English Language and Linguistics, 28.2: 305–339.
doi:10.1017/S1360674323000631

When to (not) split the infinitive: factors governing patterns of syntactic variation in Twitter-style Philippine English[1]

WILKINSON DANIEL WONG GONZALES
The Chinese University of Hong Kong
(Received 23 March 2023; revised 6 October 2023)

The variability of adverbial placement in the modified infinitive construction (i.e. split infinitives vs. full infinitives with adverbial pre- and post-modification) has been widely discussed in the (American English) literature. Yet a convincing generalized explanation for the variation that simultaneously incorporates language-internal and language-external factors has yet to be found, particularly in English varieties that have not received as much scholarly attention as standardized varieties. This article investigates modified infinitive syntactic variation in Twitter-style Philippine English (PhE) using a 135-million-word Twitter corpus. It adopts a Bayesian approach in conducting a multiple multinomial regression analysis of the said variation, with the help of Deep-Learning-based demographic inference tools. Although the conditioning effects of some factors diverge from patterns discussed in prior work, the results generally show that language-internal (e.g. stress and rhythm, adverb type, adverb length) and language-external factors (i.e. time, age, sex, geography) jointly shape the choice to split the infinitive in this linguistic style of PhE.

Keywords: split infinitives, syntactic variation and change, computational approaches to sociolinguistics, internal variation in Philippine English(es), Twitter-style Philippine English, multinomial Bayesian regression

1 Introduction

In English, the adverbial modification of *to* + infinitive phrase constructions is syntactically variable (Quirk *et al.* 1985: 497). The adverb can come immediately before the *to*-phrase, in an ADV + *to* + INFINITIVE VERB construction (1a). It can also be placed after the *to*-infinitive phrase, in a *to* + INFINITIVE VERB + (…) + ADV construction, as in (1b) and (1c) (Kostadinova 2020: 103). Another syntactic alternative would be to place the adverb between *to* and the verb in the *to*-phrase, in a *to* + ADV + INFINITIVE VERB construction, similar to the example in (1d). The last construction is more

[1] This project would not have been possible without the support of The Chinese University of Hong Kong Faculty of Arts Direct Grant (4051228) 'Exploring Variation and Change in Chinese-related Multilingual Practices in East Asia'.

commonly known by laypeople and prescriptive grammarians as the 'split infinitive' construction.[2]

(1) (a) She has tried *consciously* to stop worrying about her career. (before *to*-phrase)
(b) She has tried to stop *consciously* worrying about her career (after *to*-phrase)
(c) She has tried to stop worrying about her career *consciously*. (after *to*-phrase)
(d) She has tried to *consciously* stop worrying about her career. (within *to*-phrase)
(Quirk *et al.* 1985: 497)

In this article, the last construction is not regarded as a grammatical error. Instead, adopting a variationist stance (Labov 1972), the study considers all these *to*-phrase constructions to be equally valid and grammatical construction variants of the modified infinitive. This perspective contrasts with what many prescriptivists have historically claimed – that the split infinitive construction is not grammatical and should be avoided at all costs. The choice to abandon such a viewpoint is justified upon examining findings in previous linguistic investigations of the construction. Early work has identified linguistic conditions that warrant the use of the split infinitive construction, or at least make the use of the split infinitive appear more idiomatic, acceptable and/or natural (Kato 2001; Mitrasca 2009; Mikulová 2011; Koivistoinen 2012). In other words, if one regards acceptability, naturalness and idiomaticity as correlates of grammaticality, then previous studies have identified grammatical 'rules' for when to split the *to* + infinitive construction. Researchers have found language-internal factors that constrain the syntactic variation in *to*-phrase use (e.g. ambiguity resolution, adverb type, etc.) (see section 2). However, the vast majority of the work on the linguistic conditioning of split infinitive use is limited to standardized varieties such as American English (AmE) and British English (BrE), which are often regarded by many as the standards of English grammar. Given that English is used around the world and has taken various developmental trajectories depending on the sociohistorical context of the region, it is unclear whether the patterns described in previous work would hold for lesser-known varieties of English, which have also been documented to exhibit variable usage in the infinitive construction (Gonzales & Dita 2018).

The present study will focus on Philippine English (PhE), operationalized here as a cluster of English varieties that primarily emerged out of contact between (American) English and historically indigenous languages such as Tagalog (Thompson 2003; Gonzales 2017; Gonzales 2022a). The choice to investigate PhE was partially motivated by a previous finding that PhE has the highest incidence rates of split infinitives in a comparison of twelve world Englishes (Gonzales & Dita 2018: 254), and thus, higher rates of potential variation for examination. However, unlike prior

[2] Perales-Escudero (2011) has proposed that English's prohibition of splitting infinitives can be understood as a nineteenth-century response linked to the ideology of Teutonic purity, since languages like German do not have the 'split' construction. He reports that there is no evidence for the proscription stemming from the enforcing of Latinate standards.

work, the current study will analyze syntactic variation of infinitive constructions in an understudied 'style' or way of using PhE – Twitter-style PhE (Eckert & Rickford 2001: 121). Instead of viewing PhE on Twitter as an inaccurate representation of PhE not worthy of study, I follow the lead of contemporary variationists (Eckert & Rickford 2001; Ilbury 2020) and adopt the view of Twitter-style PhE as a one of the many styles (e.g. essay style, casual dialogues style, educated style) (Bautista 2000) that constitute PhE. From this perspective, it is imperative to study the multiple styles of PhE and how they interact with social and linguistic variables to get a more holistic understanding of the English variety (Eckert & Rickford 2001: 1).

Using the findings of previous work as a guide, I plan to identify language-internal factors that condition the alternation between split constructions and those that are not, that is, I investigate whether certain linguistic constraints identified in earlier related work on split infinitives in English apply to Twitter-style PhE. The extent to which these linguistic factors condition the variation in split infinitive use will also be examined. These findings are expected to contribute to our understanding of syntactic conventions involving the infinitive in PhE – an area of inquiry that has received very little scholarly attention (Gonzales & Dita 2018).

In addition to language-internal factors, the present study will also attempt to enrich existing research and identify potential language-external factors (i.e. social, geographical and diachronic) that may also constrain the said variation. This is, to my knowledge, something that has not been investigated in Twitter-style PhE. Language-external factors have been shown to be robust predictors of linguistic variation in PhE (Gonzales 2023b) and other neighboring English varieties (Starr & Balasubramaniam 2019; Leimgruber *et al.* 2021; Gonzales 2022b, 2023a) (see section 2). So, this article will examine the possible effects of these factors while also *jointly* considering the potential effects of language-internal factors. The following questions guide the article:

1. Which language-internal and language-external factors condition the variation in modified infinitives in Twitter-style PhE?
2. To what extent do these factors condition the variation in adverb[3] placement?

By answering these questions, the study hopes to narrow the glaring gap between variationist sociolinguistic research in East Asia and that of many well-known Anglophone territories (e.g. the United States). It also hopes to contribute to our understanding of internal variation in Twitter-style PhE and PhE in general, which has often been thought to be negligible, if not non-existent (Llamzon 1997; Lee & Borlongan 2022). Finally, from a methodological perspective, the article also hopes to (i) normalize the inclusion of language-external variables in variationist analyses in the region; (ii) popularize the use of social media data for linguistic analyses; and (iii) encourage the use of state-of-the-art machine learning techniques (e.g. Deep Learning)

[3] I acknowledge that 'adverb phrase' might be a more accurate term, to account for constructions like *to **quite literally** fly*. However, I have decided to use 'adverb' in this article to avoid confusion, as I will only be investigating lexical adverb splitters.

and stochastic and probability-based statistical modeling (e.g. Bayesian regression) in the analysis of (socio)linguistic variation in the region (Gonzales 2004). The article adds to the growing body of work that utilizes these methods (MacKenzie 2020; Hiramoto *et al.* 2022; Levshina 2022; Gonzales 2023b).

The rest of the article is structured as follows: section 2 briefly summarizes literature that examines language-internal and language-external conditions for variation in *to*-infinitive constructions. This is followed by a description of the methods employed (section 3), where I present the data source and data analysis procedures. The results and discussion can be found in section 4. In section 5, I summarize the article, answer the questions posed earlier and provide some concluding remarks, including notes on future research directions.

2 Variation in the modified infinitive construction

Most of the studies on adverbial constructions address variation in the modified infinitive construction. Although some differences can be found between the data sources of these studies – English usage guides vs. novels, spoken vs. written registers – some key factors are observed to shape the syntactic alternation in modified infinitives consistently (Kato 2001; Mitrasca 2009; Perales-Escudero 2011; Kostadinova 2020). The bulk of these factors are language-internal – some are relevant to the infinitive, while others have to do with the nature of the adverb. Also pertinent are the prosodic and semantic properties of the utterance containing the modified infinitive.

Perhaps one of the strongest effects on modified infinitive variation is the presence of ambiguity (Quirk *et al.* 1985: 497; Calle-Martín & Miranda-García 2009: 361). Speakers tend to split the infinitive if doing so would resolve the ambiguity caused by not placing the adverb directly before the verb (Mikulová 2011: 20). In sentences like (1b), for example, it is unclear whether the sentence refers to a conscious stop or a conscious worry. However, if the adverb *consciously* was placed between *to* and the infinitive, then the meaning of the utterance would be clearer (Quirk *et al.* 1985).[4] Mikulová corroborates this in her study of the construction in popular electronic corpora (e.g. *British National Corpus*); she found that 'non-native' participants tend to use the split infinitive among themselves to achieve clarity of expression.

Another notable factor that conditions infinitive variation is the type of adverb. Scholars have observed that certain adverb class types tend to favor the splitting of the infinitive over others (Koivistoinen 2012). Analyzing corpus data of historical and contemporary American English, Kostadinova (2020) found that adverbs of degree, manner, stance and time (e.g. *really*, *always*, *usually*, *actually*, *maybe*) – based on Biber & Quirk's (2012) adverbial typology – tend to encourage the splitting strategy, whereas additive and restrictive adverbs tend to discourage it.[5] In other words, the odds that a

[4] Note that splitting is not required for disambiguation, as in (1a).

[5] Additive adverbs show that one constituent is being added to another constituent either at a clausal level (e.g. *Oh, my dad was a great guy,* ***too***) or at a phrasal level (e.g. *I can hear the hatred but* ***also*** *the need*). Restrictive adverbs and

to-phrase is split are lower if the adverb belongs to the additive or restrictive class of adverbs (i.e. *just*, *only*, *even*, *too*, *else*, *also*, *especially*, *particularly*). Quirk *et al.* (1985) in their analysis of split infinitives in contemporary English observed that split infinitive constructions tend to occur with 'subjuncts[6] of narrow orientation', particularly those with a grading and focus orientation (Quirk *et al*. 1985: 497; Perales-Escudero 2011: 333). They observed that infinitives are often split by adverbs that do not modify how we perceive the verb, but perform a subordinate role in the sentence (e.g. amplification, emphasis) (e.g. *to* ***actually*** *go* as opposed to *to* ***willingly*** *go*) (Quirk *et al.* 1985). Koivistoinen's (2012: 17) investigation of split infinitives shows that the split infinitive construction also tends to correlate with focusing adverbs (e.g. *even*, *merely*, *only*) in present-day English.

In addition to adverb type, the length of the adverb may also play a role in conditioning the syntactic variation in modified infinitives. The placement of the adverb has been found to be sensitive to the number of syllables in the adverb relative to the number of syllables in the verb. The numbers show that adverbs tend to be placed in a *to* + ADV + VERB construction if the adverb is shorter than the verb or if the adverb and verb have the same number of syllables; however, if the adverb has more syllables than the verb, it tends to come after the verb (Kostadinova 2020). This is at least true in the case of American English.

An effect on split infinitive variation that tends to be discounted or tip-toed around in the literature is the prosody of modified infinitives, particularly the rhythm and stress of the adverb and the verb (Crystal 1984; Calle-Martín & Miranda-García 2009: 356). It has been claimed that split infinitives tend to appear in rhythmically neat constructions (i.e. the 'natural' *te-tum te-tum* rhythm as reflected in the unstressed–stressed unstressed–stressed or U–S U–S stress sequence). Correlations between the split infinitive construction and 'natural' rhythm patterns have been commonly cited as evidence for this despite the existence of some outlier cases (e.g. *to better prepare*, to BE-tter pre-PARE, U S-**U** **U**-S) (Crystal 1984; Perales-Escudero 2011). For example, the phrase *to* ***boldly*** *go* has a rhythmic pattern that is very 'neat' (e.g. to BOLD-ly GO, U S-U S) compared to the other constructions (e.g. ***boldly*** *to go*, S-U U S; *to go* ***boldly***, U S S-U), which contain a consecutive sequence of weak or strong syllables (Crystal 1984: 30; Calle-Martín & Miranda-García 2009). The tendency for split constructions to appear in constructions that are rhythmically 'natural' suggests that the split construction tends to be used if it contributes to the 'natural' rhythm (table 1, example for Condition D) and is avoided if the placement of the adverb in between the *to*-phrase (adverb before the verb) disrupts it. The reported preference for 'natural'

additive adverbs share the characteristic of directing attention to a particular element within a clause. However, while additive adverbs highlight the addition of information, restrictive adverbs accentuate the significance of a specific part of the statement by limiting the truth value of the proposition mainly or solely to that component (e.g. ***Only*** *those who can afford the monthly payment of $1,210.05 …can be ordered to pay*) (Biber & Quirk 2012: 556).

[6] 'Subjunct' here is operationalized by Quirk *et al.* (1985) as an adverbial with a subordinate role in comparison with other clause elements. Commonly used lexical subjuncts in the split infinitive literature include *really*, *truly*, *rather*, *so*, *further*, *even*, *ever*, *either*, *better*, *actually*, *fully*, *effectively*, *quite*, *so*, *just*, *still*, *thoroughly*, *completely* and *almost*.

Table 1. *Stress and rhythm conditions for modified infinitive constructions (Calle-Martín & Miranda-García 2009: 356)*
S = stressed, U = unstressed, ? = claimed to be rare

Condition	Description	Example			
A	monosyllabic verb, regardless of number of adverb syllables		S-U	**S**	
		to	*rea-lly*	***touch***	
		TO	ADV	VERB	
B	Finally stressed disyllabic verb, with monosyllabic and disyllabic adverbs		S	U-**S**	
		to	*quite*	*for-**get***	
		TO	ADV	VERB	
C	Initially stressed disyllabic verb, with trisyllabic adverbs		S-U-U		**S**-U
		to	*su-dden-ly*		***cha**-llenge*
		TO	ADV		VERB
D	If adverb splitter contributes to 'natural' rhythm		S-U		S
		to	*fair-ly*		*judge*
		TO	ADV		VERB
		('natural rhythm', adverb splitter to avoid consecutive stress/unstressed syllables like below)			
		?	S		S-U
		to	*judge*		*fair-ly*
		TO	VERB		ADV
		(two unstressed syllables in a series)			

constructions is claimed to be reflected in the four prosodic conditions (i.e. Conditions A to D) identified by Calle-Martín & Miranda-García that appear to favor the use of split infinitive constructions. I summarize them in table 1 (see Calle-Martín & Miranda-García 2009 for a detailed explanation).

It is worth noting that some scholars have questioned the role of rhythm and stress in shaping the choice of modified infinitive construction. Using corpus evidence, Perales-Escudero (2011: 331) argues that adverb type is a better predictor of split infinitive (dis)use than prosody.

Conventionality (e.g. idiomaticity, collocations) could be another significant factor explaining split infinitive use. It has been shown that some pre-constructed phrases associate strongly with specific lects and registers (Perales-Escudero 2011: 332). In the context of academic registers, for example, certain split infinitive constructions tend to appear more frequently in academic discourse (e.g. *to better* VERB, *to effectively* VERB). Aside from register-specific conventions, there also exist collocational tendencies in dialects. The research on World Englishes, for example, has shown that different dialects of English can have preferred split infinitive patterns (Gonzales & Dita 2018). Using the *International Corpus of English* (ICE), Gonzales & Dita (2018) discovered that English varieties worldwide have different sets of infinitive collocates. In Hong Kong English (HKE), for example, *to really* VERB and *to further* VERB constructions are

the most popular, whereas in PhE, the most popular adverb splitters include *really*, *further*, *just*, *immediately*, *fully*, *better*, *finally* and *completely* (Calle-Martín & Romero-Barranco 2014; Gonzales & Dita 2018: 250). Speakers of PhE may, for instance, adopt 'endonormative' grammar standards (Schneider 2003: 255) and use local split infinitive collocational patterns, even when other linguistic factors like ambiguity and prosody favor their disuse. The tendency to use chunks would not be surprising given that collocations and idiomatic expressions are an important component of (native) language.

There is much less research on the potential effects of language-external factors on modified infinitive variation. One of the few studies that considers these factors is that of Kostadinova (2020), who explained the variation using diachronic variables as well as 'prescriptivism-related predictors' or prescriptivist ideological variables (Kostadinova 2020: 110). Kostadinova found that the odds of an infinitive being split have increased over time. Moreover, she found that ideological orientation (e.g. adherence to prescriptivism) influences variation. Speakers who use constructions that are commonly proscribed (e.g. the use of passives, sentence-initial *and/but*, *less* with plural nouns, discourse particle *like*) tend to split the infinitive as well. In other words, those who are not concerned about using non-prescriptivist forms would also be unconcerned about using split infinitives (Kostadinova 2020: 112). Where sociodemographic factors are concerned, there is evidence that region can condition variation (Calle-Martín & Romero-Barranco 2014), but, to my knowledge, evidence of the conditioning effect of age and sex on infinitive variation has yet to be discovered.

3 Methods

3.1 Data source

The *Twitter Corpus of Philippine Englishes* (TCOPE) was used for this study (Gonzales 2023b). The TCOPE is a 135-million-word corpus that was created from roughly 27 million tweets sampled from 29 major cities in the Philippines (figure 1). The data included cover the years 2010 to 2021, with roughly 5 to 17 million words per year. The corpus was selected because at least part of it is open access, and the data were sampled from many rural and urban cities in the country. In addition to geographical metadata, each utterance in the TCOPE is linked to the public Twitter profiles of the tweeter, where information about age and sex may be gleaned or derived using computational methods (Wang *et al*. 2019; Gonzales 2004). The availability of socio-demographic metadata makes TCOPE the optimal dataset to investigate sociolinguistic variation in the use of modified infinitives in PhE because other PhE corpora do not contain such information. Furthermore, because the TCOPE is available in program-readable format (e.g. spreadsheet form), coding each utterance for linguistic factors is possible with the help of natural language processing packages. The corpus is also large enough that the results can be claimed to reflect one facet of PhE use – PhE as used in computer-mediated-communication (CMC) or more specifically, Twitter. It offers an opportunity to enrich work on the nature of PhE, as

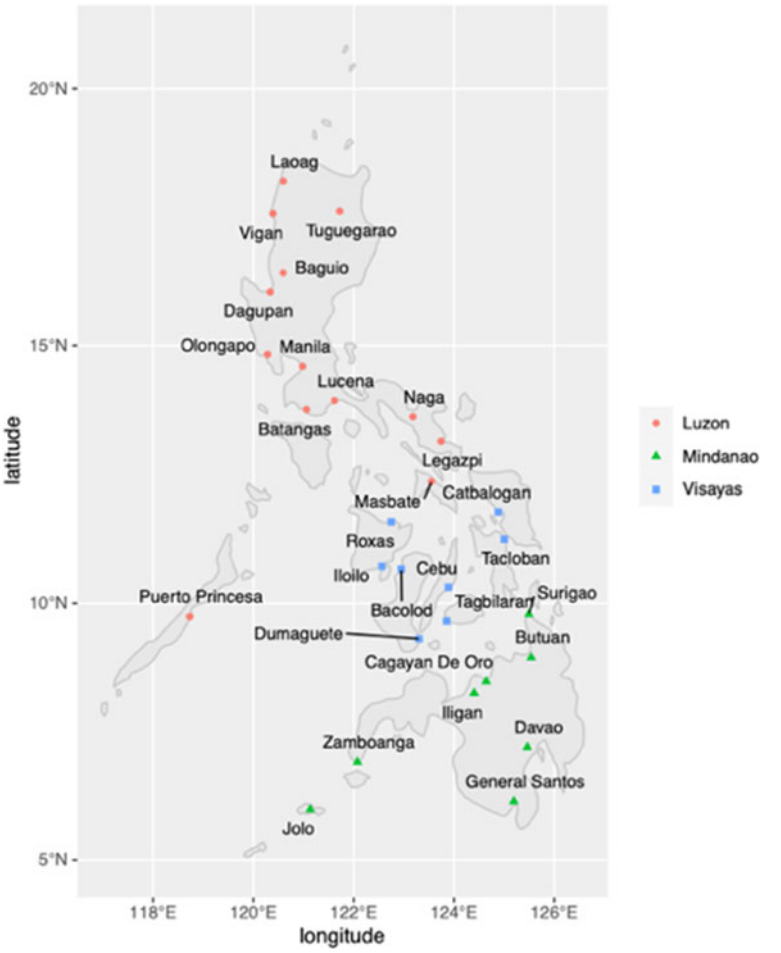


Figure 1. Cities from which the TCOPE was sampled, grouped by the three sociopolitical regions of the Philippines. Figures were adopted from the corpus overview papers (Gonzales 2023b)

most of the work on PhE have only been limited to selected written and spoken styles (e.g. essays, monologues, dialogues) used by adults (e.g. the Philippine component of the *International Corpus of English* or ICE-PH) (Bautista 2004). The TCOPE can be used for the study of Twitter-style PhE (Eckert & Rickford 2001; Ilbury 2020). Finally, the TCOPE is also one of the very few corpora in the Philippines that contains part-of-speech and dependency parsing information derived using *nltk* and *spaCy* (Bird, Klein & Loper 2009; Honnibal *et al.* 2020). This information is useful for the efficient extraction of modified infinitive constructions, as the next subsection will show.

3.2 *Data pre-processing*

In this investigation, I am interested in analyzing the syntactic variable of adverb phrase placement in modified infinitives, which has three variants:

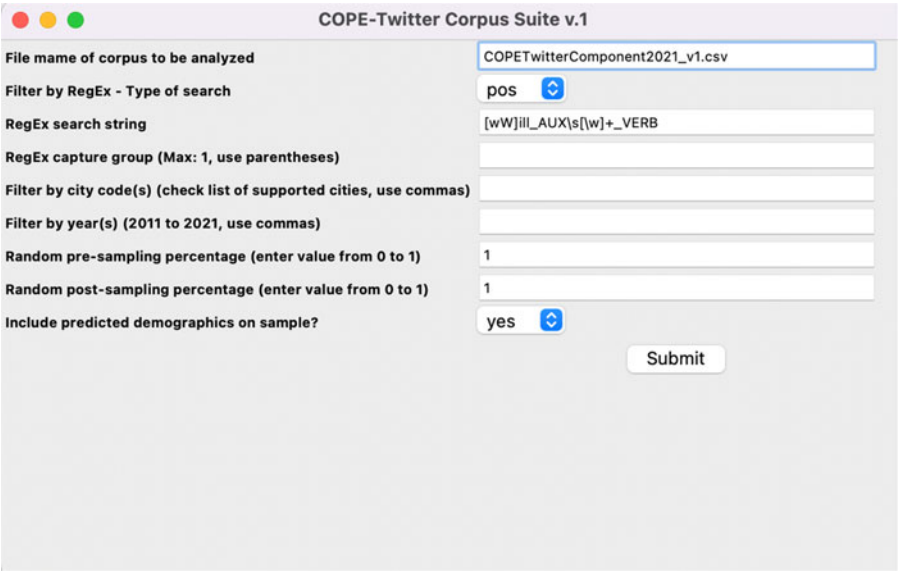


Figure 2. Graphical user interface of *Twitter Corpus Suite* (COPE)

1. **ADV** + *to* + INFINITIVE VERB (before *to*-phrase)
2. *to* + INFINITIVE VERB (+ …) + **ADV** (after *to*-phrase)
3. *to* + **ADV** + INFINITIVE VERB (within *to*-phrase)

The first step in obtaining data was to search the TCOPE for tokens that followed any of these three modified infinitive patterns. I used a self-developed program, *Twitter Corpus Suite* (TCS), to extract the utterances efficiently. TCS takes in a regular expression (RegEx) and other parameter values (e.g. sampling size, corpus to be analyzed, filtering options, type of search) and returns a spreadsheet file containing data that match the RegEx and set parameters (figure 2).

The following RegEx search strings were input in the program:

1. \s([\w]+)_ADV\sto_PART\s[\w]+_VERB (before *to*-phrase)
2. to_PART\s[\w]+_VERB\s([\w\s]+)_ADV (after *to*-phrase)
3. to_PART\s([\w]+)_ADV\s[\w]+_VERB (within *to*-phrase)

Tokens that followed the ADV + *to* + INFINITIVE VERB construction were coded as 'pre', whereas tokens following the *to* + INFINITIVE VERB + ADV construction were coded as 'post'. Those that conform to the *to* + ADV + INFINITIVE VERB pattern were coded as 'split'. Because I intend to analyze the patterns of modified infinitive variation using language-external and language-internal factors/predictors identified in the literature (see full list in section 3.3), each of the utterances was also coded for these factors using Natural Language Processing (NLP) tools in Python and in R.

For the predictors relevant to adverb type, I utilized the *grepl* function in base R. The package helped me classify modified infinitive constructions based on whether the

construction contains an adverb of a certain type (i.e. additive/restrictive, subjunct), using the adverb typology of Quirk *et al.* (1985) and Biber & Quirk (2012). I also used the same package to help me identify the adverbs that are very frequently used in PhE to split the infinitive, based on the PhE frequency list in Gonzales & Dita (2018). As for the factors related to number of syllables and prosody (e.g. stress), I utilized a prosodic parser from the *prosodic* package in Python (Heuser, Falk & Anttila 2010) to help me extract the stress patterns of the verbs and adverbs in the modified infinitives.[7] I was also able to measure the length of the adverbs and verbs with respect to number of syllables using the parser and another independent syllable counter in R – *quanteda* (Benoit *et al.* 2018). The predictor related to ambiguity and distance was derived by simply counting the number of words from the infinitive verb to the end of the sentence, minus the adverb.

Self-reported age and sex/gender data in the TCOPE cannot be directly extracted; however, as mentioned earlier, computational methods can be employed to acquire *stylistic/presented* age and gender based on Twitter profile information (Gonzales 2004). This is ideal because for my analysis, I will be viewing 'age' and 'sex/gender' as 'age presentation' and 'sex/gender presentation', aligning with the social constructionist framework. According to this perspective, gender and age are social constructs, meaning they are not inherent traits but rather emerge from how we express and perform them through various means like clothing, behavior and other aspects (Eckert 1989). This framework is well suited for analyzing Twitter data because social media serves as a significant platform for negotiating and constructing identities. Much of the stylized representation of gender and age occurs through the use of images and language on these platforms.

The tool I used is Wang *et al.*'s (2019) M3 demographic inference tool: it receives a Twitter identification number as input (available in the TCOPE), analyzes the profile image, username, screen name and biography of the Twitter user, and outputs the *stylistic/presented* age, sex and entity-type (i.e. organization, non-organization) of the user in terms of probability with relatively high precision and recall (Macro-F1: Gender = 0.918, Age = 0.522,[8] Entity-type = 0.898) (Wang *et al.* 2019; Gonzales 2004). The tool is computationally intensive, so I am unable to derive the demographic information for all extracted tokens that matched the RegEx patterns (n = 350,266). Therefore, I downsized the dataset through random sampling. Non-split utterances dominated the dataset and the utterances with split infinitives were few relative to the

[7] The parser is not variety-specific and does not consider some of the idiosyncratic prosodic patterns of PhE. However, preliminary research has shown that PhE stress conventions pattern very closely after Inner-Circle English stress patterns, with variation occurring in only a small subset of words (Shahruddin, Ao & Low 2022). For this reason, I decided to use the prosodic parser.

[8] The performance of the model in predicting age is suboptimal. Wang *et al.* (2019) did not provide specific recall and precision values, only presenting the F1 scores for age-related performance. Predicting *actual* or *self-identified* age in social media poses challenges due to various factors. Social media users often aim to present themselves as younger by using filters, wrinkle removal tools, or posting photos of younger celebrities they relate to. These practices can significantly hinder the accuracy of *actual* or *self-identified* age prediction in the analysis since the Deep Learning program relies on profile photos to estimate age likelihood. However, such a problem does not exist if one is interested in *stylistic/presented* age or gender.

total number of non-split utterances. So, the downsizing was done disproportionately to ensure that roughly 50 percent of the dataset contained split infinitives and 50 percent contained *to-phrases* that do not have an adverb within.

After acquiring the socio-demographic variables, I assessed and attempted to improve the quality of the data by removing tokens that were tweeted by organizations (e.g. corporate Twitter data) and tokens where the adverb is *too*, which is categorically placed after the infinitive in English. Retweets and duplicate tweets were also removed. Furthermore, my research assistant and I went over the dataset and filtered out utterances that were (i) mistakenly extracted by the RegEx algorithm (e.g. *I am trying* ***to demonstrate*** *that someone is* ***always*** *listening*. <COPE-TW-CEB-2020-06:232146>), (ii) contained idiomatic expressions (e.g. *to get even*) and (iii) involved verbs and adverbs in indigenous Philippine languages (e.g. *na* 'already', *din* 'also'). The original dataset contained 10,240 utterances, but after the pre-processing and careful filtering, only 7,958 tokens remained (*to* + ADV + INFINITIVE VERB = 4,425, ADV + *to* + INFINITIVE VERB = 1,927, *to* + INFINITIVE VERB (+ …) + ADV = 1,606).

3.3 *Data analysis*

In this article, I analyze and describe patterns of variation hoping to gain insights on Twitter-style PhE only. As mentioned earlier, Twitter-style PhE only constitutes a part of PhE, so I try to avoid generalizing about PhE using TCOPE data whenever possible. A conservative approach seemed to be the best way forward given that language on Twitter sometimes diverges from language in other communication platforms, partially due to platform-specific features such as character count or increased use of emojis and other multimodal resources (Tagliamonte & Denis 2008; Davies & Fuchs 2015; Jenkins 2015; Bohmann 2016).

Using the downsized and carefully curated dataset, I analyzed the variation using a multivariate Bayesian logistic regression analysis, in line with contemporary research on variationist and general sociolinguistics work (Vasishth *et al.* 2018; MacKenzie 2020). Specifically, I ran a mixed-effects multinomial model, as the syntactic variable under study has three variants (Levshina 2016). I ran the Markov chain Monte Carlo (MCMC) algorithm (Franke & Roettger 2019; Makowski *et al.* 2019; McElreath 2020) with the *brms* package in the R environment (R Core Team 2015; Bürkner 2017;). The multinomial regression model was fitted over 7,958 observations, with 30,000 iterations per chain. A total of 4 Markov chains were sampled. I also used a warm-up or burn-in period of 15,000 iterations for each chain to correct for initial sampling bias. The thinning parameter was set at 2. Weakly informative priors (i.e. normal distribution [0, 5]) were set for the intercept and slopes (Levshina 2016: 252). The choice of priors, at least the ones I tested (i.e. uniform, Cauchy [0,5]), did not influence the posteriors significantly. Following Vehtari *et al.* (2021: 683), I also monitored the $\hat{R}$ values and Effective Sample Size (ESS) to verify convergence. To ensure proper convergence, I made sure that the $\hat{R}$ value stayed within the range of 1.01 and that the ESS value remained above 400 (see model in section 4 for $\hat{R}$ and ESS values).

I did not conduct a frequentist-oriented regression because although Bayesian procedures are computationally cost-intensive and show similar results as their frequentist counterparts (i.e. models with *p*-values), the Bayesian method allows me to interpret my results intuitively in terms of probability (Levshina 2022). It also permits comments on the absence of an effect, whereas this is impossible within the frequentist framework (Vasishth & Nicenboim 2016; McElreath 2020).

The multinomial logistic regression model contains the following language-external and language-internal predictors, selected based on the literature on split infinitive use (see section 2). Multilevel categorical variables (e.g. island group) were coded using Weighted Helmert coding conventions (Sonderegger 2022). Random intercepts for verb lexeme, adverb lexeme and user were modeled in. However, following Levshina (2016: 253), only lexemes and users with adequate tokens (i.e. tokens > 5) were considered as individual factor values. All lexemes and users with less than five utterances/tokens were conflated in the 'other' category.

Language-internal predictors

- Adverb type
  - additive-restrictive vs. others
  - subjunct vs. non-subjunct
  - has *-ly* vs. others
- Reported frequency of adverb as a splitter in ICE (high vs. low)
- Length of adverb relative to the verb (syllable) – continuous
- Distance of the verb to the sentence-final boundary (degree/likelihood of potential ambiguity) – continuous
- Interaction of length and distance
- Stress and rhythm Conditions A, B, C and D (adherence/yes vs. non-adherence/no)
- Interaction of Condition D with Conditions A, B and C

Language-external factors

- Year – continuous
- Island group (Visayas vs. Luzon)
- Island group (Mindanao vs. Visayas and Luzon)
- City (Manila vs. non-Manila)
- Age presentation – continuous
- Sex/gender presentation (male-presenting vs. female-presenting)
- Interaction of age presentation and sex/gender presentation

Random intercepts

- Verb lexeme
- Adverb lexeme
- User

After obtaining the results from the summary of Bayesian posterior draws (Bürkner 2017), I identify predictors that have an effect on modified infinitive variation using the probability of direction (*pd*) measure. A predictor is said to have an effect on the dependent variable (e.g. modified infinitive variants) if it has a median value that is far away from zero or if the credibility intervals surrounding the median do not contain zero (Levshina 2016; Grafmiller, Szmrecsanyi & Hinrichs 2018; Makowski *et al.* 2019; MacKenzie 2020). The Bayesian statistical measure 'probability of direction' (*pd*) – or the proportion of posterior draws that is of the median's sign – will be used in this article to characterize the (un)certainty of existence of the effect. A higher *pd* (i.e. close to 1) indicates higher certainty that the positive or negative effect indicated in the median is present, whereas a lower *pd* (i.e. close to 0.5) indicates that the negative or positive effect could be non-existent (Makowski *et al.* 2019). For example, if the median of a predictor in this paper's model is -1.2 and its *pd* is 0.95, one can say that the proportion of posterior values that are *less* than zero is 95 percent and that only 5 percent of the values are greater than zero (Levshina 2016). In other words, there is a 95 percent chance that that predictor will have a *negative* effect on (or decrease) the likelihood to choose the pre-modification or post-modification strategy over the split infinitive strategy, as the reference category has been set to 'split'.

4 Results and discussion

4.1 Multinomial model general results

Table 2 provides the posterior median values, standard deviation, 89 percent credible intervals based on Highest Density Interval (HDI) boundaries, along with the *pd* values (see section 3.3). It also includes diagnostic measures for convergence (i.e. $\hat{R}$ and ESS). The table only shows the fixed effects and some random effects (i.e. intercepts) due to space constraints. The full results can be accessed online via the Open Science Framework (OSF): https://osf.io/sr2cy/

The model shows that many of the predictors have a high probability of influencing modified infinitive variation (table 3). It indicates that modified infinitive variation is highly sensitive to many of the language-internal and language-external factors included in the model.

4.2 Language-internal factors

4.2.1 Non-prosodic

The multinomial model of modified infinitive variation indicates high probabilities of non-zero effects of adverb type on variation. The results, for one, show that users are more likely to avoid the split infinitive strategy in favor of *pre*-infinitive modification (henceforth, pre-modification) strategy if the adverb modifier is an additive or restrictive adverb ($pd = 0.87$) (table 2, figure 3a). However, there seems to be an opposite effect of this adverb type variable on the likelihood to select the split

Table 2. *Bayesian model posterior draw estimates for predictors influencing likelihood to split the infinitive (reference levels in boldface)*

(a) *to* + ADV + INFINITIVE VERB (split infinitive) vs. ADV + *to* + INFINITIVE VERB construction (pre-modification) (reference category = **split infinitive**)

Predictors	Median	SD	89% CI (HDI)	pd	R̂	ESS
Fixed effects						
Pre_Intercept	89.88	33.36	36.66 to 142.94	1	1	24815
Pre_Adverb type (additive/restrictive vs. **others**)	1.36	1.18	−0.47 to 3.29	0.87	1	14985
Pre_Adverb type (subjunct vs. **non-subjunct**)	−2.09	0.77	−3.34 to −0.89	1	1	9427
Pre_Adverb type (-*ly* vs. **non-*ly***)	−6.04	0.94	−7.58 to −4.58	1	1	11740
Pre_Reported frequency of adverb as a splitter in ICE (high vs. **low**)	−2.42	1.25	−4.33 to −0.39	0.98	1	12107
Pre_Length of adverb relative to the verb (syllable)	−1.18	0.15	−1.42 to −0.93	1	1	18410
Pre_Likelihood of ambiguity (Distance from end of sentence)	−0.03	0.01	−0.05 to −0.02	1	1	24694
Pre_Stress and Rhythm Condition A – Monosyllabic verb (yes vs. **no**)	1.77	0.28	1.3 to 2.2	1	1	19014
Pre_Stress and Rhythm Condition B – Finally stressed disyllabic verb + monosyllabic or disyllabic adverbs (yes vs. **no**)	0.74	0.26	0.34 to 1.16	1	1	22735
Pre_Stress and Rhythm Condition C – Initially stressed disyllabic verb + trisyllabic adverb (yes vs. **no**)	−0.7	0.97	−2.32 to 0.71	0.79	1	23720
Pre_Stress and Rhythm Condition D – Series of stressed or unstressed syllables (yes vs. **no**)	−0.64	0.21	−0.98 to −0.32	1	1	21707
Pre_Length of adverb relative to the verb : Likelihood of ambiguity	0.01	0.01	0 to 0.03	0.95	1	25198
Pre_Condition A : Condition D	−0.06	0.35	−0.61 to 0.5	0.57	1	23456
Pre_Condition B : Condition D	1.77	0.56	0.9 to 2.69	1	1	22472
Pre_Condition C : Condition D	0.83	1.87	−2.03 to 3.84	0.68	1	23927
Pre_Island Group (Visayas vs. **Luzon**)	−0.09	0.11	−0.26 to 0.1	0.78	1	23039
Pre_Island Group (Mindanao vs. **Visayas and Luzon**)	0.35	0.12	0.17 to 0.54	1	1	23762
Pre_City (Manila vs. **non-Manila**)	−0.15	0.15	−0.39 to 0.09	0.85	1	22396
Pre_Gender presentation (**male-presenting** vs. female-presenting)	0.24	0.29	−0.23 to 0.71	0.8	1	19636
Pre_Age presentation	−0.02	0.01	−0.03 to −0.01	1	1	23493

(*Continued*)

Table 2. (*continued*)

(a) *to* + ADV + INFINITIVE VERB (split infinitive) vs. ADV + *to* + INFINITIVE VERB construction (pre-modification) (reference category = **split infinitive**)

Predictors	Median	SD	89% CI (HDI)	pd	R̂	ESS
Pre_Year	−0.04	0.02	−0.07 to −0.02	1	1	24847
Pre_ Gender presentation : Age presentation	−0.01	0.01	−0.03 to 0.01	0.87	1	19528
Random effects (Intercepts only)						
Pre_Adverb Lexeme (Intercept, SD)	3.22	0.38	2.64 to 3.84	1	1	10932
Pre_Verb Lexeme (Intercept, SD)	1.09	0.11	0.92 to 1.26	1	1	14944
Pre_User (Intercept, SD)	0.97	0.43	0.24 to 1.65	1	1	10398

(b) *to* + ADV + INFINITIVE VERB (split infinitive) vs. *to* + INFINITIVE VERB (+ …) + ADV construction (post-modification) (reference category = **split infinitive**)

Predictors	Median	SD	89% CI (HDI)	*pd*	$\hat{R}$	ESS
Fixed effects						
Post_Intercept	102.64	39.66	40.1 to 166.17	0.99	1	25132
Post_Adverb type (additive/restrictive vs. **others**)	−1.32	1.18	−3.21 to 0.54	0.88	1	13460
Post_Adverb type (subjunct vs. **non-subjunct**)	−1.91	0.59	−2.84 to −0.96	1	1	10816
Post_Adverb type (-*ly* vs. **non-*ly***)	−4.37	0.65	−5.4 to −3.33	1	1	10339
Post_Reported frequency of adverb as a splitter in ICE (high vs. **low**)	−0.4	0.91	−1.86 to 1.05	0.67	1	13253
Post_Length of adverb relative to the verb (syllable)	−0.58	0.12	−0.77 to −0.39	1	1	20285
Post_Likelihood of ambiguity (Distance from end of sentence)	−0.07	0.01	−0.09 to −0.05	1	1	23557
Post_Stress and Rhythm Condition A – Monosyllabic verb (yes vs. **no**)	1.51	0.26	1.11 to 1.92	1	1	19962
Post_Stress and Rhythm Condition B – Finally stressed disyllabic verb + monosyllabic or disyllabic adverbs (yes vs. **no**)	0	0.31	−0.49 to 0.49	0.5	1	24029
Post_Stress and Rhythm Condition C – Initially stressed disyllabic verb + trisyllabic adverb (yes vs. **no**)	0.4	0.6	−0.55 to 1.35	0.74	1	24642

(*Continued*)

Table 2. (*continued*)

(b) *to* + ADV + INFINITIVE VERB (split infinitive) vs. *to* + INFINITIVE VERB (+ …) + ADV construction (post-modification) (reference category = **split infinitive**)

Predictors	Median	SD	89% CI (HDI)	*pd*	$\hat{R}$	ESS
Post_Stress and Rhythm Condition D – Series of stressed or unstressed syllables (yes vs. **no**)	−0.29	0.18	−0.58 to 0.01	0.94	1	21761
Post_Length of adverb relative to the verb : Likelihood of ambiguity	0.01	0.01	0 to 0.02	0.87	1	23592
Post_Condition A : Condition D	0.24	0.38	−0.36 to 0.85	0.73	1	23393
Post_Condition B : Condition D	0.66	0.63	−0.35 to 1.66	0.85	1	23150
Post_Condition C : Condition D	0.95	1.2	−1 to 2.82	0.79	1	24463
Post_Island Group (Visayas vs. **Luzon**)	0.06	0.14	−0.15 to 0.28	0.68	1	24269
Post_Island Group (Mindanao vs. **Visayas and Luzon**)	0.46	0.14	0.25 to 0.68	1	1	23788
Post_City (Manila vs. **non-Manila**)	−0.11	0.18	−0.38 to 0.19	0.73	1	23082
Post_Gender presentation (**male-presenting** vs. female-presenting)	0.53	0.35	−0.03 to 1.09	0.93	1	19515
Post_Age presentation	−0.02	0.01	−0.03 to −0.01	0.99	1	23405
Post_Year	−0.05	0.02	−0.08 to −0.02	0.99	1	25124
Post_Gender presentation : Age presentation	−0.02	0.01	−0.04 to 0.01	0.87	1	19476
Random effects (Intercepts only)						
Post_Adverb Lexeme (Intercept, SD)	2.61	0.28	2.18 to 3.07	1	1	11685
Post_Verb Lexeme (Intercept, SD)	0.66	0.1	0.51 to 0.81	1	1	15158
Post_User (Intercept, SD)	0.41	0.34	0 to 0.92	1	1	16483

Table 3. *Proportion of factors that have a higher probability (pd > 0.79) of a non-zero effect on the likelihood of use of pre- or post-modified infinitive over the split infinitive construction*

	Pre-modification vs. split infinitive		Post-modification vs. split infinitive		
Type of variable	variables with *pd* > 0.79	% out of total	variables with *pd* > 0.79	% out of total	Total variables
Fixed only	17	80.95	14	66.67	21
Fixed and random	176	46.07	126	32.98	382

infinitive strategy over the *post*-infinitive modification (henceforth, post-modification) strategy. Users tend to prefer the split infinitive construction over post-modification when the adverb is additive or restrictive. They tend to place the adverb after the infinitive when the modified infinitive construction contains other adverb types such as adverbs of place ($pd = 0.88$) (table 2, figure 3a). The results all in all only partially confirm the findings of Kostadinova (2020) for AmE, which found the general avoidance of the use of the split infinitive construction with additive or restrictive adverbs. The differences in results are not surprising as Kostadinova studied modified infinitives in AmE across different styles (e.g. spoken, fiction, newspaper) whereas I focused on Twitter-style PhE. The comparison suggests that Twitter-style PhE has some characteristics of AmE, but that it also has its idiosyncrasies.

In addition, I also found that the likelihood of use of the split infinitive over the other two constructions is also sensitive to the presence of adverbs identified as 'subjuncts' according to Quirk *et al.*'s (1985) description of English (table 2, figure 3b). Modified infinitives with subjuncts tend to be realized as *to* + ADV + INFINITIVE VERB whereas those without tend to be realized as *to* + INFINITIVE VERB (+ …) + ADV or ADV + *to* + INFINITIVE VERB ($pd = 1$). The results are consistent with previous research on AmE, which highlight the important role of adverb subjunct status in conditioning the variation of the modified infinitive (Quirk *et al.* 1985; Perales-Escudero 2011: 331). In addition to the previous finding, the results here provide some indication that the conventions of Twitter-style PhE and contemporary AmE overlap, as the linguistic constraints governing the structure of modified infinitives in these varieties are similar.

A supplementary finding with respect to the adverb types is that adverbs with a *-ly* suffix (e.g. *evenly*) seem to favor the split infinitive construction whereas those without (e.g. *again*) tend to favor either the pre- ($pd = 1$) or post-modification constructions ($pd = 1$) (table 2, figure 3c). This finding is novel, as the effect of the *-ly* suffix on this variation has not been formally explored in previous literature. The results suggest, for Twitter-style PhE at least, that the suffix *-ly* should also be considered in explorations involving the modified infinitive.

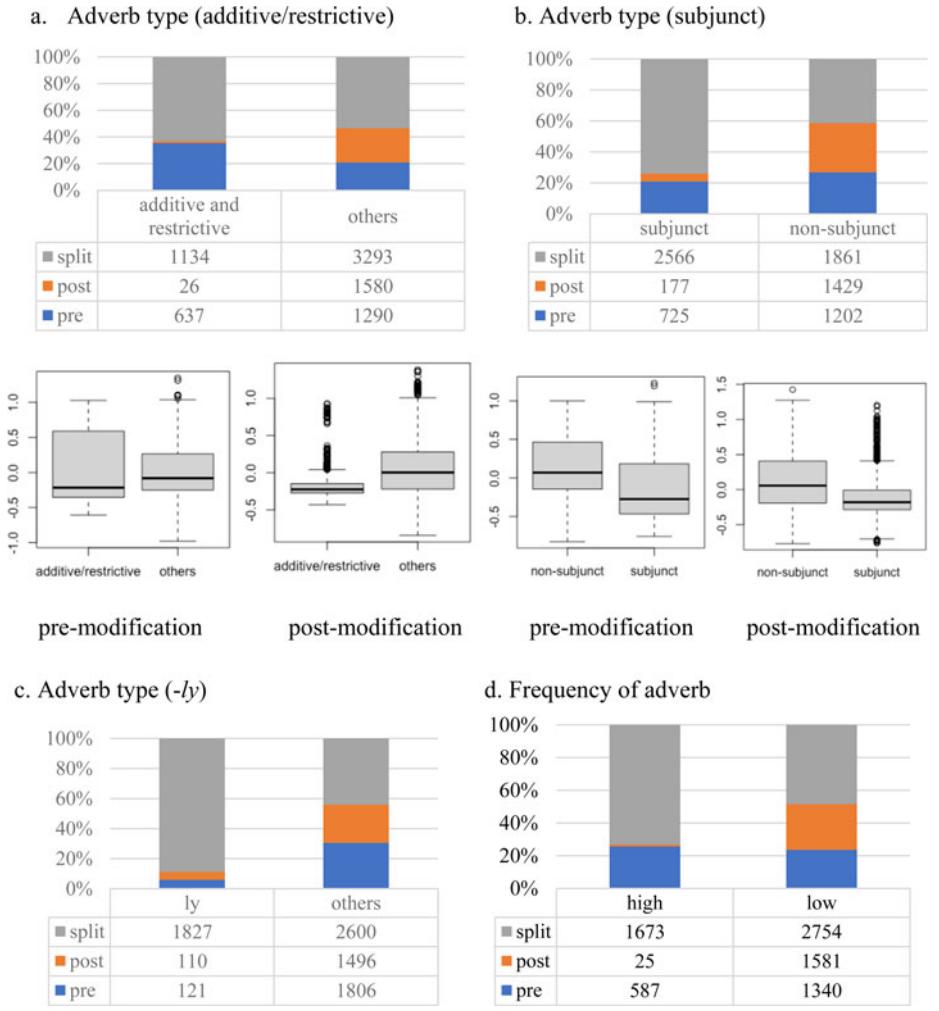
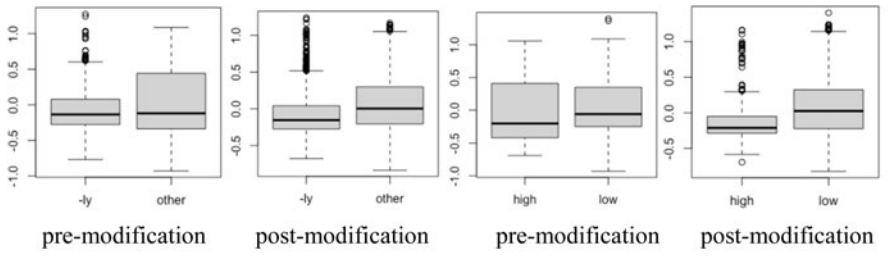
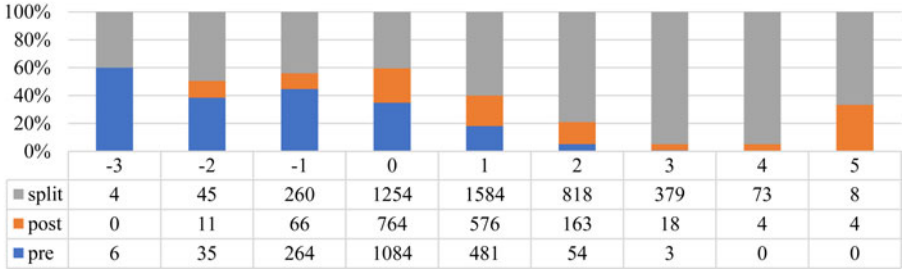


Figure 3. Distribution of modified infinitive variants by notable non-prosodic language-internal factors. Box plots and scatter plots indicate partial effects. On the y-axes of the box and scatter plots, negative values indicate likelihood to use a split infinitive, whereas positive values indicate likelihood to use pre-/post-infinitive modification.

Apart from the type of the adverb, I also found that the reported frequency of the adverb as a splitter in another PhE corpus conditions one's choice to use the split infinitive construction over the pre-modified construction ($pd = 0.98$) and the post-modified construction ($pd = 0.67$). Twitter users of English in the Philippines tend to split the infinitive if the adverb belongs to the list of most frequently used infinitive adverb splitters in the Philippine component of the ICE (see section 2) (table 2, figure 3d) (Gonzales & Dita 2018). In other words, the choice of adverb splitters in the ICE – which was sampled across genres – mirrors the choice of adverb splitters in the



e. Length of adverb relative to the verb



	-3	-2	-1	0	1	2	3	4	5
split	4	45	260	1254	1584	818	379	73	8
post	0	11	66	764	576	163	18	4	4
pre	6	35	264	1084	481	54	3	0	0

Figure 3. *Continued*

TCOPE. The finding suggests that users of PhE (e.g. Twitter users, users of spoken and written PhE etc.) generally know which adverbs are more likely to appear in split infinitive constructions and which adverbs tend not to be used in the split construction in PhE. It provides some evidence of conventionalized dialect-specific linguistic practices in PhE, which seem to condition the variation observed partially.

The current study is perhaps the first to find a clear effect of length of adverb (measured via number of syllables) on the likelihood to choose the splitting strategy over pre- ($pd = 1$) and post-modification ($pd = 1$) (table 2, figure 3e). There is a higher chance of syntactic tmesis in constructions where the adverb is longer than the verb (e.g. *to **actually** say*) compared to constructions where the adverbs and verbs are of equal length (e.g. *to **just** see*) or constructions where the verb is longer than the adverb (e.g. *to **only** actualize*). This finding is at odds with previous work, which found an effect in the opposite direction: syntactic tmesis was more likely to occur in utterances where the adverb is shorter relative to the verb or in utterances where the adverb and verb have the same number of syllables (Kostadinova 2020). This is a previously undocumented finding. However, when taken with the fact that Kostadinova's finding is on AmE, the findings are consistent with the literature on nativized varieties and contact languages, which have generally found that contact varieties – as independent entities – do not necessarily follow the developmental trajectory of their source languages (Thomason 2001; Thomason 2007; Gonzales *et al.* 2022). It is still not immediately clear why the effect of adverb length should necessarily go in this direction in Twitter-style PhE. Is the effect another indication of variety-specific norms, or is it something else? What is clear, however, is that adverb length conditions the

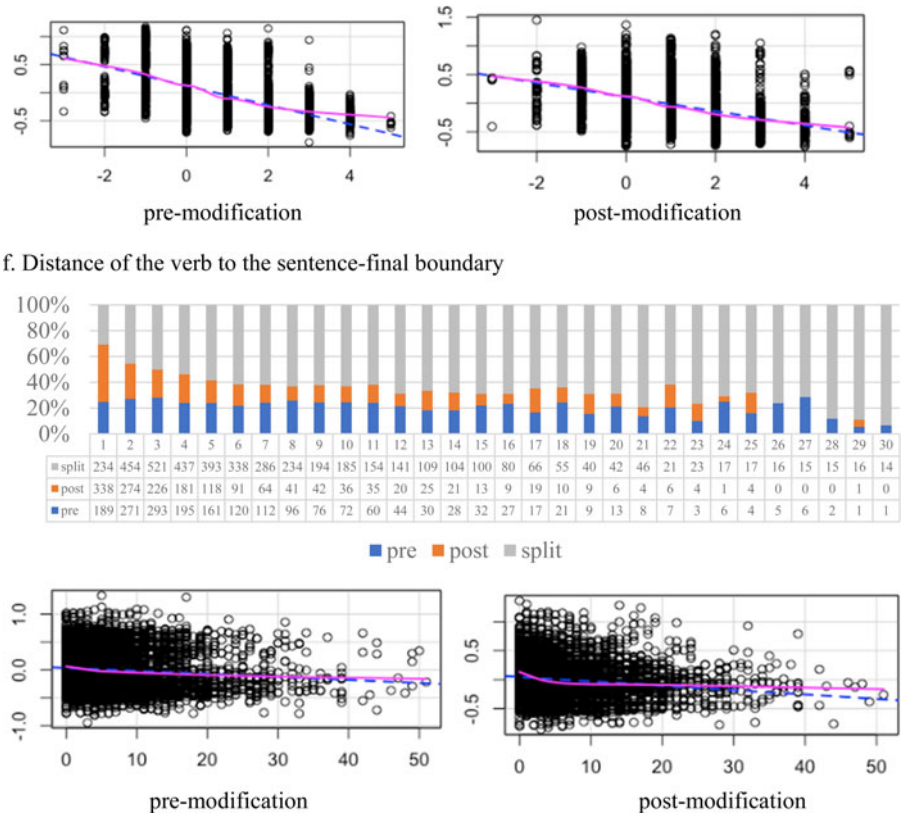


Figure 3. *Continued*

choice of modified infinitive construction, and the robust effect suggests that future work on the variation of this construction must include this predictor to fully account for the variation.

The distance between the *to*-phrase and the end of the sentence – a correlate of likelihood of ambiguity – play a role in conditioning the choice of modified infinitive construction ($pd = 1$). Utterances where the *to*-phrase is near the sentence-final boundary (i.e. utterances that leave less room for ambiguity) disfavor the use of the split infinitive, whereas utterances where the *to*-phrase is succeeded by constituents such as noun phrases (e.g. verbal complements) tend to favor the split construction (table 2, figure 3f). Under the assumption that the distance measure is reflective of (likelihood of) ambiguity, the finding is compatible with Quirk *et al.*'s (1985) description of split infinitives as an ambiguity resolution device (see section 2). Their findings suggest that if there is less ambiguity (little to no distance between to-phrase and the end of the sentence), there is less of a need to split the infinitive, which seems to be exactly what we observed. Of course, the distance measure is by no means a direct measure of ambiguity. However, I hope that the results still hold some value, as

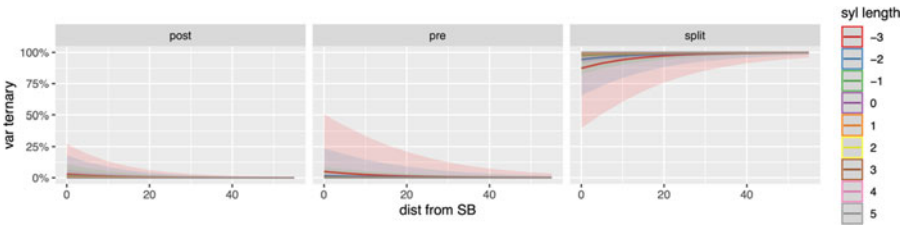


Figure 4. Marginal effects of interaction between length of adverb relative to the verb (length in syllables) and likelihood of ambiguity (distance from sentence boundary)

they present interesting correlations between split infinitive use and distance of the *to*-phrase to sentence-final boundary. They could pave the way for future research that directly analyzes ambiguity presence as a possible constraint on the use of split infinitives.

Length of adverb interacts with distance of the *to*-phrase to the sentence boundary to condition the choice to select the splitting strategy over the pre-modification strategy ($pd = 0.95$) and the post-modification strategy ($pd = 0.87$). The effect of distance is dependent on the length of the adverb (table 2, figure 4): the distance/ambiguity effect is non-existent in utterances where the adverb is longer than the infinitive verb because the split infinitive construction is almost always used. But this effect gradually becomes more salient as the adverb shortens relative to the verb. In short, the distance effect is most salient when the utterance has an adverb that is significantly shorter than the verb. I currently do not have a convincing explanation for this phenomenon, but the pattern warrants further investigation in the future.

The model also indicates high probabilities ($pd = 1$) for the conditioning effect of verb and adverb lexeme on modified infinitive variation. The verb and adverb lexeme used govern the choice to split, to pre-modify, or to post-modify the infinitive. For example, modified infinitive constructions that contain adverbs like *automatically*, *barely*, *drastically* and *possibly* and verbs like *affect*, *appear* and *consider* are always realized as the split infinitive construction, whereas constructions that contain adverbs like *enough*, *here*, *together* and *twice* and verbs like *rise*, *cheer* and *sing* tend to be realized as the pre- and/or post-modified construction. The findings point to the importance of including lexeme-related factors in models of sociolinguistic variation, as these factors could have variable effects on linguistic behavior, as demonstrated here. The full distribution of adverb and verb lexemes by modified infinitive variant can be accessed online via the Open Science Framework (OSF): https://osf.io/sr2cy/

4.2.2 Prosodic

The probabilities for three of the four factors relevant to stress and rhythm – Conditions A, B and D – to influence modified infinitive variation are high. The odds of opting for a split construction over pre-modification and post-modification decrease if the verb is monosyllabic, regardless of the number of syllables in the adverb (Condition A)

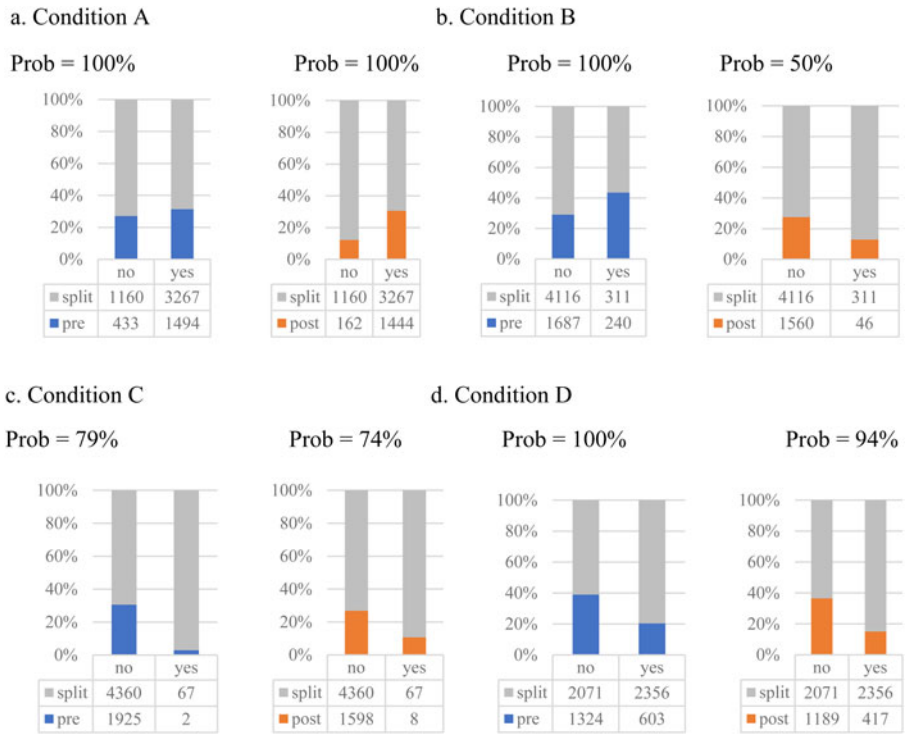


Figure 5. Distribution of modified infinitive variants by prosodic language-internal factors (main effects)

($pd = 1$) (table 2, figure 5a, example 2). While the existence of the effect is expected, the direction of the observed effect diverge from what has been documented (Calle-Martín & Miranda-García 2009): the likelihood to split the infinitive appears to have decreased instead of increased when Condition A is met.

(2) S S S (Cond. A counter)
I can't wait to see you ***soon***..
I can't wait to VERB you ADV
<COPE-TW-TAG-2017-05:195639>

A similar effect was observed when analyzing utterances using Condition B, but the effect is only highly probable for the comparisons involving pre-modification ($pd = 1$) and not post-modification ($pd = 0.5$). The rates of pre-modification are higher in constructions with a finally stressed disyllabic verb and adverb with fewer than three syllables (Condition B) compared to constructions that do not meet this condition, contra Calle-Martín & Miranda-García (2009) (table 2, figure 5b). One would find more examples like (3) in Condition B.

(3) S U-S (Cond. B counter)
Just *to* *ar-rive at TMC on time.*
ADV to VERB at TMC on time
<COPE-TW-MNL-2021-07:169842>

Effects of Condition C on modified infinitive variation have been found in the model and are consistent with Calle-Martín & Miranda-García's proposal. Modified infinitives are realized as split infinitives more in constructions with initially stressed disyllabic verbs and trisyllabic adverbs (Condition C) compared to constructions without them. However, the probability of such an effect on the likelihood to pre-modify ($pd = 0.79$) or post-modify ($pd = 0.74$) over splitting is lower than the effect probabilities of Conditions A and B.

(3) U-S-U S-U
Enabling *me* *to* *profoundly* *balance ...*
enable.PROG 1.SG to ADV VERB
<COPE-TW-BAG-2019-04:1377>

The effects of Condition D on the likelihood to pre-modify ($pd = 1$) or post-modify ($pd = 0.94$) instead of splitting are highly probable based on the model. They are also consistent with the literature (Crystal 1984; Calle-Martín & Miranda-García 2009) (table 4). Users tends to opt for the split infinitive constructions if the adverb splitter contributes to the 'natural' *te-tum te-tum* rhythm (table 2, figure 5d, examples 4 and 5). This is consistent with what Crystal and Calle-Martín & Miranda-García described (table 4).

(4) S U-S (Cond. D, no split to avoid S-S series)
its *time* *to* *sleep* ***a-gain***
it.is time to VERB ADV
<COPE-TW-GEN-2017-05:176992>

?(5) U-S S (the structure being avoided)
its *time* *to* ***a-gain*** *sleep*
it.is time to ADV VERB

The unexpected effects related to Conditions A and B in this study indicate that these conditions do indeed have influence on patterns of modified infinitive variation, but perhaps not to achieve 'natural' rhythm, as Calle-Martín & Miranda-García (2009) have argued. This is best supported by an examination of the interaction effect observed between Conditions B and D, which showed that users tend to *avoid* splitting infinitives in Condition B constructions if the adverb splitter contributes to 'natural' rhythm (table 2, figure 6c and d) ($pd_{pre} = 1$, $pd_{post} = 0.85$) – the opposite of what we would expect. There was also no strong evidence that users tend to choose split infinitives over pre- ($pd = 0.57$) or post-modification ($pd = 0.74$) in Condition A

constructions as a means to follow this natural rhythm (figure 6a and b). So, while Calle-Martín & Miranda-García's argument that Conditions A and B play a role in conditioning modified infinitive variation is supported by evidence, their argument that these two conditions stem out of an impulse to maintain 'natural' rhythm appears questionable.

Their claim that variation patterns are sensitive to Condition C due to rhythm maintenance also appears to be untenable. My findings show that speakers tend to split the infinitive instead of pre- ($pd = 0.68$) and post-modifying it ($pd = 0.79$) more in Condition C situations where the adverb splitter *disrupts* the 'natural' rhythm (figure 6e) compared to situations where the splitter contributes to the said rhythm (figure 6f).

One possible reason for the inconsistencies between the findings of this study and that of Calle-Martín & Miranda-García might have to do with the differences in the datasets, i.e. stylistic and dialectal differences between Twitter-style PhE and written and spoken AmE. However, the discrepancies might also have to do with the way the data were analyzed. Their study analyzed data using frequencies only, whereas this study incorporated a multiple multinomial statistical method that addresses potential confounding effects. It is impossible to say for certain whether identical patterns and effects would have emerged in Calle-Martín & Miranda-García's study, had the data been analyzed with the same model. What is certain, however, is that there indeed are effects of 'natural' rhythm (Condition D) and other stress/rhythm conditions (Conditions A, B) on modified infinitive syntactic variation and that these effects appear to be independent of each other. The findings altogether point to the crucial role of stress and rhythm in the structure of modified infinitives in English.

4.3 *Language-external factors*

Almost all of the language-external factors in the model have a high probability of conditioning the variation in modified infinitives. As table 2 and figure 7 show, these factors simultaneously influence the likelihood of split infinitive use. The only factor that had a relatively lower chance to condition the selection of the split infinitive strategy over pre- *and* post-modification is Island Group (Visayas vs. Luzon) ($pd_{pre} = 0.78$, $pd_{post} = 0.68$). The rates of split infinitive use between regions do not exhibit notable differences (table 2 and figure 7a), indicating a higher likelihood of homogeneity of split infinitive patterns in the English used in these regions.

There is, however, a high chance that region conditions the variation when comparing the rates of split infinitive use in Mindanao (Southern Philippines) compared to the rates in Visayas (Central Philippines) and Luzon (Northern Philippines). The numbers show that Twitter users in the Mindanao region tend to have significantly lower rates of split infinitive use compared to those in the other two island groups ($pd = 1$) (table 2 and figure 7a). In other words, Mindanao users tend to be more conservative in their use of modified infinitives than Luzon and Visayas users. Due to lack of evidence, I am unable to explain this pattern, but future work can investigate the possible reasons leading to it (e.g. regional identity, variable prescriptive censure policies across

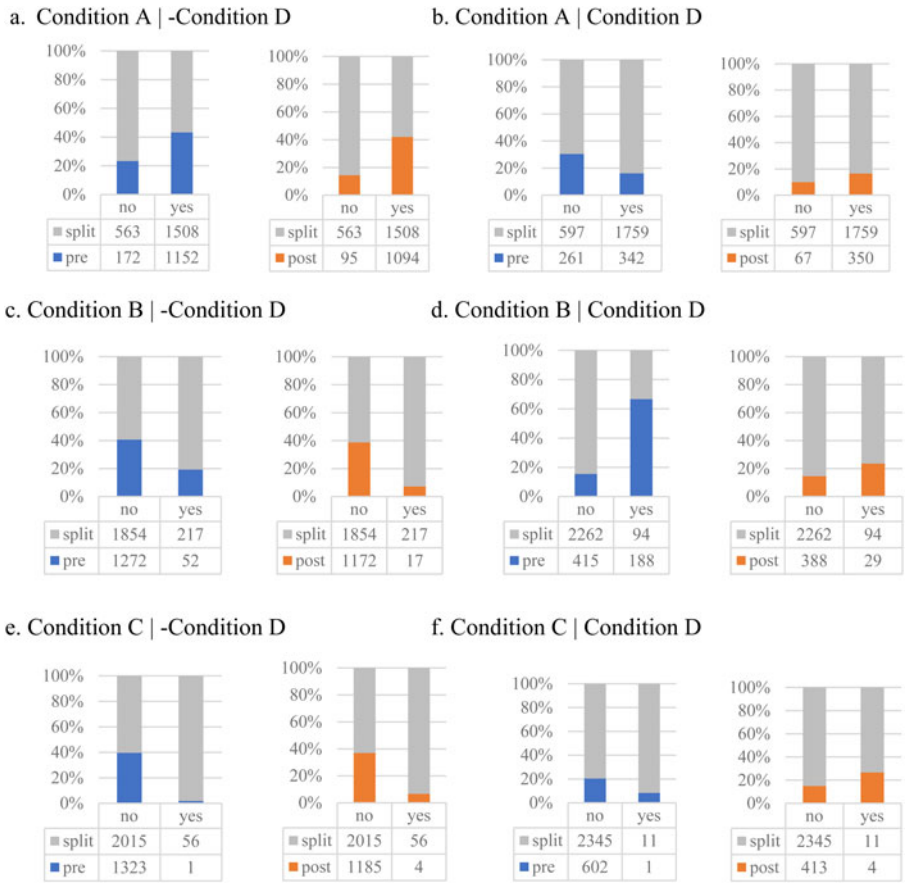


Figure 6. Distribution of modified infinitive variants – Conditions A, B and C given Condition D

regions). From a geographical perspective, the findings are, to my knowledge, perhaps among the first pieces of evidence for a north/central vs. south distinction in the PhE context.

A closer look at the data would also show the effect of city on variation. Those in Manila – the capital of the Philippines located in Luzon – tend to be more innovative and are more likely to use the split infinitive construction over the *pre*-modified construction ($pd = 0.85$) compared to non-residents of Manila (table 2 and figure 7b). Compared to non-residents, these users also tend to select the split infinitive construction over the *post*-modified construction more, but the likelihood of them doing so is noticeably lower ($pd = 0.73$). Altogether, the findings relevant to geography indicate that internal variation in Twitter-style PhE is very much present, and that the variation is constrained in part by geography. The findings corroborate previous work, which also found patterns of internal variation in (Twitter-style) PhE being sensitive to geographical location (Villanueva 2016; Gonzales 2017; Lee & Borlongan 2022; Gonzales 2023b).

Table 4. *Observed effects of stress and rhythm conditions on modified infinitive variation in previous studies vs. this study*

Condition	Crystal (1984)	Calle-Martín & Miranda-García (2009)	Perales-Escudero (2011)	This study
A monosyllabic verb, regardless of number of adverb syllables	not mentioned	+ split	not mentioned	– split
B Finally stressed disyllabic verb, with monosyllabic and disyllabic adverbs	not mentioned	+ split	not mentioned	– split (pre-modification only) – split (if contributes to 'natural' rhythm)
C Initially stressed disyllabic verb, with trisyllabic adverbs	not mentioned	+ split	not mentioned	+ split?
D If adverb splitter contributes to 'natural' rhythm	+ split	+ split	+ split (limited)	+ split

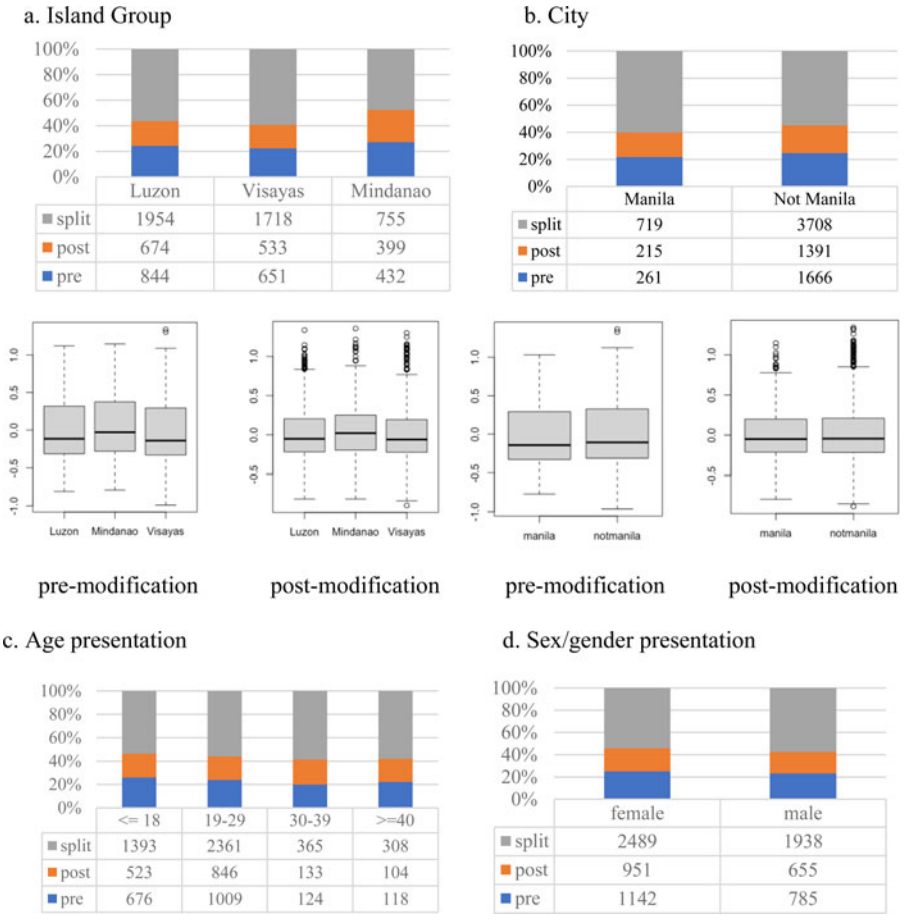
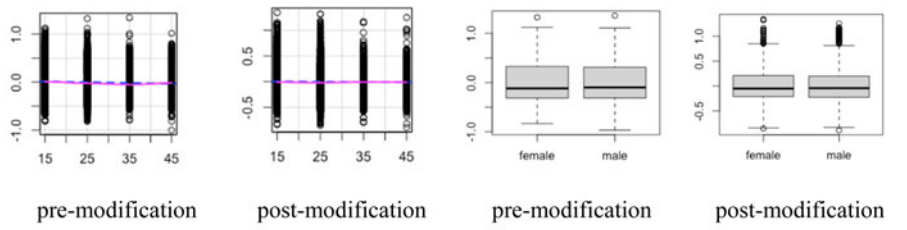


Figure 7. Distribution of modified infinitive variants by language-external factors (main effects); Box plots and scatter plots indicate partial effects. On the y-axes of the box and scatter plots, negative values indicate likelihood to use a split infinitive, whereas positive values indicate likelihood to use pre-/post-infinitive modification.

The results related to age presentation show that those who present as young (or rather, those are classified as younger by the Deep Learning algorithm) had splitting rates that are lower than those presenting as older ($pd_{pre} = 1$, $pd_{post} = 0.99$) (table 2 and figure 7c). The results on gender presentation, on the other hand, indicate a high probability that those who present as female (or classified as female) avoid splitting the infinitive over pre-modifying ($pd = 0.8$) or-postmodifying the infinitive ($pd = 0.93$) (table 2 and figure 7d). Furthermore, interestingly, individuals who present as younger and female are additionally less likely to split the infinitive compared to the rest of the population (e.g. those presenting as older males, younger males and older females) ($pd_{pre} = 0.87$, $pd_{post} = 0.87$) (table 2 and figure 7e). From a perspective of indexicality (Eckert & Rickford 2001), it is possible that the non-use of the split infinitive – the conservative



e. Age presentation | Female-presenting

	<= 18	19-29	30-39	>=40
split	809	1464	171	45
post	342	529	60	20
pre	433	631	56	22

f. Age presentation | Male-presenting

	<= 18	19-29	30-39	>=40
split	584	897	194	263
post	181	317	73	84
pre	243	378	68	96

f. Year

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
split	2	95	198	378	377	336	365	409	548	640	628	451
post	5	51	117	144	168	182	167	168	174	170	148	112
pre	3	49	119	182	188	234	179	185	231	229	190	138

f. Individual users

Figure 7. *Continued*

modified split infinitive variant – has acquired some indexical value related to youth and femininity due to the observed associations between young- and female-presenting speakers and the conservative variant. However, I hesitate to commit to this interpretation with certainty unless more evidence of this link can be gathered. We can, however, concretely observe a change in apparent time (Sankoff 2006), meaning a generational change, such that younger-presenting users of Twitter tend to use fewer split infinitive constructions compared to older-presenting users. The results suggest that younger-presenting speakers – those presenting as younger women in particular – are leading the trend towards conservative norms in Twitter-style PhE. One thing worth noting is that, if we apply Labov's (1994) generalizations for language change and gender to this situation (i.e. that young women tend to lead language change or innovation in variable features that society is not conscious of) (Meyerhoff 2018), it can be argued that the innovation in PhE is not the split infinitive construction, but

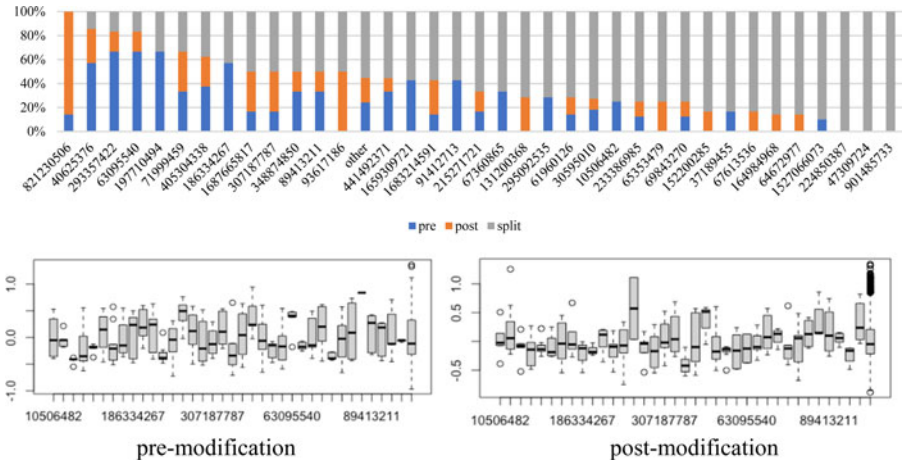


Figure 7. *Continued*

rather the avoidance of such a construction. If it is indeed true that young and female-presenting individuals are (re-)introducing syntactic conservatism as a new innovation in PhE, then the findings would parallel other variationist studies on Philippine languages like Lánnang-uè, which have found younger women also leading in established innovations like a stratified mixed language phonology (Gonzales & Starr 2020).

We also observe an effect of year on the observed internal variation – direct evidence of language change ($pd_{pre} = 1$, $pd_{post} = 0.99$). It seems that the preference for the split construction has increased over the past 11 years (table 2 and figure 7f). The findings are a welcome contribution to the relatively small amount of PhE literature discussing the diachronic dimension of PhE (Collins *et al.* 2014; Borlongan & Dita 2015; Gonzales 2023b), but they seem to be at odds with the findings on apparent time in the previous paragraph, which show a *decreasing* preference for the split infinitive construction. However, I argue that the findings are not contradictory, but complementary. If the effect of year represents change between 2010 and 2021, and the effect of age presentation represents a change-in-progress in the year 2021 (or the present), then the results could suggest the recent gradual reversal of the split infinitive trend that has dominated in the past eleven years, led in part by young women.

Lastly, the results show that individual user factors have a high chance of conditioning variation in the modified infinitive ($pd_{pre} = 1$, $pd_{post} = 1$). Figure 7f shows the robust effect of these factors. One user, 821230506, did not use any split infinitives at all. Other users like 901485733 only have modified infinitives realized as the split infinitive. The bulk of users (e.g. 293356422), however, vary in their placement of the adverb in the modified infinitive construction. The results point to the importance of factoring in individual factors in explaining syntactic variability in Twitter-style PhE.

5 Conclusion

The current article has investigated syntactic variation of modified infinitives in English, focusing on the English variety used on Twitter in the Philippines (i.e. Twitter-style PhE). It centers on the syntactic alternation between the split infinitive construction (i.e. *to* + ADV + INFINITIVE VERB), the post-modified infinitive construction (i.e. *to* + INFINITIVE VERB (+ …) + ADV), and the pre-modified infinitive construction (i.e. ADV + *to* + INFINITIVE VERB). The inquiry in this article seeks to determine whether the language-internal factors previously observed in studies on split infinitives in AmE (Calle-Martín & Miranda-García 2009; Perales-Escudero 2011) and the language-external factors influencing variation in Twitter-style PhE (Gonzales 2023b) play a role in conditioning the variation in the choice of modified infinitive constructions in one of AmE's offspring varieties – PhE (Thompson 2003).

The results of a multiple multinomial regression analysis of the data from a Bayesian standpoint reveal that the factors hypothesized to influence variation in splitting do indeed shape how the modified infinitive variable is realized. The adverb type, the reported frequency of adverb as a splitter in ICE, the length of the adverb relative to the verb, the likelihood of ambiguity, verb and adverb lexeme, and prosody (e.g. preserving the 'natural' rhythm) all jointly condition the syntactic variation in the modified infinitive in varying degrees. Furthermore, I discovered that this variation is also sensitive to language-external factors such as geography, age, sex, time and individual factors, part of which aligns with previous variationist research on PhE at the morphological and lexical levels of language (Gonzales 2023b). But while there is strong evidence for the conditioning effects of these factors on infinitive variation, the directions of the effects do not necessarily follow all the patterns described in the (AmE) literature. I attributed some of these divergences – particularly language-internal ones – to stylistic differences between AmE and Twitter-style PhE. Or if we consider Twitter-style PhE to be partially representative of PhE, it can be said that, as a nativized[9] variety of English, PhE can choose to pattern after its parent variety AmE; however, it can also chart its own trajectory and form its own set of conventions independent of AmE norms (Schneider 2003).

All in all, my findings stress the importance of the inclusion of language-internal and language-external variables in the study of split infinitives in English, particularly in PhE on Twitter. My study connects with other recent work urging more consideration of the combined role of these factors in conditioning variable processes (MacKenzie 2020; Kostadinova 2020). Although this study is already comprehensive, it would be beneficial to investigate other possible predictors of modified infinitive variation. The factors that have been investigated in this study only constitute a small subset of factors

[9] There is a current debate on which stage of Schneider's (2003) model of development of postcolonial Englishes PhE belongs to. Some argue that PhE is at Phase 3 (nativization) (Martin 2014) whereas others situate PhE at Phase 4 (endormative stabilization) (Borlongan 2016) or Phase 5 (differentiation) (Gonzales 2017). In this article, I use the term 'nativized', not in the Schneiderian sense, but more generally to refer to a variety that has gained local characteristics over time.

that condition the syntactic variation, so future work can attempt to incorporate other predictors of splitting such as discourse focus and stylistic context into their models of modified infinitive variation, as these have been shown in the literature to robustly condition splitting (Kostadinova 2020). Furthermore, follow-up work on variation in the modified infinitive construction utilizing multiple sources of PhE data across genres would be an important step in enhancing our understanding of how sociolinguistic factors structure variation in the use of split infinitives in PhE. In conjunction with the current study, such work would shed light on the complex nature of sociolinguistic variation in PhE as well as other Englishes and linguistic varieties around the world.

Like all research endeavors, this article is not without its limitations. One notable concern is the assurance of whether Twitter-style PhE truly represents an exemplar of standard or general PhE. Currently, there is a lack of extensive research on PhE that examines the degree to which Twitter-style PhE shares linguistic characteristics with other styles of PhE, such as the more conventional written and spoken forms. In other words, there is much we still do not know about Twitter-style PhE for us to comment on or compare with general PhE. Nevertheless, it is reasonable to anticipate that Twitter-style PhE would exhibit traits from both written and spoken PhE, such as a higher frequency of abbreviations, along with certain features specific to Twitter's platform, like shorter sentences and hashtags. Consequently, at this juncture, it remains challenging to determine the extent to which Twitter-style PhE represents a definitive example of general or standard PhE, assuming such a standard even exists. Aside from the difficulty of comparisons to a PhE 'standard', using the corpus to study Twitter-style PhE or PhE as a whole faces other challenges. A major issue arises from the fact that not all tweets collected originate from PhE users, and distinguishing PhE users from non-PhE users is not a straightforward task. Factors like ethnicity (Filipino versus others), linguistic exposure (native speakers of Philippine-type languages), identity (self-identification as Filipino), education, residence (whether born and raised in the Philippines), accent (Filipino-accented English) and others play a role in defining PhE users. However, this information is not publicly available on Twitter, making it difficult to profile users solely based on their names and profile descriptions. Furthermore, providing an operational definition of what qualifies as a PhE speaker becomes complex due to the intricate sociolinguistics of PhE. For instance, does a White Caucasian tweeter who has lived in the Philippines since birth, speaks Tagalog, but lacks a Filipino-accent when speaking English count as a PhE speaker? Similarly, what about an ethnic Filipino tweeter who attended an international school in the Philippines and speaks English with a blend of Filipino and British accents? Determining whether these individuals are PhE speakers introduces complexities. The article acknowledges the challenges and intricacies involved in identifying PhE users within Twitter corpora (or any other corpora) and attempts to simplify the analysis by assuming that the corpus is predominantly representative of PhE users, broadly defined as individuals who identify as Filipino. Despite the inevitable limitations and baggage associated with this approach, it is hoped that the substantial volume of data would

compensate for the inclusion of 'non-PhE' user data, thereby mitigating potential noise in the analysis, as is often encountered with social media corpora.

Despite some limitations with respect to generalizability and the like, the present study has theoretical and methodological significance for the fields of variationist sociolinguistics and world Englishes. It provides empirical evidence that can help one assess the robustness of dominant theories in the fields (e.g. Labov's principles of change, developmental theories of post-colonial Englishes) (Labov 1994; Schneider 2003). The findings have relevance not only for studies of split infinitives in English, but also for longstanding questions of the representation and nature of variable phenomena (MacKenzie 2020). Furthermore, the study has adopted the less commonly used Bayesian approach to analyzing data, which has been shown to facilitate intuitive, statistically sound and nuanced interpretations of the statistical results (Vasishth *et al.* 2018; McElreath 2020). Tools that utilize Deep Learning (Wang *et al.* 2019) were also used to help solve the problem of missing data in sociolinguistics (Gonzales 2004). It is hoped that the current study will contribute to theory and method in these aspects.

Author's address:

Department of English
3/F Fung King Hey Building
The Chinese University of Hong Kong
Shatin, New Territories
Hong Kong SAR, People's Republic of China
wdwonggonzales@cuhk.edu.hk

References

Bautista, Ma. Lourdes S. 2000. *Defining Standard Philippine English: Its status and grammatical features*. Manila: De La Salle University Press.

Bautista, Ma. Lourdes S. 2004. An overview of the Philippine component of the International Corpus of English (ICE-PHI). *Asian Englishes* 7(2), 8–26.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller & Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30), 774. https://doi.org/10.21105/joss.00774

Biber, Douglas & Randolph Quirk. 2012. *Longman grammar of spoken and written English*, 10th impression. Harlow: Longman.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python*. Sebastopol, CA: O'Reilly.

Bohmann, Axel. 2016. Language change because Twitter? Factors motivating innovative uses of *because* across the English-speaking Twittersphere. In Lauren Squires (ed.), *English in computer-mediated communication*, 149–78. Berlin: De Gruyter.

Borlongan, Ariane Macalinga. 2016. Relocating Philippine English in Schneider's dynamic model. *Asian Englishes* 18(3), 232–41.

Borlongan, Ariane Macalinga & Shirley N. Dita. 2015. Taking a look at expanded predicates in Philippine English across time. *Asian Englishes* 17(3), 240–7.

Bürkner, Paul-Christian. 2017. **brms**: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). https://doi.org/10.18637/jss.v080.i01

Calle-Martín, Javier & Antonio Miranda-García. 2009. On the use of split infinitives in English. In Antoinette Renouf & Andrew Kehoe (eds.), *Corpus linguistics: Refinements and reassessments*, 347–64. Amsterdam: Rodopi.
Calle-Martín, Javier & Jesús Romero-Barranco. 2014. On the use of the split infinitive in the Asian varieties of English. *Nordic Journal of English Studies* 13(1), 130. https://doi.org/10.35360/njes.296
Collins, Peter, Ariane Macalinga Borlongan, Joo-Hyuk Lim & Xinyue Yao. 2014. The subjunctive mood in Philippine English: A diachronic analysis. In Simone E. Pfenninger, Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja, Marianne Hundt & Daniel Schreier (eds.), *Contact, Variation, and Change in the History of English* (Studies in Language Companion Series 159), 259–80. Amsterdam: John Benjamins.
Crystal, David. 1984. *Who cares about English usage?* London: Longman.
Davies, Mark & Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide. A Journal of Varieties of English* 36(1), 1–28. https://doi.org/10.1075/eww.36.1.01dav
Eckert, Penelope. 1989. The whole woman: Sex and gender differences in variation. *Language Variation and Change* 1, 245–67.
Eckert, Penelope & John R. Rickford. 2001. *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.
Franke, Michael & Timo Benjamin Roettger. 2019. *Bayesian regression modeling (for factorial designs): A tutorial*. Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/cdxv3
Gonzales, Wilkinson Daniel Wong. 2017. Philippine Englishes. *Asian Englishes* 19(1), 79–95.
Gonzales, Wilkinson Daniel Wong. 2022a. Hybridization. In Ariane Macalinga Borlongan (ed.), *Philippine English: Development, Structure, and Sociology of English in the Philippines*, 170–83. Abingdon: Routledge.
Gonzales, Wilkinson Daniel Wong. 2022b. 'Truly a language of our own': A corpus-based, experimental, and variationist account of Lánnang-uè in Manila. PhD dissertation. University of Michigan.
Gonzales, Wilkinson Daniel Wong. 2023a. Spread, stability, and sociolinguistic variation in multilingual practices: The case of Lánnang-uè. *International Journal of Multilingualism*. https://doi.org/10.1080/14790718.2023.2199998
Gonzales, Wilkinson Daniel Wong. 2023b. Broadening horizons in the diachronic and sociolinguistic study of Philippine English with the Twitter Corpus of Philippine Englishes (TCOPE). *English World-Wide. A Journal of Varieties of English*. https://doi.org/10.1075/eww.22047.gon
Gonzales, Wilkinson Daniel Wong. 2024. Sociolinguistic analysis with missing metadata? Leveraging linguistic and semiotic resources through deep learning to investigate English variation and change on Twitter. *Applied Linguistics.* https://doi.org/10.1093/applin/amad086
Gonzales, Wilkinson Daniel Wong & Shirley N. Dita. 2018. Split infinitives across World Englishes: A corpus-based investigation. *Asian Englishes* 20(3), 242–67. https://doi.org/10.1080/13488678.2017.1349858
Gonzales, Wilkinson Daniel Wong, Mie Hiramoto, Jakob R. E. Leimgruber & Jun Jie Lim. 2022. *Is it* in Colloquial Singapore English: What variation can tell us about its conventions and development. *English Today.* First View, 1–14. https://doi.org/10.1017/S0266078422000141
Gonzales, Wilkinson Daniel Wong & Rebecca Lurie Starr. 2020. Vowel system or vowel systems? Variation in the monophthongs of Philippine Hybrid Hokkien in Manila. *Journal of Pidgin and Creole Languages* 35(2), 253–92.
Grafmiller, Jason, Benedikt Szmrecsanyi & Lars Hinrichs. 2018. Restricting the restrictive relativizer. *Corpus Linguistics and Linguistic Theory* 14(2), 309–55.
Heuser, Ryan, Josh Falk & Arto Anttila. 2010. *Prosodic: A metrical–phonological parser, written in Python*. https://github.com/quadrismegistus/prosodic

Hiramoto, Mie, Wilkinson Daniel Wong Gonzales, Jakob Leimgruber, Jun Jie Lim & Jessica Xue Ming Choo. 2022. From Malay to Colloquial Singapore English: A case study of sentence-final particle *sia*. In Aloysius Ngefac, Hans-Georg Wolf & Thomas Hoffman (eds.), *World Englishes and creole languages today existing paradigms and current trends in action* (LINCOM Studies in English Linguistics 24), 117–30. Lincom Europa.
Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303
Ilbury, Christian. 2020. 'Sassy Queens': Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics* 24(2), 245–64.
Jenkins, Jennifer. 2015. *Global Englishes: A resource book for students* (Routledge English Language Introductions), 3rd edn. Abingdon and New York: Routledge.
Kato, Kazuo. 2001. Not to be or to not be: More on split negative infinitives. *American Speech* 76(3), 312–15.
Koivistoinen, Erika. 2012. Acceptable or not? Split infinitives in American English. Bachelor's thesis, University of Jyväskylä.
Kostadinova, Viktorija. 2020. Examining the split infinitive: Prescriptivism as a constraint in language variation and change. In Don Chapman & Jacob D. Rawlins (eds.), *Language prescription*, 26. Bristol: Multilingual Matters.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.
Labov, William. 1994. *Principles of linguistic change*. Oxford: Blackwell.
Lee, Aldrin P. & Ariane Macalinga Borlongan. 2022. Internal variation. In Ariane Macalinga Borlongan (ed.), *Philippine English: Development, structure, and sociology of English in the Philippines*, 125–34. Abingdon: Routledge.
Leimgruber, Jakob, Jun Jie Lim, Wilkinson Daniel Wong Gonzales & Mie Hiramoto. 2021. Ethnic and gender variation in the use of Colloquial Singapore English discourse particles. *English Language and Linguistics* 25(3), 601–20.
Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2), 235–68.
Levshina, Natalia. 2022. Comparing Bayesian and frequentist models of language variation: The case of *help* + (*to*-)infinitive. In Ole Schützler & Julia Schlüter (eds.), *Data and methods in corpus linguistics*, 224–58. Cambridge: Cambridge University Press.
Llamzon, Teodoro. 1997. The phonology of Philippine English. In Ma. Lourdes S. Bautista (ed.), *English is an Asian language*, 41–8. Sydney: The Macquarie Library.
MacKenzie, Laurel. 2020. Comparing constraints on contraction using Bayesian regression modeling. *Frontiers in Artificial Intelligence* 3, 58. https://doi.org/10.3389/frai.2020.00058
Makowski, Dominique, Mattan S. Ben-Shachar, S. H. Annabel Chen & Daniel Lüdecke. 2019. Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology* 10, 2767. https://doi.org/10.3389/fpsyg.2019.02767
Martin, Isabel Pefianco. 2014. Beyond nativization? Philippine English in Schneider's dynamic model. In Sarah Buschfeld, Thomas Hoffman, Magnus Huber & Alexander Kautzsch (eds.), *The evolution of Englishes: The dynamic model and beyond*, 70–85. Amsterdam: John Benjamins.
McElreath, Richard. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan* (CRC Texts in Statistical Science), 2nd edn. Boca Raton, FL: CRC Press.
Meyerhoff, Miriam. 2018. *Introducing sociolinguistics*, 3rd edn. Abingdon and New York: Routledge.
Mikulová, Hana. 2011. Split infinitive – corpus analysis. Bachelor's thesis, Univerzita Palackého v Olomouci.
Mitrasca, Marcel. 2009. The split infinitive in electronic corpora: Should there be a rule? *Concordia Working Papers in Applied Linguistics* 2, 99–131.

Perales-Escudero, Moisés D. 2011. To split or to not split: The split infinitive past and present. *Journal of English Linguistics* 39(4), 313–34.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow: Longman.
R Core Team. 2015. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. www.R-project.org
Sankoff, Gillian. 2006. Age: Apparent time and real time. In Keith Brown (ed.), *The encyclopedia of language & linguistics*, 2nd edn, 110–16. Oxford: Elsevier.
Schneider, Edgar. 2003. The dynamics of new Englishes: From identity construction to dialect birth. *Language* 79(2), 233–81.
Shahruddin, Irwan Shah, Ran Ao & Ee Ling Low. 2022. Phonology. In Ariane Macalinga Borlongan (ed.), *Philippine English: Development, structure, and sociology of English in the Philippines*. Abingdon: Routledge.
Sonderegger, Morgan. 2022. *Regression modeling for linguistic data*. Open Science Framework. https://osf.io/pnumg/ (31 January 2022).
Starr, Rebecca Lurie & Brinda Balasubramaniam. 2019. Variation and change in English /r/ among Tamil Indian Singaporeans. *World Englishes* 38(4), 630–43.
Tagliamonte, Sali A. & Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83(1), 3–34.
Thomason, Sarah. 2001. *Language contact: An introduction*. Washington, DC: Georgetown University Press.
Thomason, Sarah. 2007. Language contact and deliberate change. *Journal of Language Contact* 1(1), 41–62.
Thompson, Roger Mark. 2003. *Filipino English and Taglish: Language switching from multiple perspectives* (Varieties of English around the World General Series G 31). Amsterdam: John Benjamins.
Vasishth, Shravan & Bruno Nicenboim. 2016. Statistical methods for linguistic research: Foundational ideas – Part I: Statistical methods for linguistics. *Language and Linguistics Compass* 10(8), 349–69.
Vasishth, Shravan, Bruno Nicenboim, Mary E. Beckman, Fangfang Li & Eun Jong Kong. 2018. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 71, 147–61.
Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter & Paul-Christian Bürkner. 2021. Rank-normalization, folding, and localization: An improved R^ for assessing convergence of MCMC (with discussion). *Bayesian Analysis* 16(2). https://doi.org/10.1214/20-BA1221
Villanueva, Rey John Castro. 2016. The features of Philippine English across regions. PhD dissertation, De La Salle University.
Wang, Zijian, Scott A. Hale, David Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck & David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. *The World Wide Web Conference* 2056–7. https://doi.org/10.1145/3308558.3313684