

# Holistic Approaches to Syntactic Variation: Wh-all Questions in English

Mary Robinson and Daniel Duncan\*

## 1 Introduction

A sociolinguistic variable is commonly understood as ‘two ways of saying the same thing’ (Labov 1972). One classic problem in the field of sociolinguistics is whether this definition of a sociolinguistic variable translates to syntax. Generative grammarians generally think that each syntactic structure has its own semantic meaning. Therefore, the argument goes, how could two different syntactic structures ‘mean’ the same thing? There are two answers to this: minimalist and variationist. Some recent work in the Minimalist framework has shown that different syntactic structures can have the same Logical Form (LF) (Myler 2016). From a variationist perspective, Dines (1980) provides the useful notion of functional equivalence, which states that syntactic variants have the same role in the discourse, even if they do not ‘mean’ the same thing (i.e., have the same LF).

Our interest here is not in relitigating whether syntactic variables exist, but rather a more practical concern: How can we study syntactic variables? We suggest that a complete picture of a syntactic variable involves its use, which includes language-internal and language-external conditioning of variation, and its grammaticality, which includes whether and in what environments a variant is licit. In this paper, we explore methodological approaches to use and grammaticality of syntactic variables. We argue that no single method can viably contribute this complete picture in and of itself. Rather, multiple methodologies are required to obtain a holistic understanding of a syntactic variable.

We illustrate this through a case study of *wh-all* questions, *wh*-questions in which the presence of *all* indicates that a plural answer is sought, in North American English (AmE). We take a three-pronged approach to the variable: a corpus study, a Twitter study, and a grammaticality judgment survey. This approach allows us to gather and analyze different kinds of data to create a complete picture of the variable in question. We show that *wh-all* questions are widespread in AmE, yet subject to regional variation, and that their use is restricted to informal registers. Such findings are a result of the combination of methodologies; no single methodology could yield every result we report. The advantages and disadvantages of each method are discussed, with the aim of establishing multi-pronged approaches as best practice in future syntactic variation research.

## 2 Approaches to Syntactic Variation

In order to better understand syntactic variation, we want to know about both the usage and the grammaticality of different syntactic variants. This distinction is roughly equal to the performance - competence distinction from Chomsky (1965). Usage of a given variant (a.k.a. performance) refers to the factors that condition the use of one variant over another. These factors can be language-internal (e.g., the surrounding syntactic context, the type of clause) or language-external (e.g., social situation, rate of speech). This can be determined through positive evidence from a data sample. The grammaticality of a variant (a.k.a. competence) refers to what is acceptable in a speaker’s grammar, as an abstract system without reference to whether the speaker ever says the possible sentences. Unlike usage, this cannot be ascertained through positive examples but rather needs to be explicated.

It is the usage of the variable in which we may expect to treat syntactic variables roughly similarly to phonological variables. There is one major issue in this respect, however: syntactic variables are infrequent, which makes finding enough tokens of the variants in question to do a quantitative analysis difficult. Studying a classic phonological variable, such as the short-a split in Philadelphia (for example, see Labov et al. 2013), is fairly straightforward in this respect. In the course of a sociolinguistic interview, one is almost guaranteed to collect enough tokens of this vowel in most contexts relevant to the short-a system, and thus with a sufficient sample of speakers will find statistically significant results. In contrast, with a syntactic variable such as the dative presentative (Wood et al. 2015), researchers would be lucky to have even one or two instances of the variants in

---

\*Thank you to audiences at PLC 42 and BICLCE-7, Renée Blake, Stephanie Harves, and Laurel MacKenzie for helpful questions and commentary.

an hour of recording natural speech. The difficulties of such rarity extend beyond the need for robust results; the infrequent occurrence of syntactic variables also poses a challenge to establishing the envelope of variation. After all, when tokens of the relevant construction are rare, the presence of one variant and absence of another in a given environment does not necessarily entail that one variant must (or cannot) be used in that environment.

One method of overcoming this difficulty is to use large corpora such as the Corpus of Contemporary American English (COCA) and others (Davies 2008-). These corpora can give the contexts in which speakers use one variant or another, and can lead to insights about extra-linguistic factors that favor one variant. However, using large corpora has two major drawbacks: First, when the variable is restricted to colloquial speech or subject to regional variation, there may not be enough tokens to draw meaningful conclusions about who uses the variant when. That is, even a massive corpus may be insufficient to address the usage problem, and at a maximum can only account for a set of pre-specified conditioning factors. Second, it is impossible to tell which corpora tokens that use variant A could have used variant B instead, because corpora do not allow the researcher to form a complete picture of what is in a speaker's grammar. In other words, a corpus of any size cannot address the grammaticality problem. These two difficulties are considered in turn.

One solution to the problem of register or regional variation is to use Twitter as a large corpus, as tweets are generally written in an informal style similar to colloquial speech. In addition, many users publicly list a location in their profile, or geotag their tweets. This makes it possible to map regional variation (Huang et al 2016, among others). Twitter has the advantage that it can capture language changes as they happen in real time (Eisenstein et al 2014) and can be used to track variation among minority groups (as in Jones 2015). In addition, Twitter can be used for both broad and narrow research. For example, tweets can be collected in large batches using Twitter's streaming API to give a broad overview of usage, or individuals with public profiles can give insight into syntactic variation within an idiolect. Data from Twitter thus has the potential to unearth tokens of colloquial or regional syntactic variants missed by a more formal corpus while at the same time shedding light on usage patterns conditioned by factors that are unaccounted for in other corpora.

Although data from corpora and Twitter may be used to study some of the internal and external language factors that condition the use of one variant over another, neither contribute to an understanding of the grammaticality of variants for a speaker. This question of grammaticality is important for syntactic variation research, because it is impossible to tell what variants are preferred if we do not know which variants the speaker controls and which variants are disallowed in the speaker's grammar. Even when we have a considerable amount of data from one idiolect (e.g., a Twitter user who is a prolific tweeter and often uses the variable of interest), we still cannot be said to know the speaker's grammar, because we do not know what is not allowed. One way to access a speaker's grammar is through experimental techniques like acceptability judgment tasks, as in Sprouse (2011) or Wood et al. (2015). However, this technique by design does not reveal any information about the usage of the variable.

In the following case study, we adopt all three of the above methodologies in an effort to obtain a holistic view of the usage and grammaticality of a syntactic variable. For each methodology, we offer both results and commentary on the successes and weaknesses of the approach. We contend that while none of the methods alone are an ideal way to study syntactic variation, a holistic approach which combines these methods yields a more complete picture of a syntactic variable. In particular, we will show that combining these methods is important for teasing apart the grammaticality of the variable across speech communities.

### 3 Case Study: *wh-all* Questions in North American English

To illustrate the importance and interactions of different approaches to studying syntactic variation, we present the *wh-all* variable as a case study. This variable represents an alternation of functionally equivalent plural *wh*-questions, given in (1).

- (1) a. **Who** bought a car?
- b. **Who all** bought a car?.

Both of the questions in (1) ask for a list of contextually relevant people who have bought a car. (1b) demands a plural answer; (1a) may be used in this way as well. Although (1a) can also be used to ask for a singleton answer, that use of the question is not at issue in this paper. Questions as in (1a) will be referred to in this paper as *wh*-questions, used here to mean specifically questions that require a plural list as an answer. Questions as in (1b) will be referred to as *wh-all* questions, used here to mean questions that indicate plurality with the use of the quantifier *all* immediately following the *wh*-word.

There is little discussion of *wh-all* in the literature. McCloskey (2000) offers a formal account of the *wh-all* variant in West Ulster English. The variant has been mentioned in conjunction with several North American dialects as well. Murray & Simon (2006) list “interrogative pronoun + *all*” among the features of Midland English, and Hendrickson (1997) lists it as a feature of Appalachian English. The most complete description of *wh-all* is found in the Dictionary of American Regional English (DARE) (2013). DARE includes entries for *who-all*, *what-all*, *where-all*, and *why-all* as lexical items, not syntactic constructions, and notes that these lexical items are scattered throughout the United States, but chiefly found in the South and the South Midland. We can thus formulate a usage question: where in North America exactly is the *wh-all* variant found? What social groups use it? Is usage stable or part of an ongoing change?

There are syntactic and semantic factors that constrain the use of the *wh-all* variant. McCloskey (2000) observes that in West Ulster English, sentences like (1b) are only acceptable with *who/what/where* as the *wh*-word, but not *when/why/how*. Additionally, the *wh*-word and *all* may be separated, as in (2).

- (2) a. **What all** did he say (that) he wanted?  
 b. **What** did he say **all** (that) he wanted?  
 c. **What** did he say (that) he wanted **all**? (McCloskey 2000: 61)

Tilleson (2018) finds evidence speakers in the Upper Midwest can similarly float *all* in *wh-all* constructions. The authors (North American, but from the Midland and South) find our own judgments in conflict with these results: we permit *when/why/how* (particularly *when*) as the *wh*-word in the construction, and rate *wh-all* sentences in which *all* is floated as ungrammatical. We thus can formulate a grammaticality question: For which *wh*-words is the construction licit? Is this the same for all speakers? Do all users of the construction have the same syntax for it?

Here we take a variant-centered approach to *wh-all* (à la Dinkin 2016) as opposed to drawing conclusions from comparisons of contexts where *wh*-questions are used versus those contexts where *wh-all* questions are used. This is due to the ambiguity of questions like (1a). Even given the surrounding context, it is difficult to reliably ascertain whether a speaker intended to use (1a) as a question that requires only one answer, or a question that requires a plural or exhaustive answer, for any given token. Therefore, collecting tokens like (1a) and (1b) and comparing the environments in which they occur is not a particularly productive vein of research. There is a way around this, however. Every time a *wh-all* question is used, the corresponding *wh*-question will always be acceptable in the same context. That is to say, there are no cases in which *Who all bought a car?* would be acceptable, but the corresponding *Who bought a car?* would not be acceptable. Therefore, this paper adopts the strategy of analyzing the contexts in which *wh-all* tokens occur, as it will naturally be a subset of those contexts in which *wh*-questions occur. It is assumed that outside of the contexts where *wh-all* questions occur, *wh*-questions occur.

#### 4 A Three-Pronged Study

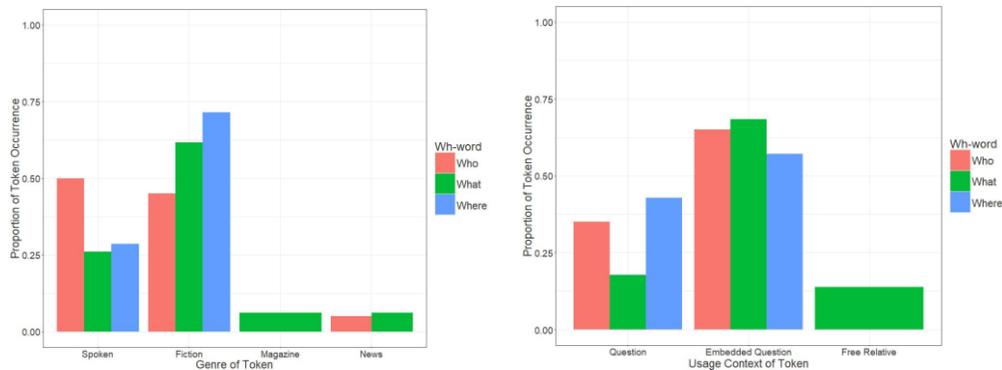
As discussed above, questions of usage and grammaticality are two different types of question necessitating different methodological approaches. We contend that a multi-pronged approach, which combines traditional corpora, corpora of more colloquial speech, and grammaticality judgment tasks, are vital for gathering a holistic picture of the usage and grammaticality of a syntactic variable. This need is immediately evident upon considering the *wh-all* phenomenon. We essentially have three problems: firstly, it is a rare construction. As such, a study using a large corpus like COCA is ideal. Secondly, *wh-all* questions appear to be used more frequently in informal speech, meaning that it

occurs infrequently even in large written corpora. This means that a more colloquial data source such as Twitter will be useful in providing context on how the usage of the variable is conditioned. Finally, data sources like COCA and Twitter only show who is using *wh-all*, but not who *can* use *wh-all*. An experimental task like a grammaticality judgment survey would thus be quite useful. For these reasons, below we take a multi-pronged approach to *wh-all* in AmE involving a COCA study, and Twitter study, and a grammaticality judgment task.

#### 4.1 COCA Study

The Corpus of Contemporary American English (COCA) is a useful tool for research on syntactic variation, as it contains 560 million words across a range of stylistic genres (Davies 2018). In the present study, the corpus’s collocate finder was used, which provided the 100 most common collocates that appeared after the search terms *who all*, *what all*, and *where all*. Those that could not be genuine *wh-all* tokens were excluded. Examples of this included “who all the” or “what all of” as this indicated that a noun phrase would follow, and therefore that it would not be a token. The remaining collocates were checked by hand and individual tokens were collected. In total, 227 *wh-all* tokens were found. These tokens were coded for the following factors: Genre, which was taken from COCA and included the categories Fiction, Spoken, News, and Magazine; *wh*-word, which was Who (n= 180), What (n=40), and Where (n=7); and *wh*-use, which was Question (*Who all is going?*), Embedded question (*I don't know who all is going*), and Free relative (*I'll eat what all she cooks*). In addition, tokens were coded according to the linguistic context following the token. These categories were Adverb (*what all else*), Auxiliary verb (*what all is*), End (*I don't know what all.*), Preposition (*I don't know what all in the past two years*), Nominal (*I don't know what all she did*), and Verb (*I don't know who all went*).

Among the tokens, *wh-all* questions were limited to *who*, *what*, and *where*. *Wh-all* is used almost exclusively in the more informal spoken and fiction genres (Figure 1). *Where-all* is absent from the news genre, and *where-all* and *who-all* are both absent from the magazine genre. This suggests that the usage of *wh-all* is restricted to informal registers of speech. It is worth noting that the few *wh-all* tokens from the news and magazine genres come from quotations from an interview or from a piece of narrative (non)fiction. As such, these are examples of relatively informal speech in those genres, and corroborate the overall trend. Figure 2 shows the distribution of *wh-all* tokens by the usage context in which they appear. As seen, the majority of tokens occur in embedded clauses, regardless of *wh*-word. *What-all* alone appears in free relatives.



Figures 1-2. *Wh-all* tokens in COCA by genre and usage context.

Figure 3 shows the interaction between the following context and usage context. While *wh-all* can precede several environments as an embedded question, as a question it (n=46) overwhelmingly appears before auxiliary verbs (87%). This is expected, as most *wh*-questions in AmE have an auxiliary verb marking tense immediately after the *wh*-word. However, there were 3 cases (7%) of a nominal following a *wh-all* question, but in all three cases, the question had a null auxiliary verb (*What all she tell you?*).

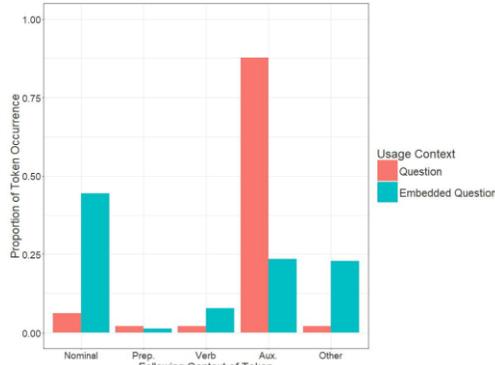


Figure 3. *Wh-all* tokens in COCA by following context and usage context.

There are significant limitations to the conclusions that can be drawn from this corpus work, however. This is in part because relatively few *wh-all* tokens were found given the large size of COCA. In addition, the corpus shows that *wh-all* is used in an informal spoken style, but the design of the corpus does not include many of these types of tokens. Furthermore, collecting the tokens via a collocate finder may have exaggerated how often AmE speakers use an auxiliary verb with a matrix *wh-all* question.

#### 4.2 Twitter Study

The second prong of this research project looked at the use of *wh-all* on Twitter. Although Twitter is not an organized and annotated corpus, it can be used as a de facto corpus of informal speech, as tweets are generally written in an informal style similar to colloquial speech. Twitter has the added benefit of including more speakers from historically underrepresented minorities than corpora do, and it can be used to track language variation among minority groups (see Jones 2015 for variation in AAVE). A linear string search of *wh-all* terms was performed, with tweets sorted by ‘latest’, to show the most recent ones. The first 100 tweets for the term “where all” did not contain any *wh-all* tokens, so that search term was excluded from further searches. For the terms “who all” and “what all”, the first 400 tweets were collected and analyzed. First, the tweets were coded for whether or not they contained a token of *wh-all*. Then the *wh-all* tokens were coded by *wh-all* use (Question vs. Embedded Question) and by the following linguistic context (Auxiliary verb, Gerund, Preposition, Nominal, Verb, and Other). Tokens were also coded for Addressee. Tweets which did not tag another user were considered addressed to No one, and tweets that were directed at other users were categorized as to a Person, a Company, an Event, or a Politician. In addition, information about the location of the user was collected if possible.

Of the 800 tweets collected, 129 were *who-all* tokens and 53 were *what-all* tokens, for a total of 182 tokens. *Wh-all* usage on Twitter differs in important ways from its use in the COCA tokens. Unlike in COCA, the majority of *wh-all* tokens on Twitter occur as questions, with a minority of tokens appearing as embedded questions or free relatives (Figure 4).

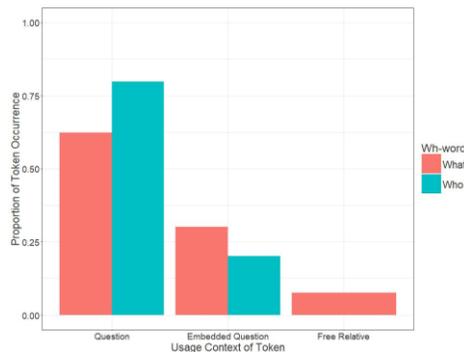


Figure 4. *Wh-all* tokens on Twitter by usage context.

The results of the following linguistic context provide a striking contrast to the COCA results. As Figure 5 shows, the rate of auxiliary verbs appearing after *wh-all* tokens in questions is only a little over 50%. This is a steep drop from the 87% rate of auxiliary verbs after *wh-all* tokens in COCA. This is because there is a larger distribution of non-auxiliary following contexts like Gerunds, as in *Wait a minute who all attending this*. In examples like this, Gerunds surface as a following context when there is a null copula. Because null copula is a feature of African American Language (AAL), we hypothesize that these tokens were collected from African American users. If this is indeed the case, we have evidence that underrepresented dialects, such as AAL, are present on Twitter but largely absent from larger corpora such as COCA.

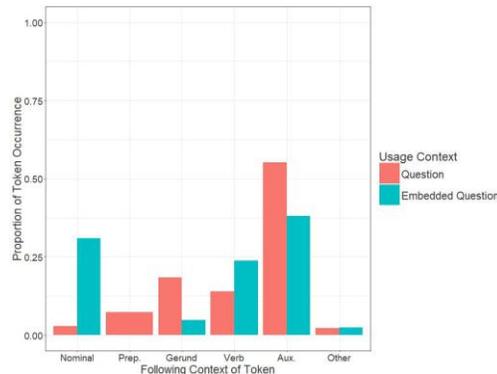
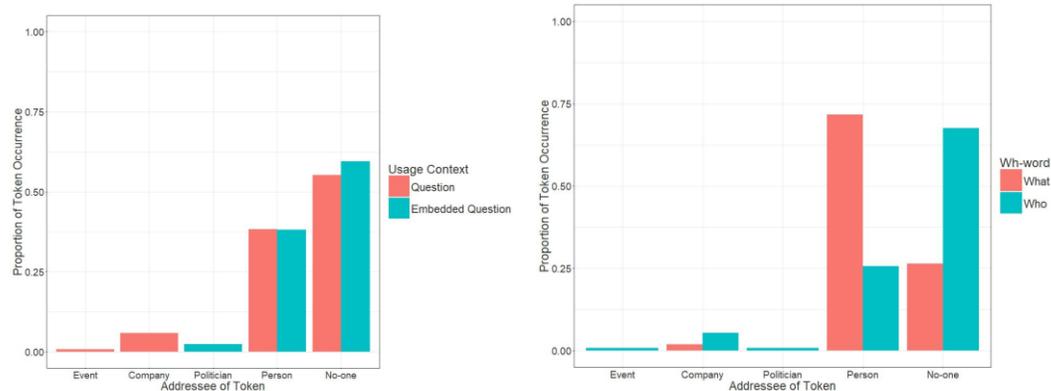


Figure 5. *Wh-all* tokens in Twitter by following context and usage context.

Although the usage context does not appear to influence or be influenced by addressee (Figure 6), we find a large difference in addressees by *wh-word* (Figure 7). Regardless of *wh-all* usage, more than 50% of tokens are directed at no one in particular, with much of the remaining tokens directed at another user. With respect to *wh-word*, we see that tweets using *who-all* are mostly (67%) directed at no one, such as *Who all's going to the Kendrick Lamar concert?* In this way, the tweets are available for anyone to respond to. By contrast, the majority of *what-all* tweets (72%) are directed at another user, with only about a quarter (26%) without an addressee. This reflects the more general tendency on Twitter to begin a conversation with an open-ended *who-all* question, and to use *what-all* questions for clarification when already in conversation with another user.



Figures 6-7. *Wh-all* tokens in Twitter by addressee, usage context, and *wh-word*.

Location information was taken from tweets that were geo-tagged, or, when that was not available, from the location the user displays as part of the profile. This information is mapped below in Figure 8. As seen, the density of *wh-all* tweets largely approximates the population density of the United States: many tokens are from major metropolitan areas like Atlanta, Chicago, and Houston, while the rural West has few tokens. Note, however, the relative absence of *wh-all* tokens in the Northeast Corridor. That cities like New York are underrepresented suggests that *wh-all* may not be used in this region.

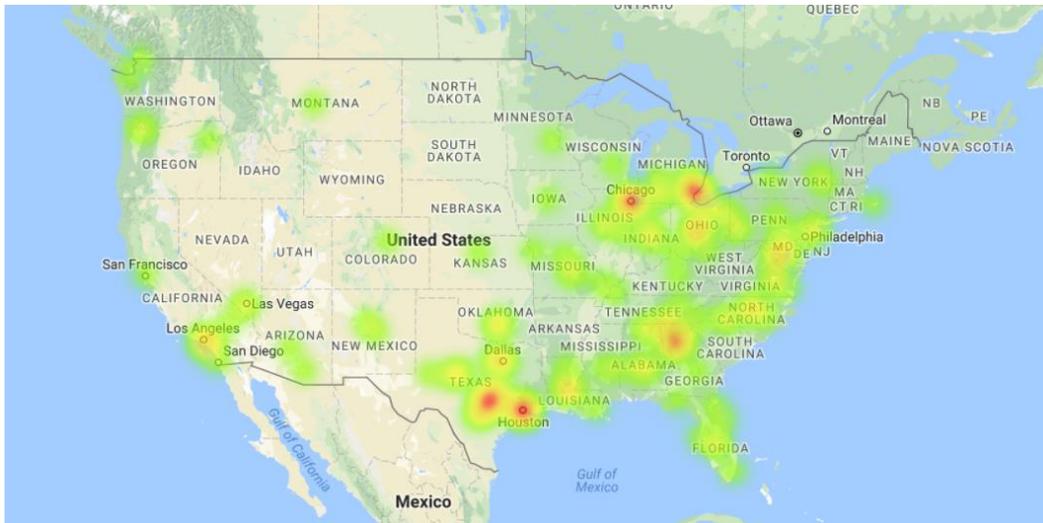


Figure 8. Distribution of *wh-all* tokens in United States.

While the data from Twitter gives us insights into the usage of the *wh-all* variable that were not available from the COCA data, there are still several limitations: it is difficult to verify that speakers are members of any given speech community, it is time-consuming to confirm that an apparent speaker is not a bot, and fewer tweets are geotagged than is ideal. Furthermore, while Twitter data shows us how *wh-all* is used in real time, it is difficult to achieve insight into language-external conditioning of variation due to the limited verifiable information we have about speakers.

### 4.3 Grammaticality Judgment Survey

An acceptability judgment experiment was conducted on Amazon Mechanical Turk (MTurk, see Wood et al. 2015 for previous work investigating grammaticality judgments of dialect features using this service). Participants ( $n=568$ ) who grew up (ages 4-14) in the United States and were native speakers of AmE answered a 47-question grammaticality survey, in which they rated different randomly presented questions along a 7-point Likert scale. To test language-internal factors of *wh*-word and position of *all*, test sentences were created based on the frame in (3).

- (3) What (all) did he (all) buy at the store (all)?

Each *wh*-word was inserted into the frame with *all* in one of the three positions to create the target questions. The remaining questions in the survey were a mixture of grammatical and ungrammatical fillers. Participants' demographic information, including where they grew up was collected upon completion of the survey. The location information was coded by the AmE dialect regions identified in Labov et al. (2006).

The responses to the MTurk survey were normalized to z-scores and modeled using linear mixed effects regression, in which participant was a random effect. The language-internal factors of position of *all* and *wh*-word, and language-external factor of region (where speakers grew up, coded based on dialects described in Labov et al. 2006) were fixed effects with a significant effect on grammaticality rating ( $p < 0.05$  for all discussed results, see Table 1). Participants preferred *all* to be next to the *wh*-word, and disfavored *when*, *how* and *why* as *wh*-words. Tukey's HSD post-hoc tests showed a hierarchy of preference for *wh*-words: *who* > *what* > *where* > *when/why/how* ( $p < 0.05$ ). While most regions rated *wh-all* questions as grammatical, participants from the Northeast United States (coded as New York City, Western New England, and Eastern New England) rated them poorly. Figure 9 compares the ratings of these northeastern regions with the ratings of the South and Midland, where *wh-all* questions were strongly preferred compared to other regions, ( $p < 0.05$ ), and Western Pennsylvania (PA), where *wh-all* questions were preferred ( $p=0.055$ ). The figure shows the density of ratings from each region in relation to the participant's own baseline.

Fixed effects:	Estimate	Pr(> t )	Significance
Intercept ( <i>what all...</i> )	0.765	<<0.0001	***
FQ position-- <i>wh-AUX-all...</i>	-0.827	<<0.0001	***
FQ position--... <i>all-end</i>	-1.31	<<0.0001	***
<i>Wh-word--who</i>	0.197	0.0128	*
<i>Wh-word--where</i>	-0.165	0.0368	*
<i>Wh-word--why</i>	-1.06	<<0.0001	***
<i>Wh-word--when</i>	-1.07	<<0.0001	***
<i>Wh-word--how</i>	-1.07	<<0.0001	***

Table 1. Language-internal factors significantly affecting acceptance of *wh-all* (\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ).

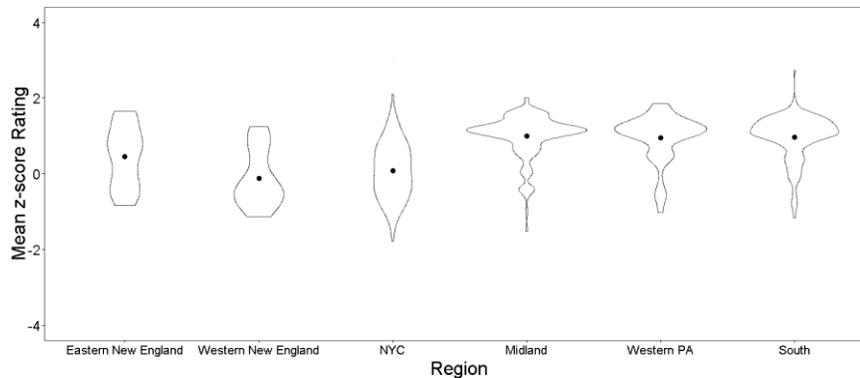


Figure 9. *Wh-all* ratings for *who-all*, *what-all*, *where-all* by region.

The grammaticality judgment survey is the best tool we have to probe participants' judgments and find out what is in their grammar. This is more straightforward evidence of who finds *wh-all* grammatical, with which *wh*-words, in what syntactic environments. This more direct probing allows us to conclude that there is evidence for micro-parametric variation in AmE with respect to *wh-all*. One advantage of this kind of experiment is that it lets us compare results across studies. For example, this study could not replicate Tilleson's (2018) finding that speakers in the Upper Midwest accept *wh-AUX-all* (i.e., *What did he all buy at the store?*) more than speakers in other parts of the United States. An advantage of the grammaticality judgment survey is to find out not only what is grammatical in a participant's grammar, but also what is ungrammatical. The results of this study indicate that *when-all*, *how-all*, and *why-all* are disliked by most speakers of AmE, which explains their absence from corpora. Note, however, that such a survey can only make predictions about usage; the corpus study is necessary to provide usage data corroborating the grammaticality data.

## 5 Discussion and Concluding Remarks on Methodology

This paper has provided a quantitative look at *wh-all* usage in the United States, and has argued that syntactic variation data should come from varied sources in order to get the most complete picture possible of the variation. Data for this paper comes from three sources: COCA, Twitter, and MTurk, each of which made unique contributions to the understanding of *wh-all*. The COCA data showed that *wh-all* is largely confined to informal genres and suggested that the variant is more common in embedded questions than in matrix questions. While the Twitter data, by virtue of coming from an informal source, supported the former claim, it did not support the latter claim. On Twitter, the majority of *wh-all* tokens are in matrix questions. Twitter data also appears to include *wh-all* users who were not included in COCA; whereas the vast majority of *wh-all* tokens are followed by an auxiliary in COCA, there are a sizable number of tokens in Twitter that are in a context involving a

null copula. For this reason we suggest that AAL speakers use *wh-all*, but are not included in COCA. We also found that there is an interaction between addressee and *wh*-word: *who-all* is typically addressed to no one in particular, while *what-all* typically has a specific addressee. Although COCA and Twitter provided valuable insights into the usage of *wh-all*, understanding the grammaticality of it would not have been possible without the grammaticality judgment survey. The survey revealed that not all speakers have the same grammar for *wh-all*. Some speakers, mostly in the Northeast, do not accept *wh-all* questions at all, while many speakers in the South, Midland, and Western PA prefer *wh-all* questions compared to speakers from other regions. In addition, there is variation across speakers in all regions in which *wh*-words are acceptable in *wh-all* questions, as well as how far the quantifier *all* can float from the *wh*-word.

It is important to note that the findings here are more informative when taken in conjunction. Much of what we learn from COCA and Twitter is not directly from the source but from comparing one source with the other. For example, we have suggested that the rate of an auxiliary following a *wh-all* token in matrix questions is lower in Twitter than in COCA because Twitter includes more examples of non-standard speech in which *wh-all* is followed by a null copula. However, because *wh-all* tokens in matrix questions are most commonly followed by auxiliaries in the Twitter data, we cannot reach the above conclusion without comparison to the baseline in COCA. To claim that Twitter includes a wider range of *wh-all* users than COCA would seem to imply that *wh-all* tokens occur more frequently in Twitter than in COCA. To determine if this is indeed the case, we took a small sample from COCA in the same manner as the Twitter sample. As such, we collected instances of the linear strings *who-all* and *what-all* (n=200) and identified which of this sample were *wh-all* tokens. We then compared the frequency at which the strings were *wh-all* tokens in COCA vs. Twitter using a chi square test. Although the string *what-all* is not significantly different between the two corpora ( $p=0.1416$ ), the string *who-all* is significantly more likely to be a *wh-all* token in Twitter than in COCA ( $p<<0.0001$ ). This is promising, as the *wh-all* tokens in matrix questions followed by a nominal (a non-standard form) in the Twitter data were typically *wh-all* questions in which *who* was the *wh*-word. It thus appears that by virtue of including more non-standard speakers, Twitter contains more of the *wh-all* variant.

The grammaticality judgment survey similarly is more informative when taken in conjunction with the corpora. While the absence of *wh-all* tokens in which *when/why/how* is the *wh*-word from COCA and Twitter may suggest a grammatical constraint, the result alone could be an artifact of the rarity of syntactic variants. The grammaticality judgment survey offers evidence that speakers disprefer *when/why/how-all*, and that for many speakers such tokens may be ungrammatical. However, this study alone cannot prove that the dispreference results in non-usage. The corpora and survey suggest a conclusion--that *when/why/how-all* are generally ungrammatical, so do not occur in speech--by providing evidence from both grammaticality and usage that the individual studies alone could not reach. Likewise, the survey and Twitter data combine to suggest an isogloss excluding the Northeast Corridor from *wh-all* usage. There is also a straightforward additive effect of multiple methodologies. The corpus study and Twitter data show that the use of *wh-all* questions is subject to register-based variation, which was not a possibility considered in the survey. Likewise, the grammaticality judgment survey showed that the position of *all* strongly affects the acceptability of a *wh-all* question, something that was unexplorable through the corpus work. By engaging in multiple methodologies, we obtain a wider range of results.

As this case study has shown, no single method gives a complete picture of the *wh-all* variable. Syntactic variation is about both the usage and grammar of a variable, so it is necessary to use tools from both sociolinguistics and syntax. Corpora like COCA are readily available and provide useful information for the researcher. At the same time, the field is quickly standardizing best practices for corpus analysis using Twitter and other social media as data sources. Further, Sprouse (2011) has established that online experiments are a fast and reliable way to elicit syntactic judgments. This has implications not only for formal theorists, but variationists interested in collecting judgments on stigmatized syntactic variants (a la Buchstaller and Corrigan 2009) as well. While each of these techniques are useful, we have shown that they are stronger together. Although we recognize that a multi-pronged study of a variable is both time- and resource-consuming, it is our contention that where possible the best practice for studying syntactic variation is to engage in such an effort.

This is, of course, easier said than done. It is not difficult to conduct a multi-pronged study of variation in a well-studied, widely-used language like English. The ideal, although not impossible,

takes more effort to achieve in other languages. Many communities of minority languages use the internet to foster their speech community, and it may be possible to conduct surveys with them using online platforms. Corpora do exist or are being built for a wide number of understudied and minority languages, and where readymade corpora do not exist, it is possible to continue to use communities on Twitter, or to create a corpus based on chatroom conversations in the target language (a la Bleaman 2018). That said, this is certainly not the case for all languages. The ideal of multiple methodologies thus highlights the need for richer sources of variationist data for researchers on minority and understudied languages.

## References

- Bleaman, Isaac. 2018. Big data in a low resource language: Syntactic variation in Hasidic Yiddish on the web. Paper presented at the 2018 Annual Meeting of the LSA, Salt Lake City.
- Buchstaller, Isabelle and Karen P. Corrigan. 2009. "Judge not lest ye be judged": Exploring methods for the collection of socio-syntactic data. In *Language Variation -- European perspectives III*, ed. F. Gregersen, J. K. Parrott and P. Quist, 149–160. Philadelphia: John Benjamins.
- Cassidy, Frederic. 1996. *Dictionary of American Regional English*. Cambridge, MA: Harvard University Press. Available online at [www.daredictionary.com](http://www.daredictionary.com).
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Davies, Mark. 2008-. The Corpus of Contemporary American English (COCA): 520 million words, 1990-present. Available online at <https://corpus.byu.edu/coca/>.
- Dines, Elizabeth R. 1980. Variation in discourse—"and stuff like that." *Language in Society* 9(1):13–31.
- Dinkin, Aaron. 2016. Variant-centered variation and the like conspiracy. *Linguistic Variation* 16(2):221–246.
- Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9:e113114. <http://dx.doi.org/10.1371/journal.pone.0113114>.
- Hendrickson, Robert. 1997. *Mountain Range: A Dictionary of Expressions from Appalachia to the Ozarks*. New York: Facts on File.
- Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding U.S. regional variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59:244–55.
- Jones, Taylor. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech* 90(4):403–440.
- Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English*. New York: Mouton de Gruyter.
- Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89(1):30–65.
- McCloskey, James. 2000. Quantifier float and wh-movement in an Irish English. *Linguistic Inquiry* 31(1):57–84.
- Murray, Thomas E., and Beth Lee Simon. 2006. What is dialect? Revisiting the Midland. In *Language Variation and Change in the American Midland: A New Look at 'Heartland' English*, ed. T. Murray and B.L. Simon, 1–30. Amsterdam: John Benjamins.
- Myler, Neil. 2016. *Building and Interpreting Possession Sentences*. Cambridge, MA: MIT Press.
- Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1):155–167
- Tillessen, Paul. 2018. Floating all In the Upper Midwest dialect of English. Poster presentation at the 2018 Annual Meeting of the LSA, Salt Lake City.
- Wood, Jim, Laurence Horn, Raffaella Zanuttini, and Luke Lindemann. 2015. The Southern dative presentative meets Mechanical Turk. *American Speech* 90(3):291–320.

Mary Robinson  
Department of Linguistics  
New York University  
New York, NY 10003  
[mkr361@nyu.edu](mailto:mkr361@nyu.edu)

Daniel Duncan  
School of English Literature, Language and Linguistics  
Percy Building  
Newcastle University  
NE1 7RU  
United Kingdom  
[daniel.duncan@ncl.ac.uk](mailto:daniel.duncan@ncl.ac.uk)