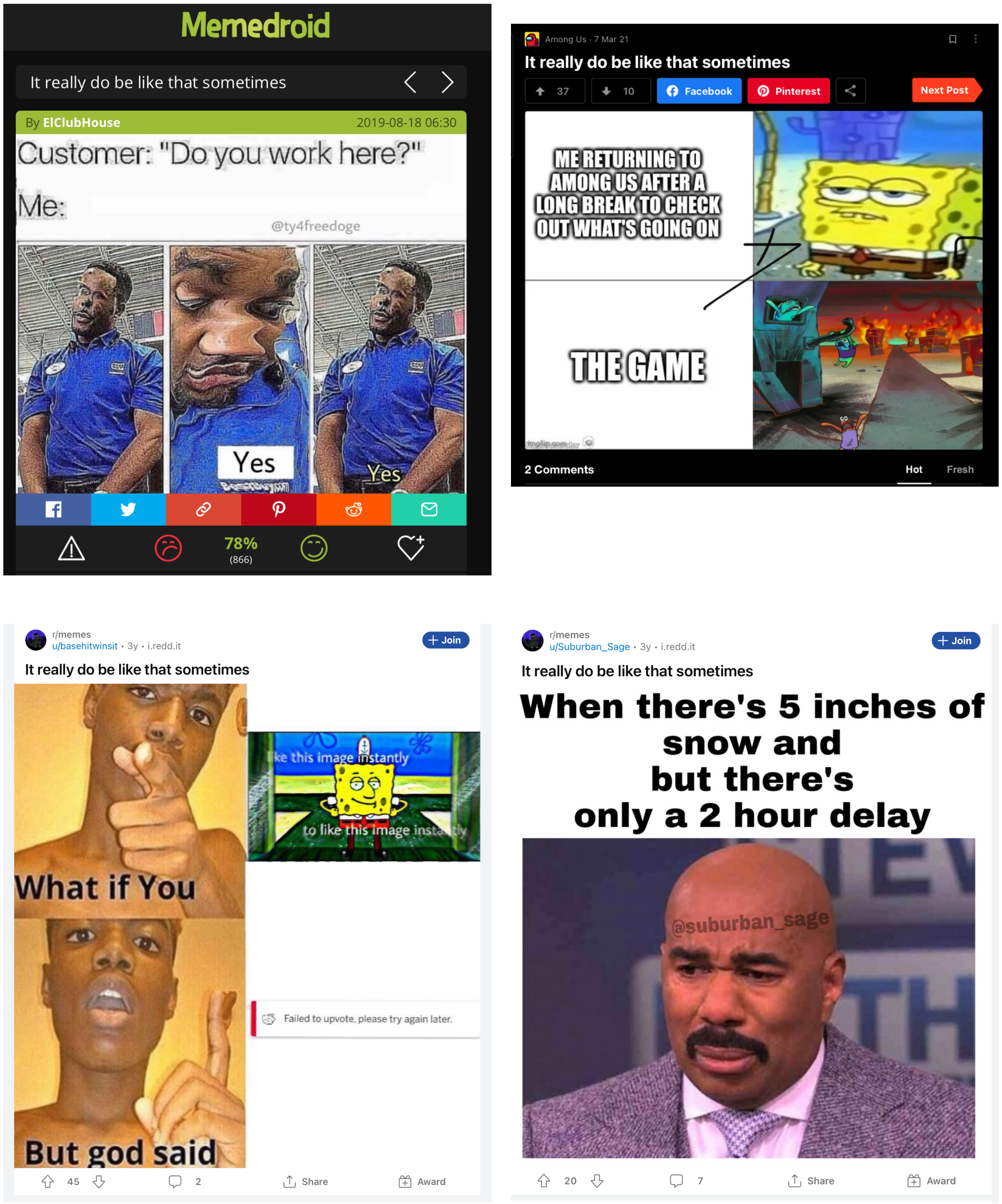


when the homie forgot you now that he up 😔

it really do be like dat sometimes 💯

Figure 4.1 - Early meme usage

In this meme, much of the meaning is created textually. “It really do be like dat sometimes” is presented as an evaluation of a situation in which a presumably now-former friend has found success and stopped associating with the “author” of the meme. The evaluation being offered is a kind of situational disappointment that represents a broader, continual existential disappointment. This sentiment is reenforced and colored by the image component of the meme, which includes an M&M representation of Dr. Phil, which emerged in an advertisement and was subsequently used in memes as a “disturbing” character (Knowyourmeme b). This emphasizes the discomfort of the situation, creating a disappointment that’s typical and expected as a result of some fundamental cruelty in how the world works. This is the general schema associated with meme use of the construction “it really do be like that sometimes,” which has since been used frequently, as further examples in figure 4.2 show, in subsequent memes expressing unsurprising situational/existential disappointment.



Memedroid

It really do be like that sometimes

By ElClubHouse 2019-08-18 06:30

78% (866)

Among Us · 7 Mar 21

It really do be like that sometimes

37 10 Facebook Pinterest Next Post

2 Comments Hot Fresh

r/memes
u/basehitwinsit · 3y · i.redd.it

It really do be like that sometimes

Failed to upvote, please try again later.

45 2 Share Award

r/memes
u/Suburban_Sage · 3y · i.redd.it

It really do be like that sometimes

20 7 Share Award

Customer: "Do you work here?"
Me:
@ty4freedoge
Yes
Yes
ME RETURNING TO AMONG US AFTER A LONG BREAK TO CHECK OUT WHAT'S GOING ON
THE GAME
What if You
But god said
to like this image instantly
When there's 5 inches of snow and but there's only a 2 hour delay
@suburban_sage
Join

Figure 4.2 - Additional meme-usage

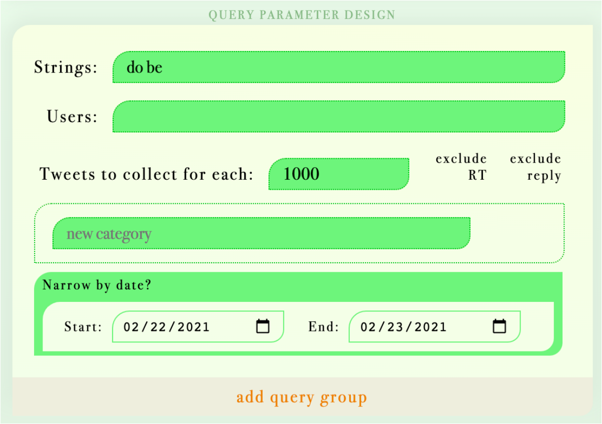
It is worth noting that in these examples, “It really do be like that sometimes” is not formally part of the meme, but rather appears as a title for the meme, which the users would have separately typed from the images they attached when posting on the various social media platforms on they elected to share the memes. This further cements the broader construction's use as an evaluation that can be applied to a variety of disappointing, but not surprising, situations.

Data collection

To explore this meme construction's possible effect on more general language use, a corpus of 23,853 tweets was collected containing the string "do be." Twitter is a suitable source of data for study for a variety of reasons. Twitter is not only generally regarded as a valuable source of natural language data (Page et al., 2014), but is particularly valuable for a study involving the language of memes, being itself a frequent site for the distribution thereof—including the original location of the meme central to this study.

The query "do be" was selected rather than "really do be" because it allows the tracking of both constructions, particularly the relative rise of "really do be" over time. To track the construction over time, 1000 tweets were collected for every six months beginning Feb 22, 2010 and ending Aug 23, 2021. This time period was selected to include the time period of and after the circa 2010 meme "they don't think it be like it is, but it do," which is often discussed in relation to "it really do be like that sometimes" (knowyourmeme 3). This did not prove to be particularly salient, but the extra data was useful in that it provided additional pre-2017 "really do be" examples.

The construction of this corpus in TWIG was completed by creating a project for the study and, in the Project Design Environment, defining 24 temporally-distinct query groups set with "de be" as a string query and a collection size of 1000 tweets, as shown in Figure 4.1.



Strings: do be

Users:

Tweets to collect for each: 1000

Narrow by date?

Start: 02/22/2021 End: 02/23/2021

Figure 4.3 - Query design prototype

Sampling 1000 tweets every 6 months was preferable to simply defining one query with a 12-year collection period for largely practical reasons. The Twitter API collects tweets in reverse-chronological order, and based on the collected corpus size and the amount of time represented, collecting enough tweets to represent the full 12-year time period would require collecting more than 7 million tweets. This would be greatly increase both data collection and processing time with no likely benefit to the kind of analysis performed in this study. The total corpus size not being exactly 24,000 tweets (an even 1000 per sample) is a product of the Twitter API's never returning exactly the requested totals, but I believe the variance represents a sufficiently balanced corpus for this study's purpose.

Really do be

Before discussing specific uses of "really do be" in this dataset, it is useful to provide a degree of quantitative context for the current study. The goal in doing so is not to provide a statistical argument, but rather to illustrate that the frequencies with which the constructions "do be" and "really do be" are used on Twitter change over time and that certain of these changes coincide in ways that seem salient with the 2017 "it really do be like that" meme. These changes are shown in Figures 4.4 and 4.5 below. These figures were plotted in R using numbers identified in TWIG by narrowing the dataset for each year one at a time, and copying the statistics (including standard deviation) from above the ten-minute interval frequency tables, as shown in Figure 4.6. The ability to track temporal metadata in this way illustrates the value of both TWIG's inclusion of statistical measures of variance and dispersion and the way in which those measures are visualized.

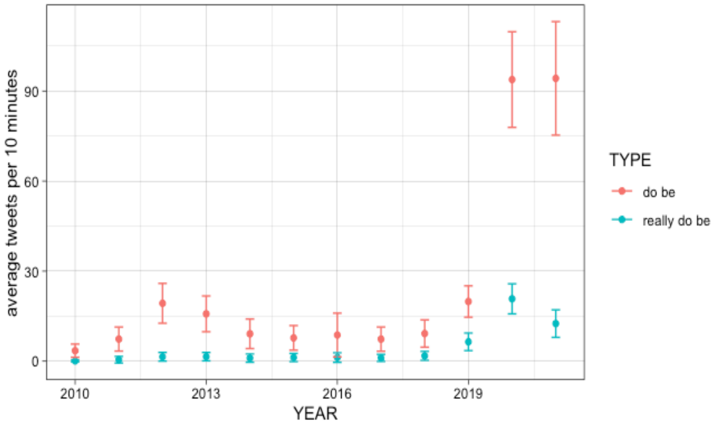


Figure 4.4 - Rates of "do be" and "really do be" use

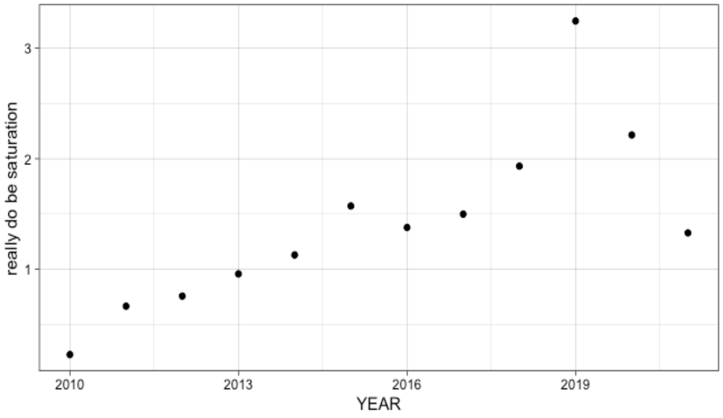
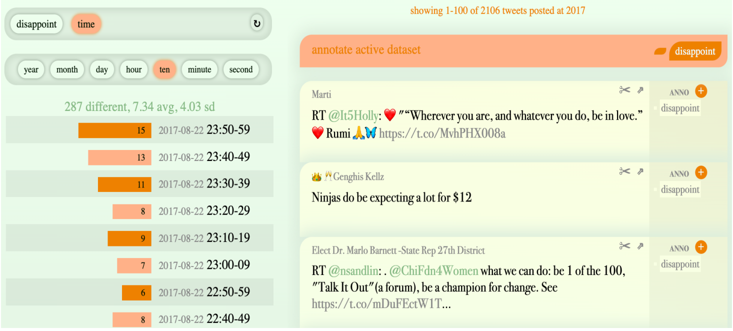


Figure 4.5 - Distribution of "really do be" / "do be"



showing 1-100 of 2106 tweets posted at 2017

287 different, 7.34 avg, 4.03 sd

15	2017-08-22 23:50-59
13	2017-08-22 23:40-49
11	2017-08-22 23:30-39
8	2017-08-22 23:20-29
9	2017-08-22 23:10-19
7	2017-08-22 23:00-09
6	2017-08-22 22:50-59
8	2017-08-22 22:40-49

Marti

RT @It5Holly: ❤️ "“Wherever you are, and whatever you do, be in love.” ❤️ Rumi 🙏🦋 https://t.co/MvhPHX008a

👑🥂Genghis Kellz

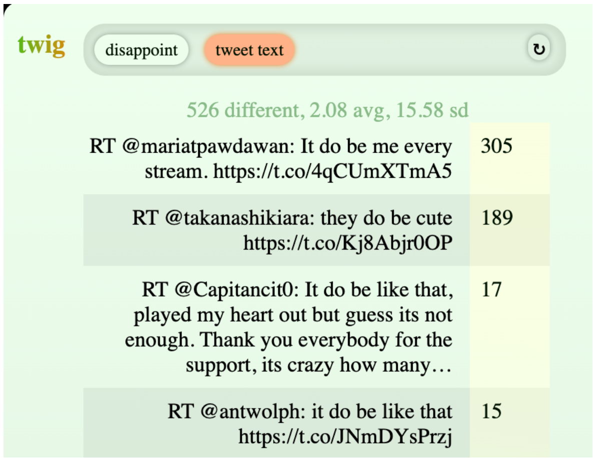
Ninjas do be expecting a lot for $12

Elect Dr. Marlo Barnett -State Rep 27th District

RT @nsandlin: . @ChiFdn4Women what we can do: be 1 of the 100, "Talk It Out" (a forum), be a champion for change. See https://t.co/mDuFEctW1T...

Figure 4.6 - “do be” 10-minute statistics for 2017

Figure 4.4 shows the rate of tweeting as tweets per ten-minutes for the years represented in this study’s corpus. While the most glaring finding on this figure is a rapid rise in “do be” posting rate for the years 2020 and 2021, it is worth noting that this jump was preceded by an upward trend beginning in 2017 and that, unlike with the 2012-2013 bump, this rise coincides with a rise in “really do be” usage. Figure 4.5 examines this trend by tracking “really do be” use as a percentage of “do be use.” Note how before 2013 “really do be” accounted for less than 10 percent of “do be” use, and since 2016 it’s been as high as around 33 percent. Note also that the dip in 2021 is skewed as a product of two bare “do be” tweets with a combined retweet count of 494. These retweets were identified in TWIG by looking at “tweet text” metadata, as shown in Figure 4.7 below. Overall, however, the numbers suggest that people on Twitter use “really do be” more than they used to, which suggests that the chunk might be becoming more fixed and that it might provide a meaning for which there is a need.



526 different, 2.08 avg, 15.58 sd

RT @mariatpawdawan: It do be me every stream. https://t.co/4qCUmXTmA5	305
RT @takanashikiara: they do be cute https://t.co/Kj8Abjr0OP	189
RT @Capitancit0: It do be like that, played my heart out but guess its not enough. Thank you everybody for the support, its crazy how many…	17
RT @antwolph: it do be like that https://t.co/JNmDYsPrzj	15

Figure 4.7 - Identification of frequent retweets in TWIG

Existing usage

As discussed above, “really do be” is a grammatical AAVE construction. It existed before the meme, and, as such, has uses before the meme, which persist. These uses seem largely compositional, using *really do* to emphasize (as shown in examples 1-3 below) or to counter a hypothetical counterfactual proposition (examples 4-6).

Emphasis

1. _platinumLEXISS: OMG I hate when hella people over! It **really do be** irritating me, bcos they hella loud like they can't hear eachother. (2010-08-22T23:38:41.000Z)
2. AshlynManghane: people **really do be** testing my patience . i HAVE to be a string person to put up with al the nonsense in my life . (2010-08-22T02:07:34.000Z)
3. LonnieVee_: Damn, I **really do be** fuked up on the low (2011-08-22T17:58:15.000Z)

Counter-counterfactual

4. HtownSprizzy: On the freeway and it hits me that I **really do be** busy as hell. Always on the move, hopefully it gets me somewhere (2010-08-22T13:24:14.000Z)
5. Official_jayy: @_chaianna omggggg. no. haha it was a joke for the guys **that really DO be** taking it that serious. The ones with big egos. lol (2011-08-22T09:26:04.000Z)
6. _aChade: Which is where when I say “idc” I **really do be** meaning it sometimes .. (2014-02-22T06:25:52.000Z)

It is worth noting that some of these examples contain additional AAE features, such as the null copula in example 1 ("hella people over" and "they hella loud") and the use of "on the low" in example 3. This suggests that these users might be authentic AAE speakers, in likely contrast to many of the "really do be" users presented below. This significance of this contrast is the spread of "really do be" to speakers of other dialects. Another interesting aspect of example 3 is that this use does seem to carry a connotation of disappointment, three years before the meme, suggesting the meme use might have actually popularized an existing but infrequent use.

When it do be like that

In the meme, *it really do be like that* is the full evaluative expression. *It* is the situation being evaluated, the situation that is textually or visually represented in the meme, and the *that* that it *really do be like* is some fundamental nature of existence. So while I have thus far been discussing the chunk "really do be," which (accurately) suggests a primary interest in uses in which that copular chunk is used with more specific subjects and predicates, the full meme is chunkable. This study's corpus contains a large number of tweets that use the full expression either verbatim or switch the position of *sometimes*, as shown in examples 7 and 8 below.

7. gameshed_: i was called braindead and autistic multiple times for saying i don't mind the new additions that much, i genuinely despise that shithole of a subreddit
   ITalkFortnite: I'm sorry to hear that.
   gameshed_: @ITalkFortnite **it really do be like that sometimes** (2019-08-22T21:31:34.000Z)
8. ajpOneThree: I really took a 3 minute nap in my car before walking in to work this morning. Set a whole alarm and everything
   Powerflare: @ajpOneThree **Sometimes it really do be like that** (2019-08-22T22:29:08.000Z)

These examples are shown in interaction, which makes sense, because it aligns with the structure of the original meme, in which *it really do be like that sometimes* is attached to an otherwise developed situation. This use is naturally mirrored in interaction. The replied-to tweet becomes a situation for the replier's evaluation, and I believe the evaluation maintains the connotation of situationally-expected disappointment constructed by the meme. In example 7, cyberbullying is evaluated as inevitable because of the essential nature of the "shithole of a subreddit" (a subreddit is a specific community on the social-media platform Reddit). In example 8, being exhausted enough to need a nap before work is an unfortunate feature of existence. It is worth also noting that neither of these interactions include AAE use other than that from the meme.

In terms of form, examples 7 and 8 show relatively fixed application of the original full expression, with the movement of *sometimes* in example 8 likely being a product of the user analyzing the chunk as a grammatical entity and recognizing a potential for movement. While such verbatim uses of the real expression are frequent, the chunk is ultimately not very fixed, allowing, as examples 9, 10, and 11 show below, for the deletion of *really, it,* and *like that*.

9. ViRosefall: I was going to tweet something smart, but then I was reminded Im a dumb donkey. Darn.
   ReikaShirogane: @ViRosefall **It do be like that sometimes** (2021-08-22T23:52:32.000Z)

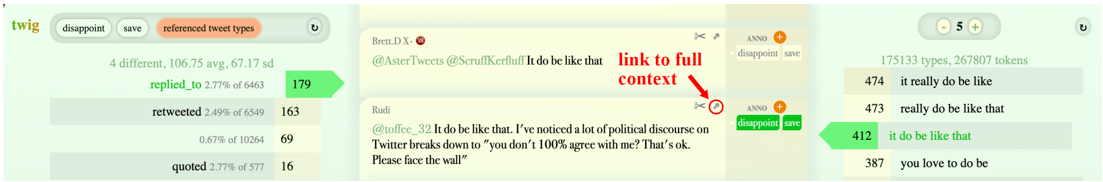
10. ConnorGilgallon: "Was George Washington born in Vermont?"
    EHernan11: Bro obviously he was born in Washington 🤦🏽‍♂️
    ConnorGilgallon: Need his REAL birth certificate
    [deleted tweet]
    ConnorGilgallon: @kdonne14 @EHernan11 **Putin really do be like that sometimes** (2018-08-22T18:52:33.000Z)

11. *[a British wrestling organization tweeting a link to their upcoming matches]*
    nickl104: **It really do be that hard not to book abusers**, apparently https://t.co/F7zAqzZWm9 (2021-02-22T22:15:06.000Z)

Examples 10 and 11 are particularly interesting because an analysis of the grammar of the chunk has located subject and complement positions occupied by the unnecessary *it* and *like that*, so the chunk has lost words but not a grammatical positions in which to place other words. What remains is a construction with two slots, with the effect of creating a copularized chunk. Again, I maintain that these examples retain meaning from the meme, regarding the disappointing inevitabilities involved in (9) not being able to say something smart, (10) a global political landscape that involves Vladmir Putin, and (11) the number of abusers apparently involved in professional wrestling.

Locating examples in TWIG

Before discussing the copularized *really do be*, let's pause to look at how the above examples were identified in TWIG. They were not found by an unstructured reading of the full 23,853 tweets in the corpus. This would create an unnecessarily grueling and (for me, at least) disorganized workflow. Knowing the kinds of tweet I was looking for—those used in interaction and those that contain various combinations of words from the full expression—filters from TWIG's metadata and text analysis tools were leveraged to isolate such tweets. Tweets used in interaction were isolated by selecting "replied to" from the "referenced tweet types" metadata frequency table, as shown in Figure 4.8 below. To view these replies in context, the external link on the top right of each tweet shows the tweet on Twitter, within the reply chain of the tweet to which it is a reply, as shown in Figure 4.9.



4 different, 106.75 avg, 67.17 sd

replied_to 2.77% of 6463	179
retweeted 2.49% of 6549	163
0.67% of 10264	69
quoted 2.77% of 577	16

Brett.D X·

@AsterTweets @ScruffKerfluff It do be like that

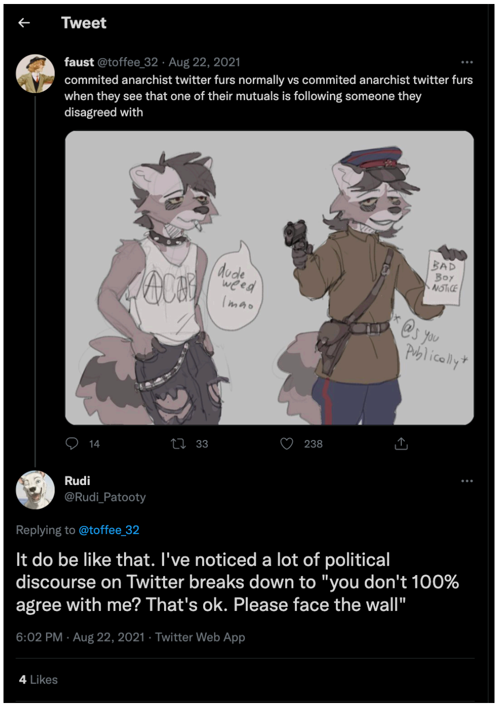
Rudi

@toffee_32 It do be like that. I've noticed a lot of political discourse on Twitter breaks down to "you don't 100% agree with me? That's ok. Please face the wall"

175133 types, 267807 tokens

474	it really do be like
473	really do be like that
412	it do be like that
387	you love to do be

Figure 4.8 - Locating tweets in interaction



Tweet

faust @toffee_32 · Aug 22, 2021

commited anarchist twitter furs normally vs commited anarchist twitter furs when they see that one of their mutuals is following someone they disagreed with

14 33 238

Rudi
@Rudi_Patooty

Replying to @toffee_32

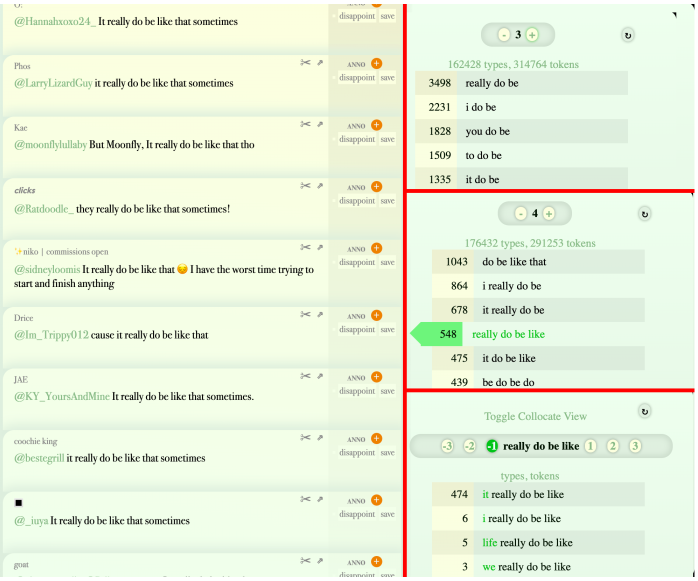
It do be like that. I've noticed a lot of political discourse on Twitter breaks down to "you don't 100% agree with me? That's ok. Please face the wall"

6:02 PM · Aug 22, 2021 · Twitter Web App

4 Likes

Figure 4.9 - The tweet from 5.5 in context

Tweets containing various chunked representations of *it really do be like that sometimes* were found using the Text Analysis Pane to find relevant n-grams and collocates thereof, as shown in Figure 4.10 below.



@Hannahxoxo24_ It really do be like that sometimes

Phos
@LarryLizardGuy it really do be like that sometimes

Kae
@moonflylullaby But Moonfly, It really do be like that tho

clicks
@Ratdoodle_ they really do be like that sometimes!

✨niko | commissions open
@sidneyloomis It really do be like that 😔 I have the worst time trying to start and finish anything

Drice
@Im_Trippy012 cause it really do be like that

JAE
@KY_YoursAndMine It really do be like that sometimes.

coochie king
@bestegrill it really do be like that sometimes

■
@_iuya It really do be like that sometimes

goat

- 3 +

162428 types, 314764 tokens

3498	really do be
2231	i do be
1828	you do be
1509	to do be
1335	it do be

- 4 +

176432 types, 291253 tokens

1043	do be like that
864	i really do be
678	it really do be
548	really do be like
475	it do be like
439	be do be do

Toggle Collocate View

-3 -2 -1 **really do be like** 1 2 3

types, tokens

474	it really do be like
6	i really do be like
5	life really do be like
3	we really do be like

Figure 4.10 - Locating different chunks

This figure shows one process of locating _ *really do be like* constructions that do not begin with "it." First, the n-gram size was increased to four, then "really do be like" was selected, then the collocate view was enabled and set to show preceding words. Similar versions of this process were employed to find the other constructions shown above.

The copularized *really do be*

While there might be an argument for a *do be* chunk carrying the disappointed evaluative meaning, I believe the *really do be* form to be much stronger. Even if it's not, what's interesting is that copularization of the chunk enables specific things to be evaluated in domain-specific ways, as shown in examples 11 and 12 below, which are, as above, in interaction.

12. Lahumi__: Why would you willingly put cow's milk in your coffee when oat milk exists
    ElegantiaeArbit: I actually like the taste but oat milk normally contains some type of oil (canola, sunflower) that I try to avoid
    Lahumi__: Oh I didn't know about this - jzk for highlighting, need to look into it!
    ElegantiaeArbit: @Lahumi__ Seed oils **really do be** in everything 😥 (2021-08-22T23:19:20.000Z)

13. chenicew: Why are bags of chips so loud?? Like, now everyone knows I got snacks in my bag 🙄
    AriesTrill: @chenicew They **really do be** trying to let the world know you got chips 🤣🤣 (2021-08-22T22:50:23.000Z)

In these examples, the scope of the disappointment is restricted. The "life is like that" sentiment of the full expression is replaced with much more concrete disturbing but inevitable situations: seed oils are in everything, and chips bags are loud. But these are still evaluative statements, and they both express disappointment without surprise in a way that I believe is more similar to examples 7-10 above than to examples 1-6. This reading of the *really do be* use in examples 12 and 13 as evaluative rather than simply emphatic is bolstered by the emoji use in the tweets, which represent crying and laughing, further suggesting that the tweets have evaluative intents.

In terms of form, the ability of the copular chunk to include concrete situations and domain-specific evaluations further allows the construction to be presented without an outside situation to evaluate. The situation and evaluation components that are discreet in the meme and in all of the other chunkings of the meme are merged. This enables usage out of interaction, that functions as intended in standalone tweets, as the following examples show.

14. amandakerril: live photos **really do be** catching the worst mom voice hahaha (2021-08-22T23:57:18.000Z)
15. enriquiem: They **really do be** making musicals about anything these days. Can't wait for Greggs on the West End (2021-08-22T23:31:06.000Z)
16. HONEYMILK2TASTE: they **really do be** playing elton john in maccas (2020-02-22T23:54:14.000Z)

As in examples 12 and 13 above, these examples—while containing the fewest elements of the meme expression, losing even the externally-initiated situations of 12 and 13—seem to maintain the use of *really do be* to evaluate, not simply to emphasize. This claim is admittedly complicated by the fact that in these examples *really do be* is followed by *ing* forms, which is more clear-cut AAE habitual *be* use than is present even in the meme form. The meaning of the habitual *be* is retained as well, because these examples all represent recurring situations. I believe however, that these uses do maintain an evaluative meaning that is not present in all habitual *be* use. The meaning conveyed by *really do be* in 15, for example, is more than an emphasized version of the *be* in the traditional "he be running" example of habitual *be.* Example 15 presents a negative evaluation of the fact that musicals are being made about "anything." This evaluation is emphasized by the tweet's second sentence, which sarcastically anticipates a musical based on Greggs, a British bakery chain. The mention of Greggs is interesting in itself. None of the three examples above include any AAE features other than *really do be*, and 15 and 16 include references that suggest, respectively, British and Australian users. This represents the spread of a form that in 2010 was mostly used by what seem to be authentic AAE speakers to, by 2021, speakers in English-speaking countries other than the United States.

Conclusions

As noted above, it is not the position of this paper that the novel usage of *really do be* I've described has become a dominant use of the construction, nor even that it will be a lasting use. The construction is, however, definitely being used, at an elevated rate, and apparently within multiple communities that might not typically use AAE features. Regarding the construction's semantic relationship to the "it really do be like that sometimes" meme, I believe the variations shown in examples 7-11 are sufficient to illustrate that the expression is being chunked and various additional features of the presented tweets' text suggest an evaluative connotation in line with the construction of the meme form. These arguments would likely be strengthened, however, by a subsequent, more fine-grained tracking of the various chunked forms. Nevertheless, I believe the above analysis contributes illustrations of how the chunking of a multimodal construction can affect change in monomodal language use and spread language features typically associated with a specific dialect to a broader community of users.

More broadly, this study offers a demonstration of TWIG's effectiveness in collecting and navigating datasets for studies interested in specific constructions. TWIG's real-time filtering capabilities in the analysis environment were particularly valuable for a usage-based study, for which the metadata analysis functionality allowed me to isolate tweets used in interaction and the text analysis functions allowed me to locate tweets that contain variations of an unfixed target chunk. The metadata-framed dispersion statistics TWIG provides proved valuable in roughly tracking the relative distribution of related forms. However, this analysis would have likely benefitted from more fine-grained tracking of emergent forms than TWIG allows for. This is to say that, while TWIG allows individual the tracking of the complete "it really do be like that sometimes," the copular "really do be" form, and those forms in-between, to look at the relationship between these forms over time would require manually recording statistics for each chunked form and entering them in an external plotting software. This represents an avenue for future development, though such multivariate plotting would likely require a separate analytical environment for visualizations.

CHAPTER V

CONCLUSIONS

The broad goal of this project was to develop a tool with which language and media researchers, particularly discourse analysts, can, without any coding, collect and analyze data from Twitter in a way that meets the needs of a variety of Twitter-centric studies. TWIG, as demonstrated in Chapters II-IV of this document, meets this objective. This alone, I believe, fills a clear technological gap in more than one research area.

However, the ultimate aim of TWIG is more complex than technological accessibility. The original motivation for the development of what became TWIG's analysis environment was personal dissatisfaction with the technical capabilities of existing corpus and discourse tools and with the fragmented workflow of conducting exploratory analysis within coding environments. To these ends, the goal of TWIG was to provide an environment for analyzing datasets of Twitter data that would be of use to even researchers with extensive coding experience. Most basically in this regard, TWIG provides a level of metadata access that is not available in other standalone corpus tools—with a level of immediacy that is not available in any coding environment. From a technical perspective, TWIG does in just a few clicks what would take dozens of lines of code in a coding environment. Each time a user applies a text or metadata filter to the active dataset, this represents the assembly of at least one frequency table, the extraction from a dataset of all matching texts, and the assembly of at least one updated frequency table, along with statistics representing means and variance for the values in the active metadata frequency table.

From a theoretical perspective, TWIG's handling of metadata enables researchers who lack coding experience to perform analyses that meaningfully include all of the data that API-collected and user-annotated datasets contain, much of which is important discourse-contextual information. This enables the conducting of studies in which extratextual aspects of texts can be analyzed along with the textual content. In this document's sample studies, TWIG's metadata functions were specifically used to isolate and compare bicycle racers from different teams and with different roles within those teams in Chapter III, to isolate tweets created in interaction in Chapter IV, and to facilitate the temporal aspects of both studies. Similar approaches can be used to address research questions regarding such phenomena as methods of interaction, affiliation types, identity-construction within and between communities; temporal and interactive factors in how, when, and to what effect specific constructions are used.

TWIG's analysis environment was also designed to address several theoretical concerns regarding both the use of corpus methods in discourse analysis and the general calculation and representation of corpus frequency statistics. At the center of these theoretical concerns is the idea that analyses of corpus data should not, at a quantitative level, describe simply the corpus itself, but rather the corpus as a representation of the texts it contains. Egburt and Schur (2018) suggest that part of why such problematic methodologies prevail is that existing analytical software is designed with corpus-level, not text-level, analysis in mind. In this regard, TWIG contributes to the fields of both corpus and discourse analysis by providing an analytical environment designed from the ground up for analyzing corpora as representations of their constituent texts. Briefly, it does this by providing frequencies at the text level, providing descriptive statistics, when applicable, with measures of dispersion variance, and calculating these statistics based on a variety of metadata-based sampling frames—e.g. by tweet author or temporal range. The fact that TWIG provides these statistics often provides the option of a quantitative grounding to qualitative claims, such is the case in the study presented in Chapter IV, in which the study's semantic claims coincide with changes in temporal frequencies related to the construction. Beyond this, however, the value of the statistics TWIG provides is not in the kinds of study it enables, but rather the more accurate picture that is provided by these more involved and informative methods. Put simply, TWIG provides statistical measures that can help researchers qualify discussions of frequency and avoid misrepresenting their data.

In addition to these contributions, I am compelled to assert that the effectiveness of TWIG as an analytical tool emerges as more than the sum of the technical accessibility it provides and the theoretically-conscious approaches to metadata and quantification it enables. While in one sense TWIG can be boiled down to about 4,500 lines of code, it has emerged as a product of concurrent development and use—its development driven by my using it to conduct research, namely, at this point, the studies presented in Chapters III and IV, and by discussions with peers regarding Twitter-based studies they are interested in. This symbiotic process has led to the targeted inclusion of features in search of not only theoretically-driven functionality, but an efficient analytical workflow—a search that will continue as development continues beyond what is described in this document.

Moving forward

The future of TWIG will involve two main avenues of work: continued development of features and a public release. Regarding feature development, discussions with peers suggest that the most valuable next steps will be to enable a greater degree of project customization. For example, while TWIG by default calculates text frequency tables at the text level, showing how many tweets contain each construction, some researchers might prefer these frequencies be corpus level, showing the raw frequencies of each construction. Allowing this to be configurable in the Project Design Environment would be relatively simple. Similarly, for keyness analyses, researchers might prefer to have more control over the reference corpus, especially given the downsides of the default BNC corpus discussed above. To this end, development has already begun to allow researchers to select other reference corpora—or even upload their own. A final option that has emerged as valuable to analytical workflow is the ability to configure what will be copied to the clipboard when the copy button on individual tweets is clicked. This defaults to include the username, tweet text, and timestamp for the selected tweet, but username and timestamp might not be salient to all studies, where other information, such as annotations or like count, might. Having this configurable for each project could save time and burden for researchers collecting examples as they write.

Other areas in which TWIG's development are likely to continue are larger in scale. The conclusion to the second sample study (Chapter IV here) illustrated the potential value for visualizing more complex, multivariate statistical information that TWIG's real-time exploratory analysis environment can provide. Furthermore, TWIG currently lacks a framework for collaboration among researchers. A small way in which this can be facilitated would be the ability to export the kind of CSV—with tweet IDs and researcher-provided metadata—that TWIG currently has the ability to import and use in data collection. This would enable the sharing of datasets in a way that is within Twitter's developer user agreements.

While these plans for development represent significant change, TWIG is already, as demonstrated above, an incredibly functional and usable tool, and, as such, the most significant next step is beginning to make it available to other researchers. Before this can happen, certain formalities need to be addressed, namely the drafting of a privacy policy to assure researchers of the security of their data, an agreement that TWIG is not responsible for lost data in the case of a server or software error, provided alongside a suggestion to use TWIG's various data download options to save local backups, and finally a warning to not use TWIG to break Twitter's developer agreement. These can all be integrated into the registration process. Once these things are in place, the work of promotion, getting users, and continuing based on their feedback can begin via publication and presentation in media-focused discourse and corpus journals and conferences, particularly those with a focus on social media.

REFERENCES

Anthony, L., (2018). Introducing Fireant: A Freeware, Multiplatform Social Media Data-Analysis Tool. IEEE Transactions on Professional Communication, 61(4), 428-442.

Auxier, B. & Anderson, M. (2021). Social media use in 2021. Pew Research. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

boyd, d. Golder, S. Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *Proceedings of the 43rd Hawaii International Conference on System Sciences (pp. 1-10)*. IEEE. doi:10.1109/HICSS.2010412

Brown, G. & Yule, G. (1983). *Discourse analysis*. Cambridge University Press.

Bucholtz, M. & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies* 7, 585–614.

Burr, V. (1995). *An introduction to social constructionism.* London: Routledge.

Bybee, J. (2010). *Language, usage and cognition.* Cambridge: Cambridge University Press.

Bybee, J. (2015). *Language change.* Cambridge: Cambridge University Press.

Bybee, J., & Moder, C. (2017). Chunking and Changes in Compositionality in Context. In M. Hundt, S. Mollin, & S. Pfenninger (Eds.), *The Changing English Language: Psycholinguistic Perspectives* (Studies in English Language, pp. 148-170). Cambridge: Cambridge University Press. doi:10.1017/9781316091746.007

Dancygier, B. & Vandelanotte, L. (2017). Internet memes as multimodal constructions. *Cognitive Linguistics 28:3,* 565-598.

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Davies, M. (2015). The Wikipedia Corpus: 4.6 million articles, 1.9 billion words. Adapted from Wikipedia. Available online at https://www.english-corpora.org/wiki/.

Egburt, J., & Schur, E. (2018). The role of the text in corpus and discourse analysis: Missing the trees for the forest. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical view* (pp. 159-173). Routledge.

Ellis, N. (2003). Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. In *The Handbook of Second Language Acquisition* (eds C.J. Doughty and M.H. Long). doi:10.1002/9780470756492.ch4

Ellis, N. (2017). Chunking in Language Usage, Learning and Change: I Don’t Know. In M. Hundt, S. Mollin, & S. Pfenninger (Eds.), *The Changing English Language: Psycholinguistic Perspectives* (Studies in English Language, pp. 113-147). Cambridge: Cambridge University Press. doi:10.1017/9781316091746.006

Hemsley, J., Ceskavich, B., & Tanupabrungsun, S. (2014). Syracuse social media collection toolkit (Version 0.1). Retrieved from https://github.com/jhemsley/Syr-SM- Collection-Toolkit

Hemsley, J. Jackson, S., Tanupabrungsun, S. & Ceskavich, B. (2019). bitslabsyr/stack: STACKS 3.1 (Version 3.1). http://doi.org/10.5281/zenodo.2638848

Hemsley, J., Stromer-Galley, J., Semaan, B & Tanupabrungsun, S. (2018). Tweeting to the Target: Candidates 'Use of Strategic Messages and @Mentions on Twitter. *Journal of Information Technology & Politics 15(1)*, 3-18.

Huyssen, A. (2000). Present pasts: Media, politics, amnesia. *Public Culture* 12(1), 21–38.

Ilbury, C. (2019). "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics* (24)*,* 245–264.

Iveson, M. (2017). Gendered dimensions of Catalan nationalism and identity construction on Twitter. *Discourse & Communication 11(1),* 51–68.

Jutel, A. (2002). Olympic road cycling and national identity: where is Germany? *Journal of Sport and Social Issues 26*(2), 195-208.

Knowyourmeme a. It really do be like that sometimes. https://knowyourmeme.com/memes/it-really-do-be-like-that-sometimes

Knowyourmeme b. Dr. Phil M&M. https://knowyourmeme.com/memes/dr-phil-mm

Knowyourmeme c. They don't think it be like it is, but it do. https://knowyourmeme.com/memes/they-dont-think-it-be-like-it-is-but-it-do

Labov, W. (1998). Coexistent systems in African-American English. In *Africam-American English: Structure, history and use* (eds Bailey, G.,Baugh, J., Mufwene, S.S., and Rickford, J.R.)

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media & Society, 14(5), 729–747. https://doi.org/10.1177/1461444811422894

Leech, G., Rayson, P., Wilson, A. (2001). Frequency lists. Word Frequencies in Written and Spoken English: based on the British National Corpus. Companion website: http://ucrel.lancs.ac.uk/bncfreq/

Murthy, D. (2012). Towards a Sociological Understanding of Social Media: Theorizing Twitter. Sociology 46(6), 1059-1073.

Page, R. (2012). The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. Discourse \& Communication 6(2), 181-201.

Page, R. & D. Barton, J.W. Unger, M. Zappavigna. (2014). Researching Language and Social Media: A Student Guide. Routledge: New York.

Parker, I. (1990). Discourse: Definitions and contradictions. Philosophical Psychology, 3(2/3), 189. https://doi-org.argo.library.okstate.edu/10.1080/09515089008572998

Marwick and boyd (2011). To See and Be Seen: Celebrity Practice on Twitter. *Convergence: The International Journal of Research into New Media Technologies 17(2),* 139–158

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. Research & Politics. https://doi.org/10.1177/2053168017720008

Moder, C.L. (2016). Begging the question: chunking, compositionality and language change. *European Journal of English Studies* (20:1), 35-46, doi: 10.1080/13825577.2015.1136161

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2011). Understanding the Demographics of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 554-557. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14168

Murthy, D. (2012). Towards a Sociological Understanding of Social Media: Theorizing Twitter. *Sociology 46(6),* 1059-1073.

Page, R. (2012). The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & Communication 6(2)*, 181-201.

Quin, Liam. (2016). Extensible Markup Language (XML). World Wide Web Consortium. https://www.w3.org/XML/

Scott, Kate (2015). The pragmatics of hashtags: Inference and conversational style on Twitter. *Journal of Pragmatics 81*, 8-20

Seargeant, P. & C. Tagg (eds.). (2014). *The Language of Social Media: Identity and Community on the Internet*. Palgrave Macmillan: New York.

Smokoski, H.L. (2016). Voicing the other: mock AAVE on social media. Masters thesis, City University of New York.

Stricker. G. (2014). The 2014 #yearontwitter. Twitter Blog. https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html

Toffoletti, K. & Thorpe, H. (2018). Female Athletes' self-representation on social media: a feminist analysis of neoliberal marketing strategies in "economies of visibility". Feminism \& Psychology 28(1), 11-31.

Torness & Tujillo. (2021). Enabling the future of academic research with the Twitter API. Twitter developer blog. https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.html\

Tweepy. (2017). Streaming With Tweepy -- Tweepy 3.5.0 Documentation. Tweepy.readthedocs.io.

Twitter. (2022a). Academic research: Preparing for the application. https://developer.twitter.com/en/products/twitter-api/academic-research/application-info

Twitter. (2020a). Data dictionary: Standard v1.1 -- Tweet Object. Docs: Twitter developer platform. https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

Twitter. (2021a). Data dictionary: Tweet. Docs: Twitter developer platform. https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet

Twitter. (2021b) Data dictionary: User. Docs: Twitter developer platform. https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/user

Twitter. (2020b). Developer agreement and policy. https://developer.twitter.com/en/developer-terms/agreement-and-policy

Twitter. (2022b). Investor relations: Quarterly results. https://investor.twitterinc.com/financial-information/quarterly-results/default.aspx

Twitter. (2020c). Reports: Information Operations. https://transparency.twitter.com/en/reports/information-operations.html

Twitter. (2021c). Search Tweets: API Reference, GET /2/tweets/search/all. Docs: Twitter developer platform. https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

Twitter. (2021d). Timelines: API Reference. Docs: Twitter developer platform. https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference

Valesco-Sacristán, M. & Fuertes-Olivera, P (2006). Towards a critical cognitive-pragmatic approach to gender metaphors in Advertising English. *Journal of Pragmatics 38*, 1982-2002.

Widdowson, H. G. (2000). On the Limitations of Linguistics Applied. *Applied Linguistics* 21(1), 3-25.

Williams, T. (1989). Sport, hegemony and subcultural reproduction: The process of accommodation in bicycle road racing. *International Review for Sociology of Sport 24(4),* 315-332.

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society 13(5), 788-806.*

Zappavigna, M. (2012). *Discourse of Twitter and Social Media*. Continuum International Publishing Group: New York.

Zappavigna, M. (2014). Enacting identity in microblogging through ambient affiliation. *Discourse & Communication 8(2),* 209-228.

Zappavigna, M. & Martin, J.R. (2018). #Communing affiliation: Social tagging as a resource for aligning around values in social media. *Discourse, Context & Media 22*, 4–12

Yaqub, U., Ae Chun, S., Atluri, V., and Vaidya, J. (2017). Sentiment based Analysis of Tweets during the US Presidential Elections. In *Proceedings of the 18th Annual International Conference on Digital Government Research*,.1–10. https://doi.org/10.1145/3085228.3085285

VITA

Robert J Redmon III

Candidate for the Degree of

Doctor of Philosophy

Dissertation: ON THE DEVELOPMENT OF A SUITE OF TOOLS FOR THE ANALYSIS OF TWITTER DISCOURSE

Major Field: English

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in English at Oklahoma State University, Stillwater, Oklahoma in May, 2022.

Completed the requirements for the Master of Arts in English at Midwestern State University, Wichita Falls, Texas in 2013.

Completed the requirements for the Bachelor of Arts in Mass Communication at at Midwestern State University, Wichita Falls, Texas in 2007.

Experience:

Research & Teaching Assistant at Oklahoma State University, 2015-2022

Teaching Assistant at Midwestern State University, 2010-2013