David Paul Gerards*, Patrick Steinmetz and
Levin Ludwig Schwarzkopf

Discovering and mapping predicted but undocumented morphosyntactic variation through Twitter: evidence from Spanish bare count singulars with *ir a*

https://doi.org/10.1515/flin-2025-0045
Received March 11, 2025; accepted October 8, 2025; published online January 30, 2026

Abstract: This paper investigates – to the best of our knowledge, for the first time – whether social media platforms can be used to detect and geographically map theoretically predicted but previously undocumented morphosyntactic variants. Specifically, it examines Spanish bare singular count nouns with *ir a* (lit. 'to go to + [noun]') across different national varieties of Spanish, as opposed to Standard Spanish *ir al* + [noun] (lit. 'to go to the + [noun]'). Based on Twitter data (n = 6,206), we show that such bare singular count nouns exist and are particularly prevalent in Colombian Spanish, a fact not previously reported in the literature. However, a follow-up acceptability judgment task with native Spanish speakers (n = 226) suggests that social media data may produce false positives, especially for low-frequency phenomena. As a result, we argue that such studies should always be complemented by other methods, e.g., from experimental linguistics.

Keywords: bare count singulars; digital geolinguistics; experimental linguistics; Twitter/X; Spanish; weak definites

1 Introduction: Twitter/X and geolinguistics

The use of data from social media platforms has significantly increased in geolinguistics over the past fifteen years. The most widely used platform by far is Twitter (for a description, see Squires 2015), renamed X in 2023 following its acquisition by Elon Musk. Twitter/X is convenient, contains vast amounts of data, and enables

***Corresponding author: David Paul Gerards**, Department of Romance Languages, Johannes Gutenberg University Mainz, Jakob-Welder-Weg 18, 55128 Mainz, Germany, E-mail: david.gerards@uni-mainz.de. https://orcid.org/0000-0003-2380-2631
Patrick Steinmetz and Levin Ludwig Schwarzkopf, Department of Romance Languages, University of Leipzig, 04107 Leipzig, Germany

linguists to "unobtrusively observe" and study features characteristic also of informal registers and from languages with relatively few speakers (Nguyen 2021: 207). However, Twitter/X's uncontested potential to study also rare phenomena (cf. Nguyen 2021: 205; Strelluf 2022: 49; a.o.) is only one side of the coin. Problematically, for instance, the community of Twitter/X users, instead of being a representative population sample, is skewed towards young urban social groups (Duggan and Brenner 2013; Eisenstein 2017: 370; De Benito Moreno 2022: 488–489), a trend that may however be moderating as the platform is becoming less popular among younger teens, at least in the U.S. (Anderson et al. 2023). Additionally, Twitter/X usage frequency and verbal behavior on the platform varies across languages and even countries sharing the same language (Hong et al. 2011; De Benito Moreno 2022: 489).

A more fundamental issue than those noted above is that data from social media platforms like Twitter/X need to be repurposed for their use in linguistics as these platforms were not initially designed for linguistic research (Nguyen 2021: 206). Unlike many modern corpora specifically tailored for linguistic purposes, Twitter/X is, for instance, not lemmatized or part-of-speech tagged, nor does it allow searches for specific cotexts within a set range of tokens preceding or following a necessarily plain-text search string. This often leads to strenuous manual data cleaning and may help explain why most geolinguistic research using Twitter/X data tends to focus either on easily queryable lexical features (for English, see, e.g., Russ 2013; Eisenstein et al. 2014; Huang et al. 2016; Kulkarni et al. 2016; Eisenstein 2017; Grieve et al. 2018, 2019; Blaxter and Britain 2021; for Dutch, van Halteren et al. 2018; for Catalan, Perea and Ruiz Tinoco 2016) or on morphosyntactic features that can be identified through plain-text search strings (for English, see, e.g., Haddican and Johnson 2012; Doyle 2014; Strelluf 2022; for Welsh, Willis 2020; for Frisian, Dijkstra et al. 2021; for Catalan, Estrada Arráez and De Benito Moreno 2016; for French, Abitbol et al. 2018; for Portuguese – and for a morphosyntactic phenomenon not easily queryable – Gerards 2022). Research on phonetic and phonological geolinguistic variation using Twitter/X data is still in its early stages and necessarily relies on an indirect methodology that extrapolates phonic information from spelling (for English, see Jones 2015; Jørgensen et al. 2015; for Dutch, see van Halteren et al. 2018).

Spanish, the language examined in this paper, is among those for which a fairly large amount of Twitter/X-based geolinguistic research has been done. Interestingly though, and in stark contrast to English, studies on Spanish display a greater focus on morphosyntactic variables than on lexical ones (for lexical features, see Gonçalves and Sánchez 2014, 2016; Donoso and Sánchez 2017; for morphosyntactic ones, Tinoco 2013; Brown 2016, Estrada Arráez and De Benito Moreno 2016; Pato and De Benito Moreno 2017; Claes 2017; De Benito Moreno and Estrada Arráez 2018; Marttinen Larsson and Bouzouita 2018; Hoff 2020; Casanova 2020; Kellert 2024). The only

Twitter/X data-based study (marginally) addressing the geolinguistics of Spanish phonetic features we are aware of is De Benito Moreno and Estrada Arráez (2018).

An additional challenge in using Twitter/X data for linguistic research in general, and for geolinguistics in particular, is determining the geographical origin of the data. This is due to the fact that it is optional for Twitter/X users to explicitly disclose such information. Broadly speaking, there are three different approaches to address this issue. Firstly, some studies have relied exclusively on geolocated posts, which include the GPS coordinates of the user at the time the post was made (e.g., Eisenstein et al. 2014; Jørgensen et al. 2015; De Benito Moreno and Estrada 2018; Grieve et al. 2019; Hoff 2020; Willis 2020; Blaxter and Britain 2021; Strelluf 2022). Secondly, other research has inferred the geographical origin from the content of Twitter/X's localization field, where users can optionally input free text with geographical information (e.g., Russ 2013; Nguyen et al. 2015; Bland and Morgan 2020; Willis 2020). A combination of these two methods has also been explored (e.g., Doyle 2014). Thirdly and lastly, there is a growing body of studies that use diverse computational techniques to predict the geographical origin of a user based on the form, content, and/or other meta-analyses of the user's posts, as well as those of their social networks on Twitter/X (e.g., Eisenstein et al. 2010; Scheffler et al. 2014; Rahimi et al. 2017; Abitbol et al. 2018; Zola et al. 2020; Mahajan and Mansotra 2021; Lamsal et al. 2022; Julie et al. 2023; for surveys see Melo and Martins 2016; Mahajan and Mansotra 2021; Hu et al. 2023).

Comprehensively evaluating all advantages and disadvantages of these different techniques, as well as their accuracy, is beyond the scope of this paper (cf., e.g., Graham et al. 2014; Zheng et al. 2018, also for further reading). Nonetheless, a few brief observations are warranted. This is especially true as there is, to date, no established gold standard for geographically mapping Twitter/X data (Graham et al. 2014: 569). The continued relevance of this observation is demonstrated by the fact that even among the most recent studies cited above, all three types of approaches are represented.

In our opinion, the third group of approaches is certainly the most promising for future use, primarily because it is likely to allow for the retention of more data, as relatively few data points need to be discarded due to missing geographical information. However, aside from requiring significant NLP expertise, a major drawback of such approaches is that they have been developed almost exclusively for English (Graham et al. 2014: 571; but see Ambrosio Aguilar et al. 2021 for a recent attempt to adapt a content-based approach to Spanish). The first and second group of approaches, too, have their pros and cons. Relying on geolocated posts featuring GPS coordinates is, above all, convenient, as very little additional data cleaning is necessary. However, only around 1–2% of Twitter users choose to enable geolocation, meaning that the majority of posts cannot be used for research that requires

geographically localized data (Pato and de Benito Moreno 2017: 127; Eisenstein 2017: 369). This limitation may not pose a significant problem when investigating relatively common phenomena in English but can become an issue when working on languages with fewer speakers and/or with low-frequency phenomena. Additionally, rural Twitter users enable GPS geolocation less frequently than urban ones, and among urban users, females enable GPS geolocation more often than males (Hecht and Stephens 2014; Pavalanathan and Eisenstein 2015). These biases can further skew the data beyond the inherent demographic biases already associated with Twitter. Finally, GPS-geolocation indicates where a post was sent from and not necessarily where the *user* is from. While many studies tacitly accept this shortcoming as an inevitable source of noise (for an exception, see Brown 2016: 50), both Graham et al. (2014) and De Benito Moreno and Estrada Arráez (2018) have demonstrated that physical mobility in an increasingly globalized world results in a considerable number of posts being sent from locations that do not align with the user's geographical origin. This may not be a major issue when the geographic distribution of variants is at least partially known (e.g., Grieve et al. 2019), but it can become problematic in cases where this information is unknown.

The percentage of Twitter users providing valid information in the free-text localization field is significantly higher than those enabling GPS geolocation. For instance, Hecht et al. (2011) reported that 66 % of users in their dataset provided valid geographic information, while 18 % left the field blank, and 16 % entered fictional locations such as *around the corner* or *in the hood*. Similar, more recent percentages are reported by Willis (2020). However, even valid geographic information may not always be sufficiently informative, depending on the research questions involved. For instance, when examining geographic linguistic variation within a country or region, locations entered in the field that only specify countries are too coarse-grained to be useful. In combination with fantasy locations, the latter means that using the localization field to determine the geographic origin of Twitter data requires an enormous amount of time-consuming manual data inspection and cleaning (e.g., Doyle 2014: 101; Willis 2020: 4–6), also because automatic geocoders do not yet perform particularly well with input from Twitter's location field (Graham et al. 2014: 570, 576; Bland and Morgan 2020).

Despite all potential pitfalls associated with Twitter/X data outlined so far, it has been repeatedly demonstrated that data from this platform, when handled with care, are a valuable source in geolinguistics. The results of Twitter/X-based studies replicate fairly well and can even complement the findings of traditional geolinguistic studies based on fieldwork (Russ 2013; Brown 2016; van Halteren et al. 2018; Grieve et al. 2019; Willis 2020; Nguyen 2021: 211, among others).

With this conviction, and in response to Nguyen (2021: 213), who explicitly lists the "bottom-up discovery of features" as a promising future direction for Twitter

data-based linguistic research, the present study tests whether Twitter data can also be used to detect and geographically map *previously undocumented but theoretically predicted linguistic variants*. To the best of our knowledge, no such endeavor has been undertaken so far. The study most similar to ours – though not involving theoretical prediction – is Grieve et al. (2018), who identify and map lexical innovations in American English.[1] Additionally, we verify part of our Twitter-based findings through experimental judgment data. Although shown to be highly effective by Haddican and Johnson (2012) and Brown (2016), these are the only studies we are aware of that employ such a mixed-method approach in geolinguistics.

The remainder of the paper is structured as follows: Section 2 provides the theoretical and empirical background on bare count singulars, the morphosyntactic phenomenon used to assess whether social media data can serve the purposes outlined above. Section 3 presents a Twitter-based study on previously undocumented but theoretically predicted, lexically restricted Spanish bare count singulars with *ir a* + [noun] (n = 6,206). Section 4 verifies parts of the findings from Section 3 through an Acceptability Judgment Task completed by 226 native Spanish speakers. Section 5 discusses the empirical, methodological, and theoretical implications of the results. Section 6 concludes the paper.

2 Bare count singulars

This section introduces bare count singulars, the variable exemplarily chosen to investigate the research question of this paper. Section 2.1 outlines key cross-linguistic features of these nominals, emphasizing their most significant properties. Section 2.2 focuses specifically on Spanish, providing a state-of-the-art review of bare count singulars in this language.

2.1 Bare count singulars: a cross-linguistic perspective

Bare count singulars (henceforth, BCSs) are widely attested in languages with well-developed article systems, featuring both definite and indefinite articles. The examples in (1)–(6) contain instances of BCSs in English, Dutch, Italian, Brazilian Portuguese, Norwegian, and Albanian. These examples are noteworthy because regular count noun arguments in these languages generally require the use of a determiner.

1 In this context, note too that the present study differs from Russ (2013), Doyle (2014), Jørgensen et al. (2015), Eisenstein (2017), Strelluff (2022), and others who investigate variants with unknown distributions, but whose existence had already been noted.

BCSs are to sharply be distinguished from bare mass nouns, as exemplified by English (7) and German (8). As the bareness of argumental (indefinite) mass nouns, unlike that of BCSs, is the default in most article languages (but see Ihsane et al. 2025), bare mass nouns are not relevant to the present paper.

(1) English (BCS)
a. *Sue took her nephew to* {*college*/*prison*/*class*}.
b. *Mark attended* {*college*/*class*/*school*}.
c. *The ship is at* {*sea*/*port*}.
(Carlson et al. 2006: 180)

(2) Dutch (BCS)
Bert ging naar school.
Bert went to school
'Bert went to school.'
(Aguilar-Guevara and Oggiani 2023: 3)

(3) Italian (BCS)
Vado a {*teatro*/*scuola*}.
I.go to theater/school
'I am going to the theater/to school.'
(Leonetti 2019: 14; adapted)

(4) Brazilian Portuguese (BCS)
Pedro vai ler jornal.
Pedro is.going.to read newspaper
'Pedro is going to read the newspaper.'
(Espinal and Cyrino 2017a: 136)

(5) Norwegian (BCS)
Hun vasket sykkel.
she washed bike
'She washed a bike.'
(Borthen 2003: 62)

(6) Albanian (BCS)
Ana do të blejë biçikletë.
Ana wants to buy bicycle
'Ana wants to buy a bicycle.'
(Kallulli 1999: 79)

(7) English (bare mass noun)
a. *Sue drank milk.*
b. *Peter bought rice.*

(8) German (bare mass noun)
Hans aß Brot.
Hans ate bread
'Hans ate bread.'

As illustrated in (1)–(6), BCSs are typically direct objects or arguments of prepositions (e.g., Carlson et al. 2006), although some subject BCSs have also been discussed in the literature (Munn and Schmitt 2002; Stvan 2009; Wall 2017). While a detailed examination of the syntax and semantics of BCSs, including their distributional asymmetries, lies beyond the scope of this paper, it is important to highlight select cross-linguistic features of BCSs. This serves as a foundation for the discussion of Spanish BCSs in Section 2.2 and justifies their selection for investigating the research question of this paper.

Cross-linguistically, BCSs display a remarkable set of shared semantic, (morpho) syntactic, and lexical properties that distinguish them from standard argumental expressions in the respective languages. One such property, their difficulty in appearing as subjects, was already noted above. In addition, at least nine other properties have been identified (Aguilar-Guevara and Oggiani 2023; Aguilar-Guevara and Zwarts 2013; Carlson and Sussman 2005; Carlson et al. 2006). The following discussion briefly highlights three of these cross-linguistically stable BCS features: (i) meaning enrichment, (ii) lexical restrictions, and (iii) modificational restrictions.[2]

Feature (i): BCSs are characterized by semantic meaning enrichment, which is dependent on the stereotypical or habitual activities associated with a given noun and absent from regular argumental expressions in the respective languages (cf., e.g., Carlson and Sussman 2005: 74; Carlson et al. 2006: 182; Aguilar-Guevara and Zwarts 2013: 36). For example, the English BCS *in prison* in (9a) necessarily implies that Sue herself is a prison inmate. In contrast, (9a) is infelicitous if Sue is going to a prison to visit someone else. In the latter case, the only acceptable options are (9b) or (9c), depending on the discourse referential status of *prison*.

(9) a. *Sue is **in prison**.*
b. *Sue is **in a prison**.*
c. *Sue is **in the prison**.*

2 The remaining six cross-linguistic properties of BCS are as follows: (iv) sloppy identity in VP-ellipsis sentences, (v) acceptability in sluicing contexts, (vi) obligatory narrow scope, (vii) restrictions on nominal number morphology, (viii) discourse referential defectiveness, and (ix) non-unique reference (for the latter, see also below; Carlson and Sussman 2005; Carlson et al. 2006; Aguilar-Guevara and Zwarts 2013; Aguilar-Guevara and Oggiani 2023).

Feature (ii): In all languages, BCSs exhibit strong lexical restrictions, whether nominal, verbal, or prepositional. This is demonstrated in English (10a), where the BCS *in prison* is perfectly acceptable, in stark contrast to the ungrammatical BCSs *next to prison* and *in store* (10b/c) (cf. Carlson and Sussman 2005: 74; Carlson et al. 2006: 180–181; Aguilar-Guevara and Zwarts 2013: 36; Aguilar-Guevara and Oggiani 2023: 6–7).

(10) a. *Joe is **in prison**.*
b. **Joe is **next to prison**.*
c. **Joe is **in store**.*

Feature (iii): Another notable cross-linguistically stable property of BCSs is that such nominals are not freely modifiable. For example, while *in prison* is a felicitous English BCS (11a), adjectival modification with *horrible* renders the BCS ungrammatical (11b), instead requiring the use of a determiner (11c) (cf. Carlson and Sussman 2005: 74; Carlson et al. 2006: 180–181; Aguilar-Guevara and Zwarts 2013: 36; Aguilar-Guevara and Oggiani 2023: 8).

(11) a. *Joe is **in prison**.*
b. **Joe is **in horrible prison**.*
c. *Joe is **in** {**a**/**some**/**this**} **horrible prison**.*

Broadly speaking, the cross-linguistic properties of BCSs illustrated in (9)–(11), along with the additional ones listed in note 2, are best explained by modeling BCSs as semantically pseudo-incorporated nominals. In such analyses, BCSs nominals – pretheoretically speaking – form a tighter-than-usual union with a verb or preposition. For detailed discussions on how semantic pseudo-incorporation is modeled within formal frameworks and how it accounts for the cross-linguistic peculiarities of BCSs, readers are referred to an extensive body of specialized literature, including but not limited to van Geenhoven (1998), Chung and Ladusaw (2004), Farkas and de Swart (2003), Dobrovie-Sorin et al. (2006), Massam (2009), Stvan (2009), Mithun (2010), Dayal (2011), Espinal and McNally (2011), Carlson (2006), and Borik and Gehrke (2015).

A final cross-linguistic aspect relevant to the present paper is that, in article languages, BCSs are in complementary distribution with short weak definites (henceforth, SWDs; cf., e.g., Carlson et al. 2006; Leonetti 2019; Aguilar-Guevara and Oggiani 2023; for other types of weak definites, see Espinal and Cyrino 2017b; Gerards and Stark 2022: 5–10). SWDs are nominals that feature a definite article but lack the semantic properties of uniqueness and familiarity, the key semantic ingredients standardly assumed for definite articles (Christophersen 1939; Heim 1982; Russell 1905). For illustration, consider the English examples in (12):

(12) a. *Jacqueline took **the train** from Paris to Moscow.*
b. *Jacqueline watered **the plant**.*
([12a]: Carlson et al. 2006: 186)

The SWD *the train* in (12a) is felicitous even if Jacqueline changed trains several times during her journey. Furthermore, (12a) can be uttered out of the blue, meaning that the train exemplar(s) do not need to meet familiarity requirements (Zwarts 2014: 267). This renders SWDs truth-conditionally equivalent to indefinites and, therefore, fundamentally different from regular strong definites, such as *the plant* in (12b). The latter, unlike (12a), is felicitous iff the cardinality of plants watered by Jacqueline is |1| and iff this plant has been previously introduced into the discourse universe or is, at minimum, subject to bridging or cataphoric accommodation (Hawkins 1978).

Importantly, all the semantic, (morpho)syntactic, and lexical properties discussed above for BCSs also apply to SWDs, including those mentioned in note 2 (Aguilar-Guevara and Oggiani 2023; Aguilar-Guevara and Zwarts 2013; Carlson et al. 2006). The parallel behavior of BCSs and SWDs is illustrated through examples of meaning enrichment, lexical restrictions, and modificational restrictions in English (13), constructed around the English SWD *call the doctor* (Aguilar-Guevara and Zwarts 2010: 189).

(13) a. #*Alice called **the doctor**, but not for medical reasons.*
b. #*Alice called **the sailor**.*
c. #*Alice called **the nice doctor**.*
([13a]: Aguilar-Guevara and Zwarts 2010: 189)

Although all examples in (13) are fully grammatical, they lose the possibility of being interpreted as SWDs (indicated by #). In (13a), the second clause, introduced by *but*, negates the stereotypical, activity-based meaning enrichment associated with an SWD reading of *call the doctor* (namely, seeking medical advice); the same applies to (13c), where the modifying adjective *nice* has the same effect: The only possible interpretation of (13a) and (13c) is that of a regular strong definite. This is evidenced by the fact that both examples are now subject to the same uniqueness and familiarity requirements as the strong definite *the plant* in (12b): Unlike what was observed for the SWD *the train* in (12a), neither (13a) nor (13c) is felicitous if Alice called a doctor who had not previously been introduced into the discourse or if she called more than one doctor. In (13b), substituting *the doctor* with another lexical item, *the sailor*, results in the same loss of a possible SWD reading.

The parallel properties of BCSs and SWDs have led researchers to propose that these two types of nominals represent morphosyntactic surface variants of a single underlying semantic phenomenon: pseudo-incorporation (e.g., Carlson et al. 2006; Aguilar-Guevara and Zwarts 2013; Aguilar-Guevara and Oggiani 2023). Different

article languages (14) and different varieties of the same article language (15) exhibit variation between BCSs and SWDs for the same lexical item. Similarly, within a single language, the choice between BCSs and SWDs can also vary depending on the lexical item in question (16) (Aguilar-Guevara and Oggiani 2023; Carlson et al. 2006, passim; Leonetti 2019: 13–15). No linguistic factors have yet been identified to parameterize this variation (cf. especially Leonetti 2019: 12–13).

(14) a. *Hablo por teléfono*. (BCS, Spanish)
I.talk on phone
b. *Parlo al telefono.* (SWD, Italian)
I.talk on.the phone
'I talk on the phone.'
(Leonetti 2019: 14)

(15) a. *I go to* {*hospital*/*university*}. (BCS, British English)
b. *I go to the* {*hospital*/*university*}. (SWD, American English)
(Carlson et al. 2006: 185; adapted)

(16) a. *Vado a teatro.* (BCS, Italian)
I.go to theater
'I go to the theater.'
b. *Vado al cinema.* (SWD, Italian)
I.go to.the cinema
'I am going to the cinema.'
(personal knowledge)

Given their shared properties and status as two morphosyntactic variants of a single underlying semantic class, inter- and intralinguistic BCS/SWD variation of types (14)–(16) is to be expected. Consequently, BCSs serve as an ideal testing ground for investigating whether social media data can be used to detect and geographically map theoretically predicted but previously undocumented morphosyntactic variants. The following section provides an overview of BCS (and SWD) usage in Spanish, the language chosen for the Twitter- and experimentally-based studies presented in Sections 3 and 4.

2.2 Bare count singulars (and short weak definites) in Spanish

As Laca (1999: 919) observes, Spanish is a language with relatively limited BCS use (cf. also RAE/ASALE 2009: 1156; Bosque 1996: 35). Instead, many of the nominals cross-linguistically prone to feature BCSs appear as SWDs. Consider the data in (17) and compare them to their English translations and to Dutch (2) and Italian (3):

(17) a. *Luis fue a=*(l) hospital.*
Luis went to=the hospital
'Luis went to (the) hospital.'
b. *María fue a *(la) escuela.*
María went to the school
'Mary went to school.'
c. *Juan siempre va a=*(l) teatro.*
Juan always goes to=the theater
'Juan always goes to the theater.'
d. *Ana siempre va a *(la) iglesia.*
Ana always goes to the church
'Ana always goes to church.'
e. *Julia está en *(la) cárcel*
Julia is in the prison
'Julia is in prison.'

Nevertheless, there are Spanish BCSs that occur regularly without any diatopic marking. Consider, for example, the direct object BCS in (18), described by RAE/ASALE (2009: 1156; our translation) as being used under "stereotyped conditions", aligning with the property of BCS meaning enrichment (Section 2.1), and sometimes further restricted to intensional verbs (18b/c) or negation ([18d]; cf. also Laca 1999: 919–920; Bosque 1996: 39–45).

(18) a. *Llevaba falda.*
s/he.wore skirt
'S/he wore a skirt.'
b. *Estoy {buscando/*pintando} piso.*
I.am looking.for/painting apartment
'I am {looking for/painting} an apartment.'
c. *Ha {pedido/*guardado} coche nuevo.*
s/he.has ordered/kept car new
'S/he {ordered/kept} a new car.'
d. *(*No*) *hay profesor que no se haya enterado.*
not there.is teacher who not REFL has found.out
'There is no teacher who hasn't found out.'
([18a]: RAE/ASALE 2009: 1157; [18b/c]: Bosque 1996: 35; [18d]: Laca 1999: 920)

Other domains with regular, diatopically unmarked Spanish BCSs include (i) "verbal locutions, i.e., complex predicates listed in dictionaries" (RAE/ASALE 2009: 1157; our translation), as exemplified in (19); (ii) certain fossilized proverbs ([20]; cf. also RAE/ASALE 2009: 1126, 1148); and (iii) adverbial PPs, particularly those denoting instruments or means of transportation ([21], Bosque 1996: 50–54; Laca 1999: 923; RAE/ASALE 2009: 1160). In contrast, directional BCSs such as (22), are extremely rare (Álvarez Martínez 1986: 220–222; Laca 1999: 921–923).

(19) a. *tomar nota*, *echar mano*, *formar parte*
take note lend hand form part
'to take notes, to lend a hand, to take part'
b. *estar en duda*, *salir de fiesta*
be in doubt go.out of party
'to be in doubt, to go out partying'
(Laca 1999: 920; adapted)

(20) a. *Perro que ladra no muerde.*
dog that barks not bites
'A barking dog doesn't bite.'
b. *Piedra que rueda no cría moho.*
stone that rolls not creates moss
'A rolling stone gathers no moss.'
([20a]: Bosque 1996: 49; [20b]: Laca 1999: 924)

(21) a. *Cerró la puerta con llave.*
s/he.closed the door with key
'S/he locked the door.'
b. *Hicieron algunos tramos en bus.*
they.made some stretches in bus
'They traveled some stretches by bus.'
(Laca 1999: 923)

(22) a. *Hoy no fueron a clase.*
today not they.went to class
'Today, they didn't go to class.'
b. *Regresó a casa de sus padres.*
s/he.returned to house of his/her parents
'S/he returned to her/his parents' house.'
c. *lban a misa los domingos.*
they.went to mass the Sundays
'They used to go to Mass on Sundays.'
(Laca 1999: 922)

Regarding the diatopic variation of Spanish BCSs, the existing literature is very limited. Kany (1969: 39) cursorily noted a "decreasing" tendency in Latin American Spanish toward locative and directional BCS use. He provides example (23a) from Bolivia and (23b) from Mexico, both of which correspond to SWDs in Standard Spanish.

(23) a. *No voy más a colegio.*
not I.go more to school
'I don't go to school any more.'
b. *Oí misa en Catedral.*
I.heard Mass in cathedral
'I heard Mass in the cathedral.'
(Kany 1969: 39)

More recent literature offers additional references to geographically scattered, diatopically marked BCSs in Spanish. RAE/ASALE (2009: 1156, 1159) observed that *no cruces, viene carro* 'don't cross, there is a car coming' is common in Peru, *Casa Presidencial* 'presidential palace' increasingly appears as a BCS in adverbial PPs in Central American varieties, and the same applies to *palacio* 'palace' in Peninsular Spanish. Bosque (1996: 49) noted that Puerto Rican Spanish exhibits regular BCS subjects that elsewhere appear only in proverbs (cf. [20]), a pattern the author later extended to not further identified varieties of Andean Spanish for locative adverbial PPs, such as *el libro está en biblioteca*, lit. 'the book is in library' (Bosque 2021: 16). Severo (2019: 582–585) highlighted a high degree of acceptability for BCS objects in Mexican Spanish but provides only diatopically unmarked data. Lipski (2008: 84) remarked on the relatively free occurrence of BCSs in some semi-creolized varieties of Afro-Bolivian Spanish.

The variety of Spanish most frequently noted to allow diatopically marked BCSs is that of the Río de la Plata, particularly for locative BCSs "denoting dependencies, sections, or internal services of an institution" (RAE/ASALE 2009: 1159; our translation).[3] This tendency is examined by Kuguel and Oggiani (2016), who introspectively identify three groups of BCSs specific to this variety. The first group comprises

3 It has been suggested that this phenomenon could be attributed to Italian-Spanish language contact, which was intense in the region during the late 19th and early 20th centuries and is well-documented for its phonetic and lexical influence on Río de la Plata Spanish (cf. Ennis 2015: 138–139; Fontanella de Weinberg 1987: 136–142). RAE/ASALE (2009: 1159) explicitly considers an Italian-Spanish contact scenario as a possible explanation for certain locative BCS in Río de la Plata Spanish (cf. also Laca 1999: 922).

what Kuguel and Oggiani (2016: 9, 13, 20; our translations) refer to as "generic bare nouns" ([24]; cf. also Stvan 2009), while the second consists of "bare nouns of individuated reference" denoting "unique and identifiable locations within an institution". They can be directional (25) or locative (26). The third group are "activity bare nouns" (27) and are reported to display greater productivity in Uruguay compared to Argentina (Kuguel and Oggiani 2016: 26).

(24) *El querosén se vende en ferretería.*
the kerosene REFL sells in hardware.store
'Kerosene is sold in hardware stores.'

(25) *Voy a rectorado.*
I.go to rectorate
'I go to the rectorate.'

(26) *Sara está en bedelía y Facundo también*
Sara is in jenitor's.office and Facundo too
'Sara is in the janitor's office, and Facundo is too.'

(27) *Pedro no canta cuando está en ruta.*
Peter not sings when he.is on road
'Peter doesn't sing when he's on the road.'
(Kuguel and Oggiani 2016: 10, 19, 22; adapted)

In subsequent introspectively based studies, Oggiani solidifies the status of the Río de la Plata region as a BCS hotspot (Oggiani 2021a, 2021b, 2022; cf. also Aguilar-Guevara and Oggiani 2023, their note 1). Example (28) features another directional Río de la Plata Spanish BCS; (29) a direct object one:

(28) *El médico va a consultorio.*
The doctor goes to doctor's.office
'The doctor goes to the office.'

(29) *Juan tomó ómnibus.*
Juan took bus
'Juan took the bus.'
(Oggiani 2022: 250–251)

Given the cross-linguistic insight that BCSs and SWDs represent two morphosyntactic surface variants of a single underlying semantic phenomenon – semantic pseudo-incorporation (Section 2.1) – and that varieties of the same language can differ in whether they use a given noun as a BCS or an SWD, it would not be surprising if varieties of Spanish allowed for additional, still undocumented BCS distinct from (17)–(29).

3 Detecting diatopically marked Spanish BCSs on Twitter

Using data from Twitter, this section investigates the availability of previously undocumented but theoretically predicted BCSs with *ir a* 'to go to' in diatopic varieties of Spanish. Section 3.1 describes the methodology, Section 3.2 presents the results.

3.1 Twitter study: methodology

To test whether diatopic varieties of Spanish allow for previously undocumented BCSs, we conducted Twitter searches for eight nouns that (i) correspond to SWDs in standard Spanish and (ii) have been shown to exhibit cross-linguistic variation between BCSs and SWDs by Carlson and Sussman (2005), Carlson et al. (2006), Aguilar-Guevara and Zwarts (2010, 2013), Zwarts (2014), Espinal and Cyrino (2017b), Leonetti (2019), and/or Aguilar-Guevara and Oggiani (2023). Taken together, (i) and (ii) suggest that these nouns encode a stereotypical telic component as part of their lexical entry (e.g., a *school* is stereotypically understood as a place where students go to study; cf. Pustejovsky 1995). Finally, (iii) all eight nouns yielded a promising number of BCSs hits in exploratory Twitter searches. All searches were performed using the pay version of the Twitter Streaming API and the *R* package *rtweet*. They excluded retweets and quotes and were limited to Spanish-language tweets produced between November 1, 2016, and August 31, 2017. In the absence of a gold standard for retrieving geographical information from Twitter data (Section 1), the searches did not use any GPS-based geotagging but were instead limited to posts from Twitter users who provided some information in the location field. Unlike GPS-based geotagging, this approach enabled a fine-grained but time-consuming manual decision-making process to determine user origin (see below). This consideration is particularly critical for the present study, as it examines variants not previously documented in the literature and therefore aims to exclude false positives at all costs.

For financial reasons, each search was limited to a maximum of 500 tokens per query and restricted to uses with *ir a* + [noun] 'to go to' + [noun], a well-established locus of cross-linguistic variation between BCSs and SWDs (Section 2). If the initial search with the infinitive *ir* 'to go' did not yield 500 tokens, additional searches were conducted with different verb forms of the paradigm of *ir*, ensuring their presence across all varieties of Spanish. Additionally, identical string searches, again limited to

a maximum of 500 tokens, were conducted for versions with the definite article *el*$_{\text{M.SG}}$ and *la*$_{\text{F.SG}}$ in order to calculate the relative frequencies of BCSs and definite variants for all countries covered by the Twitter data set (cf. Section 3.2).[4]

The eight nouns queried are represented in Table 1; strings for which the queries, despite using additional verb forms, failed to reach the 500-token limit are marked with an asterisk.

The data obtained by the queries are exemplified in (30) for *piscina* 'pool'.

(30) a. *Que rico ir* ***a piscina*** *uhh*
how pleasant to.go **to pool** uhh
'How pleasant to go to **ø** pool'

b. *ahora quiero ir a* ***la piscina***
now I.want to.go to **the pool**
'Now, I want to go to **the** pool'
([30a]: (@alech_58, Colombia; [30b]: (@fvmnaa, Venezuela)

The data obtained were manually assigned to Spanish-speaking countries based on the information provided in the user location field and simultaneously cleaned. During this process, 16.1 % of the collected raw data had to be excluded due to the following reasons:

Table 1: Twitter queries for Spanish *ir a*-BCSs & definites (cleaned).

Noun	Verb forms	*n* BCS	*n* def. art.
escuela 'school'	*ir*$_{\text{INF}}$, *voy*$_{\text{PRS.1SG}}$	426	420
colegio 'high school'	*ir*$_{\text{INF}}$	426	442
cine 'cinema'	*ir*$_{\text{INF}}$	419	455
piscina 'pool'	*ir*$_{\text{INF}}$	375	404
médico 'doctor'	*ir*$_{\text{INF}}$, *voy*$_{\text{PRS.1SG}}$, *fui*$_{\text{PST.1SG}}$, *va*$_{\text{PRS.3SG}}$, *fue*$_{\text{PST.3.G}}$	378*	424
teatro 'theater'	*ir*$_{\text{INF}}$	420	413
peluquería 'hairdresser'	*ir*$_{\text{INF}}$, *voy*$_{\text{PRS.1SG}}$, *fui*$_{\text{PST.1SG}}$, *va*$_{\text{PRS.3SG}}$, *fue*$_{\text{PST.3SG}}$	177*	444
iglesia 'church'	*ir*$_{\text{INF}}$, *voy*$_{\text{PRS.1SG}}$, *fui*$_{\text{PST.1SG}}$, *va* $_{\text{PRS.3SG}}$, *fue*$_{\text{PST.3SG}}$	155*	428
Total		**2,776**	**3,430**
			6,206

4 The definite tokens may represent either SWDs or regular strong definites (cf. Section 2.1). Given that this applies consistently to all definites queried, this is methodologically unproblematic.

1. tweets in languages closely related to Spanish that the API-streaming algorithm did not identify. This issue arose particularly with search strings that are (near) homographs in other languages (e.g., Portuguese *ir à piscina* 'to go to the pool' vs. Spanish *ir a piscina* 'to go to pool'),
2. tweets in which the search string was not part of the tweet but appeared only in a URL included within the tweet,
3. repeated identical tweets containing identical website embeddings shared via the share function,
4. tweets containing proper names (e.g., *voy a (la) peluquería Grace* 'I go to (the) hair salon Grace'). Since proper names are inherently definite (e.g., Lyons 1999, 195–197), making the definite article semantically pleonastic, it is uncertain if the BCS/definite article variation in *voy a (la) peluquería Grace* can be analyzed in the same way as that in *voy a (la) peluquería*,
5. tweets in which, for other reasons, the information in the location field proved insufficient to unambiguously determine the user's origin.

The detailed, step-by-step manual annotation and decision-making process used to determine the geographical data origin is outlined in Table A.1 in the Appendix Section.

3.2 Twitter study: results

This section presents the results of the Twitter study described in Section 3.1, both descriptively (Section 3.2.1) and by means of conditional inference tree and random forest modelling (Section 3.2.2).

3.2.1 Twitter study: descriptive statistical results

Table 2 presents the descriptive statistical results for all eight queried nouns, ordered by country from the highest to the lowest percentage of the BCS variant. Since many countries yielded low overall *n* (see Section 5 for discussion), Table 2 includes only data from countries that meet an arbitrarily established threshold of at least 40 tokens for the BCS variant and the definite variant combined. The percentages in the rightmost column indicate the proportion of data from each country relative to the overall dataset for the respective noun, including data from countries excluded from Table 2 due to having fewer than 40 tokens. For six out of the eight nouns, the exclusion of countries with fewer than 40 tokens means that more than 80 % of the data are represented in Table 2. The two exceptions are *teatro* 'theater' (76.7 %) and

Table 2: Twitter data: BCS versus definite.

Noun	Country	BCS	Definite article	Total
escuela	ARG	243 (50.4 %)	239 (49.6 %)	482 (57.0 %)
'school'	MEX	62 (29.7 %)	147 (70.3 %)	209 (24.7 %)
			Total	691/846 (81.7 %)
colegio	ARG	342 (45.6 %)	408 (54.4 %)	750 (86.4 %)
'high school'			**Total**	750/868 (86.4 %)
cine	COL	388 (99.7 %)	1 (0.3 %)	389 (44.5 %)
'cinema'	MEX	6 (8.3 %)	66 (91.7 %)	72 (8.2 %)
	ARG	7 (2.4 %)	284 (97.6 %)	291 (33.3 %)
			Total	752/874 (86.0 %)
piscina	COL	330 (96.5 %)	12 (3.5 %)	342 (43.9 %)
'swimming pool'	VZL	10 (13.3 %)	65 (86.7 %)	75 (9.6 %)
	SPA	12 (4.3 %)	264 (96.7 %)	276 (35.4 %)
			Total	693/779 (89.0 %)
iglesia	VZL	23 (50 %)	23 (50 %)	46 (7.9 %)
'church'	COL	17 (29.3 %)	41 (70.7 %)	58 (9.9 %)
	ARG	44 (24.3 %)	137 (75.7 %)	181 (31.0 %)
	MEX	16 (18 %)	73 (82 %)	89 (15.3 %)
			Total	374/583 (64.2 %)
médico	CHL	225 (93.8 %)	15 (6.3 %)	240 (29.9 %)
'doctor'	SPA	28 (33.3 %)	56 (66.7 %)	84 (10.5 %)
	ARG	81 (22.9 %)	272 (77.1 %)	353 (44.0 %)
			Total	677/802 (84.4 %)
teatro	COL	141 (94.6 %)	8 (5.4 %)	149 (17.9 %)
'theater'	ARG	201 (55.8 %)	159 (44.2 %)	360 (43.2 %)
	SPA	12 (27.9 %)	31 (72.1 %)	43 (5.2 %)
	MEX	17 (13.6 %)	108 (86.4 %)	125 (15.0 %)
			Total	677/883 (76.7 %)
peluquería	ARG	113 (28.1 %)	289 (71.9 %)	402 (64.7 %)
'hairdresser'	SPA	11 (17.2 %)	53 (82.8 %)	64 (10.3 %)
			Total	466/621 (75.0 %)

iglesia 'church' (64.2 %). Overall, Table 2 accounts for 81.9 % of all retrieved and cleaned data (n = 5,080/6,206).[5]

Table 2 reveals some interesting initial findings: Argentina ranks as the top BCS-producing country for three of the eight queried nouns (*escuela* 'school', *colegio* 'high school', and *peluquería* 'hairdresser'). Meanwhile, Colombia leads for three nouns (*cine* 'cinema', *piscina* 'swimming pool', and *teatro* 'theater'), while Chile ranks

5 Abbreviations in Table 2: ARG = Argentina, MEX = Mexico, COL = Colombia, VZL = Venezuela, SPA = Spain, CHL = Chile.

highest for *médico* 'doctor' and Venezuela for *iglesia* 'church'. Notably, among all these countries, only Argentina has been highlighted in the literature, based on introspection, as using diatopically marked Spanish BCSs (Section 2.2). Table 3

Table 3: Twitter data: top BCS-countries (*n* BCS + definite ≥ 40)

Noun	Top-BCS country	% BCS	*n* BCS||*n* definite article
escuela	Argentina	50.4	243||239
colegio	Argentina	54.4	342||408
peluquería	Argentina	71.9	113||289
cine	Colombia	99.7	388||1
piscina	Colombia	96.5	330||12
teatro	Colombia	94.6	141||8
médico	Chile	93.8	225||15
iglesia	Venezuela	50.0	23||23

Figure 1: Graphic visualization of top BCS-countries in Twitter data (*n* BCS + definite ≥ 40).

summarizes the top BCS-producing countries by noun; Figure 1 provides a graphical representation of the top BCS percentages for each noun.

3.2.2 Twitter study: conditional inference & random forest modelling

The standard approach to inferential statistical data analysis would now typically involve regression modeling. However, in the present case, this is not feasible due to the nature of the dataset. Even when restricting the analysis to the six countries included in Table 2, the data would contain cases of complete separation, known to distort the results of regression modeling (Levshina 2015: 273). For example, the Chilean data for *escuela* 'school' and the Venezuelan data for *colegio* 'high school' consist exclusively of BCS tokens (n = 12, n = 5, respectively) with no corresponding definites. An alternative method robust in cases of complete separation is conditional inference tree modeling combined with random forests. Conditional inference tree modeling recursively partitions a dataset into subsets based on significant associations between predictor variables and the dependent variable, creating a tree-like structure. The method selects splitting variables based on statistical criteria (e.g., p-values) derived from permutation tests, ensuring unbiased variable selection and interpretability of the resulting decision rules. This approach is particularly advantageous as it performs well even when certain predictor combinations have low observation counts (Levshina 2015: 292).

For methodological reasons, the conditional inference tree and random forest modeling of the Twitter dataset must include additional tokens not listed in Table 2. These tokens come from noun-country combinations that (i) do not meet the ≥40 threshold but (ii) meet the threshold for another noun from the same country. These 436 additional tokens are detailed in Table 4. Including these additional data results in the conditional inference tree and random forest dataset covering 5,516 of the total 6,206 tokens (88.9 %).

Figure 2 displays the conditional inference tree modeled based on the data from Tables 2 and 4, where the response variable is 'BCS' versus 'definite', and the predictor variables are 'country' and 'noun'.[6] The tree is interpreted as follows: Node 1 divides the dataset into two groups of countries – Chile and Colombia on one side, and Argentina, Mexico, Spain, and Venezuela on the other. At Node 2, tokens from Chile and Colombia are further split based on nouns: *iglesia* and *peluquería* tokens from these two countries only have a 21.4 % likelihood of being BCSs (Node 3).

6 To maintain simplicity in the conditional inference tree, we did not include 'speaker' as a predictor. 5,149 of the 5,516 tokens were produced by distinct Twitter users. The highest number of posts by a single user was n = 7; additionally, there were 4 users with 6 posts, 5 users with 5 posts, 12 users with 4 posts, 47 users with 3 posts, and 299 users with 2 posts.

Table 4: Additional Twitter data included in conditional inference tree.

Noun	Country	BCS	Definite article	Total
escuela	CHL	12 (100 %)	–	12 (1.4 %)
'school'	COL	7 (77.8 %)	2 (22.2 %)	9 (1.1 %)
	SPA	14 (70 %)	6 (30 %)	20 (2.4 %)
	VZL	8 (88.9 %)	1 (11.1 %)	9 (1.1 %)
			Total	50/846 (5.9 %)
colegio	CHL	13 (72.2 %)	5 (27.8 %)	18 (2.1 %)
'high school'	COL	14 (70 %)	6 (30 %)	20 (2.3 %)
	MEX	6 (54.5 %)	5 (45.5 %)	11 (1.3 %)
	SPA	12 (75 %)	4 (25 %)	16 (1.8 %)
	VZL	5 (100 %)	–	5 (0.6 %)
			Total	70/868 (8.1 %)
cine	CHL	3 (42.9 %)	4 (57.1 %)	7 (0.8 %)
'cinema'	SPA	1 (4 %)	24 (96 %)	25 (2.9 %)
	VZL	6 (22.2 %)	21 (77.8 %)	27 (3.1 %)
			Total	59/874 (6.8 %)
piscina	ARG	2 (40 %)	3 (60 %)	5 (0.6 %)
'swimming pool'	CHL	–	6 (100 %)	6 (0.8 %)
	MEX	1 (33.3 %)	2 (67.6 %)	3 (0.4 %)
			Total	14/779 (1.8 %)
iglesia	CHL	4 (17.4 %)	19 (82.6 %)	23 (3.9 %)
'church'	SPA	3 (11.5 %)	23 (88.5 %)	26 (4.5 %)
			Total	49/583 (8.4 %)
médico	COL	10 (34.5 %)	19 (65.5 %)	29 (3.6 %)
'doctor'	MEX	10 (45.5 %)	12 (54.5 %)	22 (2.7 %)
	VZL	9 (40.9 %)	13 (59.1 %)	22 (2.7 %)
			Total	73/802 (9.1 %)
teatro	CHL	2 (8 %)	23 (92 %)	25 (2.8 %)
'theater'	VZL	5 (25 %)	15 (75 %)	20 (2.3 %)
			Total	45/883 (5.1 %)
peluquería	CHL	3 (14.3 %)	18 (85.7 %)	21 (3.4 %)
'hairdresser'	COL	3 (12.5 %)	21 (87.5 %)	24 (3.9 %)
	MEX	2 (25 %)	6 (75 %)	8 (1.3 %)
	VZL	10 (43.5 %)	13 (65.5 %)	23 (3.7 %)
			Total	76/621 (12.2 %)

Colombian and Chilean *colegio*, *escuela*, *médico*, and *teatro* tokens, in turn, have a 84.5 % likelihood of being BCSs (Node 5). Additional noteworthy insights regarding high BCS percentages in the dataset can be observed at the following nodes: Nodes 9 and 10 indicate that *cine* and *piscina* tokens from Colombia are almost exclusively BCSs; Node 16 reveals that *colegio*, *escuela*, and *teatro* tokens from Argentina, Spain,

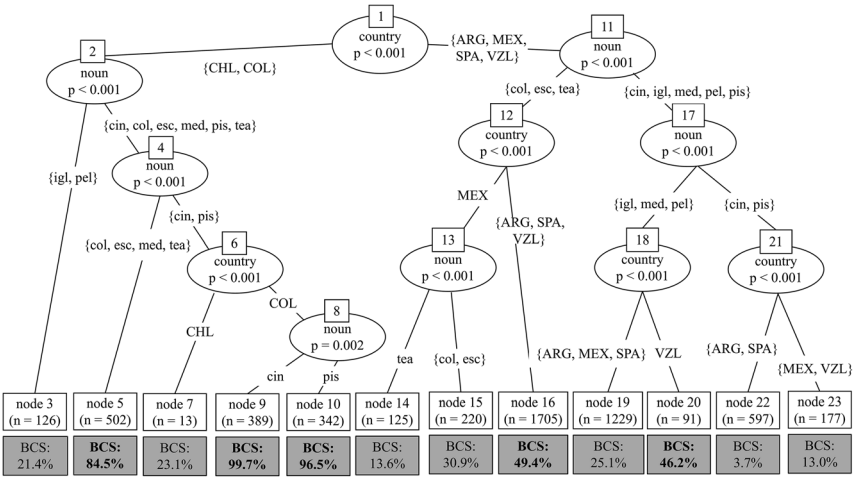

Figure 2: Twitter data (Tables 2 and 4): conditional inference tree (BCS vs. definite ~ country + noun).

and Venezuela are evenly distributed between BCS and definite forms; and Node 20 shows that *iglesia*, *médico*, and *peluquería* tokens from Venezuela are also nearly evenly distributed between BCSs and definite forms.

Building on the conditional inference tree model in Figure 2, Figure 3 displays the random forest, which yields the importance measure for every predictor variable in the model averaged over multiple conditional trees, each trained on a random subset

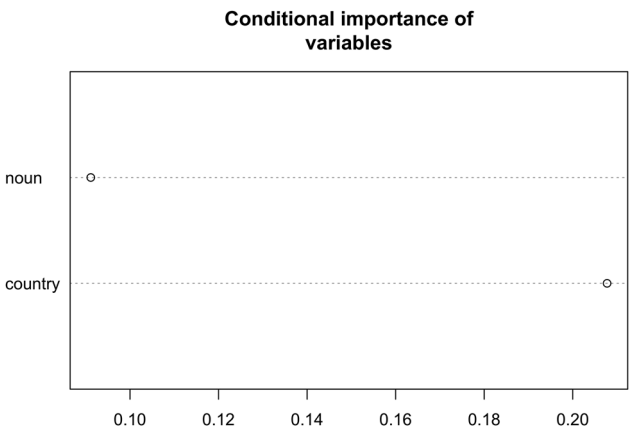

Figure 3: Twitter data (Tables 2 and 4): random forest.

of the data and predictor variables (cf. Levshina 2015: 292). Figure 3 shows that the overall importance of *country* (0.208) as a predictor is higher than that of *noun* (0.091).

In summary, the numerous cases of high percentages of the BCS variant compared to the definite variant in the Twitter data analyzed in this section suggest the widespread availability of many *ir a*-BCSs across different countries. Among these, the most prominent are *ir a* 'to go to' + *escuela* 'school', *colegio* 'high school', and *peluquería* 'hairdresser' in Argentina; *ir a* + *cine* 'cinema', *piscina* 'swimming pool', and *teatro* 'theater' in Colombia; *ir a médico* 'to go to the doctor' in Chile; and *ir a iglesia* 'to go to church' in Venezuela (Table 3). Notably, the relative frequencies of the BCS variants in the dataset are much higher for the Colombian and Chilean cases than for the Argentinian and Venezuelan ones, a descriptive statistical finding also supported inferentially by conditional inference tree modeling. Furthermore, other high percentages of the BCS variant for different noun-country combinations are observed in the Twitter data (Table 2), though severe data scarcity is evident in some cases (Table 4).

If the findings reported in the present section correspond to the acceptability of specific *ir a*-BCSs in the respective countries, this would align with previous literature in the case of Argentina (Section 2.2). By contrast, the cases of Colombia, Chile, and Venezuela would, to the best of our knowledge, represent instances of theoretically predicted but previously undocumented Spanish *ir a*-BCSs. Since the research question of this paper precisely is whether data from social media platforms can serve as a viable resource for identifying and geographically mapping such theoretically predicted but previously undocumented morphosyntactic variants, the following section aims to validate some of the insights derived from the Twitter data through an additional acceptability judgment experiment. Among other aspects discussed in Section 5, this is, above all, necessary due to a complete lack of transparency regarding the Twitter search algorithm (cf. Section 1): we do not know whether the results obtained in this section are affected by undisclosed data pre-selection processes or other operations that may potentially skew the returned Twitter data.

4 An experimental pilot study for validating diatopically marked Spanish *ir a*-BCS data obtained via Twitter

This section presents a first pilot acceptability judgment task conducted for the eight *ir a*-BCSs examined using Twitter data, focusing on a subset of the countries covered by Section 3. Section 4.1 outlines the methodology, while Section 4.2 reports the results.

4.1 Experimental pilot study: methodology

In response to the issues outlined at the end of Section 3.2.2, we designed a pilot acceptability judgment experiment to further explore the Twitter-based findings presented in Section 3. It is important to emphasize that this experiment is not intended to serve as a fully-fledged experimental study in its own right but rather as a complementary step to assess the plausibility of the Twitter-based geolinguistic results regarding previously undocumented yet theoretically predicted morphosyntactic variation.

The pilot acceptability judgment task focused on the same *ir a*-BCSs examined in Section 3 and was conducted online in July and August 2021 with 310 native Spanish speakers, each compensated $3.00 for their participation. Data cleaning, as described below, led to the exclusion of 84 participants, resulting in a final sample of 226 participants considered for analysis. The experiment included participants from the following countries, where they were required to have been born and, at the time of the experiment, reside: Spain (n = 29), Colombia (n = 38), Argentina (with a distinction between Buenos Aires [city and province] and non-Río de la Plata varieties [n = 43; n = 33]), Venezuela (n = 42), and Chile (n = 41). This selection ensured the inclusion of all countries that, according to the Twitter data, exhibited the highest percentages of the BCS variant for one or more of the eight nouns (cf. Table 3). Spain was included because the Twitter data revealed BCS percentages as high as 33.3 % compared to the definite variant (cf. Table 2), despite the fact that, prescriptively, the *ir a*-BCSs under study are not considered acceptable in European Spanish (cf. Section 2.2). Ideally, Mexico would have been included for similar reasons (cf. Table 2), but financial constraints made this infeasible. The decision to distinguish between Río de la Plata and non-Río de la Plata varieties of Argentine Spanish was motivated by prior research identifying Río de la Plata Spanish as a hotspot for BCS usage (Section 2.2). More fine-grained geolinguistic distinctions were not systematically pursued due to financial limitations; however, participants were asked to provide their specific place of birth (city/village) and current place of residence (city/village). Additionally, participants were asked to specify their age and gender and to provide information on whether their parents spoke any other first language besides Spanish.

Participants were evenly divided into two groups, each presented with four different *ir a*-BCS items embedded in short sentences in the present tense and third person singular. Together with a number of fillers (see below), this design aimed to prevent participants from detecting the phenomenon under investigation, an issue that was addressed at the end of the experiment by asking participants to guess the purpose of the study. Only 5 out of 310 participants were excluded based on responses suggesting they had noticed the lack of the definite article in some of the tested items.

After completing five practice rounds to familiarize themselves with the experimental procedure and following an explicit statement that there were no right

Table 5: Experimental *ir a*-BCS items.

Group 1	Group 2
María va a teatro.	*María va a médico.*
'María goes to theater'	'María goes to doctor'
Lucía va a piscina.	*Luis va a peluquería.*
'Lucía goes to swimming pool'	'Luis goes to hairdresser'
Alejandro va a cine.	*Lucía va a iglesia.*
'Alejandro goes to cinema'	'Lucía goes to church'
Álvaro va a colegio.	*Alberto va a escuela.*
'Álvaro goes to high school'	'Alberto goes to school'

or wrong answers, participants were instructed to rate the naturalness of each item on a 7-point Likert scale. Table 5 provides an overview of the eight test items, one for each noun, along with their distribution across participant groups.

The four *ir a*-BCS items per group were organized into four randomized blocks, each containing one test item and seven filler items, with the position of the test item and fillers within each block fixed. This arrangement ensured that no two test items appeared consecutively. The 24 filler items (4 blocks × 7) are listed in Table A.2 in the Appendix Section. Among the fillers, eight were perfectly grammatical, eight were completely ungrammatical, and eight included minor linguistic errors or constructions uncommon in most varieties of Spanish. This design served three purposes. First, the fillers further obscured the phenomenon being tested in the experiment. Second, they served as "anchor items for certain points on the scale" (Schütze and Sprouse 2014: 33), encouraging participants to utilize the full range of the 7-point Likert scale and thereby reducing the risk of response biases and rating compressions. Third, the fillers were used to assess whether participants were engaging seriously with the task. If a participant rated a perfectly grammatical filler below 5, the response was flagged as suspicious. Similarly, if a clearly ungrammatical filler was rated higher than 4, the response was also flagged as suspicious. Participants with more than three suspicious responses in total were excluded from the analysis, which accounted for the majority of the 84 exclusions mentioned above.

4.2 Experimental pilot study: results

This section presents the results of the experimental pilot acceptability judgment task described in Section 4.1, both descriptively (Section 4.2.1) and by means of ordinal regression modelling (Section 4.2.2).

4.2.1 Experimental pilot study: descriptive statistical results

Figures 4a–4d display the results of the experimental pilot acceptability judgment task described in Section 4.1, organized by noun and country. Means and medians of ≥4 are highlighted in red, while standard deviations are indicated in green.

Setting a rating of '4' as an arbitrary threshold, the *ir a*-BCSs with *teatro* 'theater', *piscina* 'pool', and *cine* 'cinema' received high experimental mean and median acceptability ratings in Colombia (M = 4.78; Mdn = 5|M = 5.89; Mdn = 7|M = 5.44; Mdn = 7). Importantly, however, the ratings for *piscina* exhibit polarization at the extremes of the scale: the relatively high Colombian mean and median result from many participants giving very high ratings, but they somewhat obscure the fact that quite a few participants also provided very low ones and, crucially, that very few assigned intermediate ratings. The remaining Colombian experimental data show rather low acceptability ratings (*iglesia* 'church': M = 2.6; Mdn = 1.5; *escuela* 'school': M = 2.73; Mdn = 2; *médico* 'doctor': M = 2.35; Mdn = 2; *colegio* 'high school': M = 3.67; Mdn = 2.5; *peluquería* 'hairdresser': M = 3.7; Mdn = 3.5). Again, *colegio*, *iglesia*, and *peluquería* show strong polarization, with some participants rating the *ir a*-BCS very highly despite the low overall means and medians.

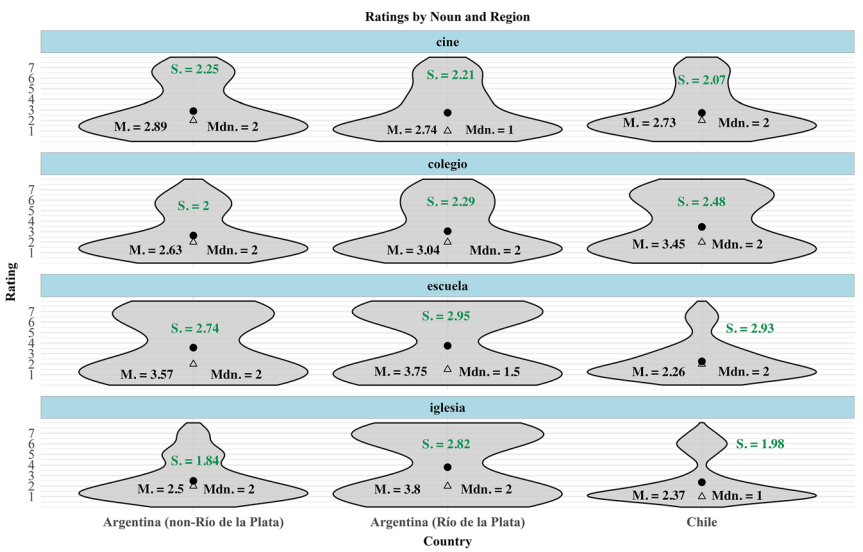

Figure 4a: Experimental results: Argentina & Chile I (means, medians, standard deviations).

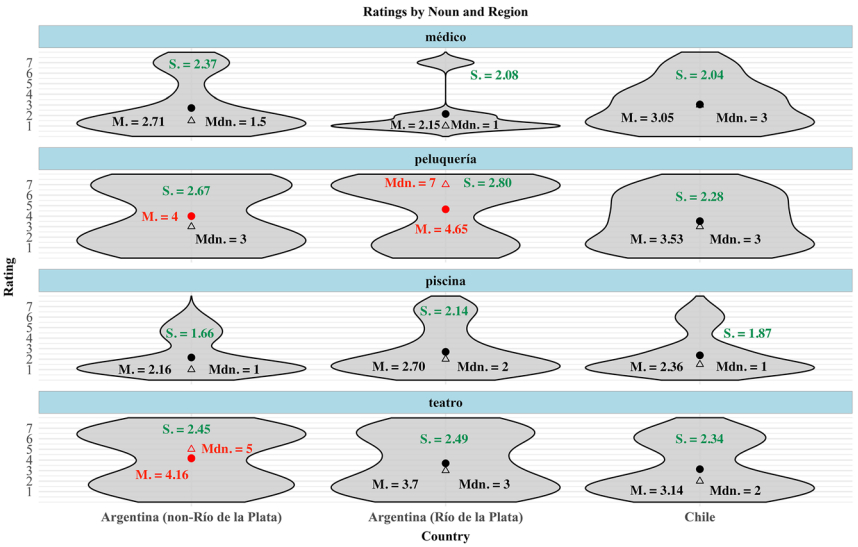

Figure 4b: Experimental results: Argentina & Chile II (means, medians, standard deviations).

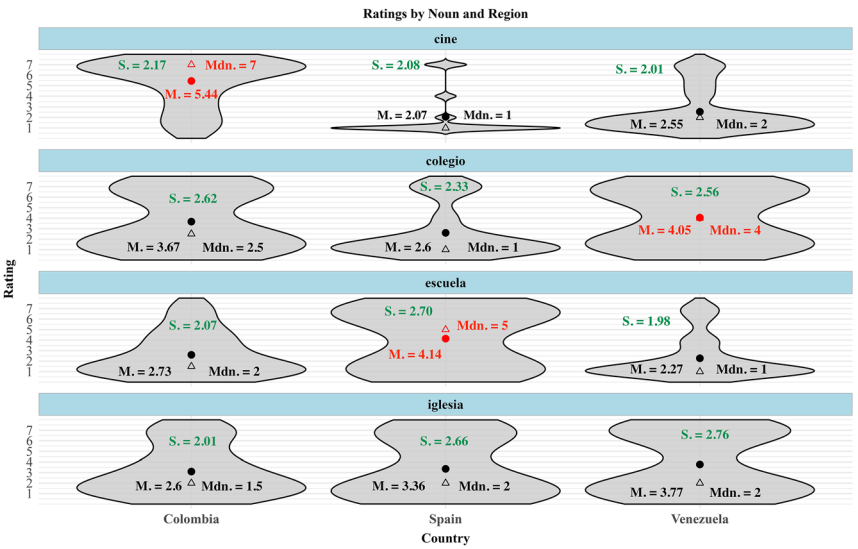

Figure 4c: Experimental results: Colombia, Spain, Venezuela I (means, medians, standard deviations).

Argentinian participants from non-Río de la Plata regions provided a mean and median greater than '4' for *teatro* (M = 4.16; Mdn = 5) and a mean of '4' for *peluquería* (Mdn = 3). Strong polarization at the extremes of the scale is evident for both nouns.

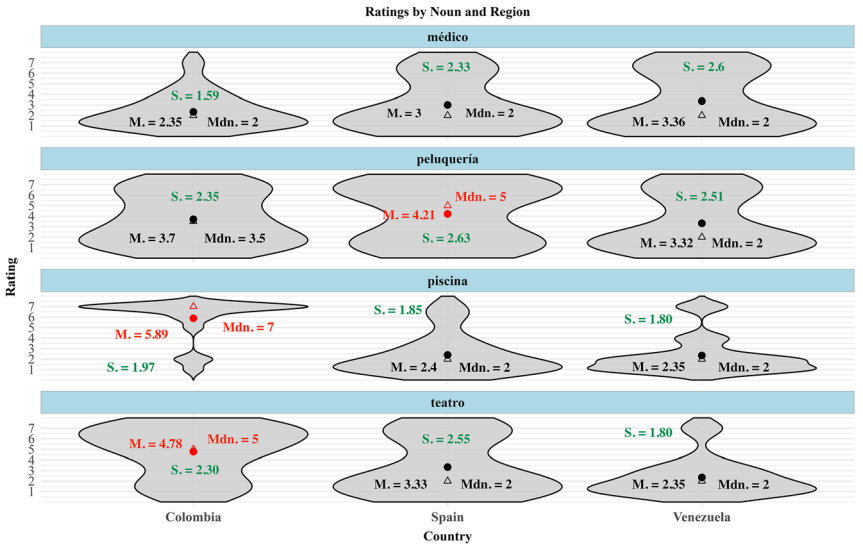

Figure 4d: Experimental results: Colombia, Spain, Venezuela II (means, medians, standard deviations).

The remaining non-Río de la Plata experimental ratings are low (*cine*: *M* = 2.89; Mdn = 2; *escuela*: *M* = 3.57; Mdn = 2; *colegio*: *M* = 2.63; Mdn = 2; *iglesia*: *M* = 2.5; Mdn = 2; *médico*: *M* = 2.71; Mdn = 1.5; *piscina*: *M* = 2.16; Mdn = 1). Polarization, with some participants providing high ratings, is true for all of these latter *ir a*-BCSs to differing extents and is especially pronounced for *escuela*.

Argentinian participants from Río de la Plata regions assigned high ratings to *peluquería* (*M* = 4.65; Mdn = 7), though, once again, with strong polarization at both extreme ends of the scale. The remaining Río de la Plata experimental ratings are low (*cine*: *M* = 2.74; Mdn = 1; *escuela*: *M* = 3.75; Mdn = 1.5; *colegio*: *M* = 3.04; Mdn = 2; *iglesia*: *M* = 3.8; Mdn = 2; *médico*: *M* = 2.15; Mdn = 1; *piscina*: *M* = 2.70; Mdn = 2). The least polarized cases among the latter *ir a*-BCS ratings are those of *médico* and *piscina*, though even for these, some participants provided high ratings.

In the Venezuelan data, the only *ir a*-BCS receiving high acceptability ratings was *colegio* (*M* = 4.05; Mdn = 4), though with pronounced polarization at the extreme ends of the rating scale. The remaining Venezuelan acceptability ratings are all below '4' (*cine*: *M* = 2.55; Mdn = 2; *escuela*: *M* = 2.27; Mdn = 1; *iglesia*: *M* = 3.77; Mdn = 2; *médico*: *M* = 3.36; Mdn = 2; *peluquería*: *M* = 3.32; Mdn = 2; *piscina*: *M* = 2.35; Mdn = 2; *teatro*: *M* = 2.35; Mdn = 2). Strong polarization, however, is notable for *iglesia*, *médico*, and *peluquería*.

For Spain, the experimental data show relatively high means and medians for *peluquería* and *escuela* (*M* = 4.21; Mdn = 5|*M* = 4.14; Mdn = 5). All other Spanish experimental means and medians are low (*cine*: *M* = 2.07; Mdn = 1; *colegio*: *M* = 2.6;

Mdn = 1; *iglesia*: *M* = 3.36; Mdn = 2; *médico*: *M* = 3; Mdn = 2; *piscina*: *M* = 2.4; Mdn = 2; *teatro*: *M* = 3.33; Mdn = 2). Among the Spanish data, the most polarized cases are those of *colegio*, *escuela*, *iglesia*, *médico*, *peluquería*, and *teatro*.

As for Chile, the experimental acceptability ratings are low across the board for all *ir a*-BCSs tested (*cine*: *M* = 2.73; Mdn = 2; *colegio*: *M* = 3.45; Mdn = 2; *escuela*: *M* = 2.26; Mdn = 2; *iglesia*: *M* = 2.37; Mdn = 1; *médico*: *M* = 3.05; Mdn = 3; *peluquería*: *M* = 3.53; Mdn = 3; *piscina*: *M* = 2.36; Mdn = 1; *teatro*: *M* = 3.14; Mdn = 2). Ratings are most polarized for *colegio*, *médico*, *peluquería*, and *teatro*, and, to a lesser extent, also for *cine*.

As a final point, note that all experimental data exhibit high standard deviations, ranging from 1.66 to 2.95.

4.2.2 Experimental pilot study: ordinal regression model

To examine the descriptive statistical experimental results from Section 4.2.1 more closely, we attempted to fit a mixed-effects ordinal regression model with 'rating' as the dependent variable and a three-way interaction between 'country', 'noun', and 'age' (modeled as a factor with three levels: '18–30', '31–42', and '>42'), while controlling for speaker as a random effect. However, this model encountered numerical issues, including undefined standard errors, a singular Hessian matrix, and warnings related to convergence criteria. These problems likely stem from the model's complexity, sparse data for certain predictor combinations, and numerical instability in estimating parameters. An alternative model with two two-way interactions ('country*noun' and 'age*country') was subsequently fitted, yielding a converging model. This second model, however, also encountered optimization issues, including step factor reductions and iteration limits, indicating challenges in reliably estimating parameters. These difficulties are again likely attributable to the model's complexity, sparse data for certain combinations of predictors, and potential overparameterization. Given the unreliability of the more complex models – and even though we briefly comment on some of the results of the second model below – we proceeded with a reduced mixed-effects ordinal regression analysis. In this third and final model, 'rating' was the dependent variable, and the predictors included 'country' and 'noun' as main effects and their interaction. Additionally, 'speaker' was included as a random effect to account for individual variability among participants. 'Chile' was chosen as the reference level for 'country' because it was the only country with neither a mean nor a median rating ≥4 (cf. Figures 4a and 4b). For 'noun', the reference level was 'médico', as this noun had the lowest mean rating across all 226 experiment participants (*M* = 2.77). The statistically significant results of this reduced model are presented in Table 6; the symbols *, **, and *** stand for 'significant', 'highly significant', and 'very highly significant', respectively. Detailed model diagnostics are provided in the Appendix Section.

Table 6: Experimental *ir a*-BCS acceptability: mixed-effects ordinal regression model (statistically significant results only).

Model	Rating ~ country * noun + (1|Response.ID)					
	Link: logit					
	Threshold: flexible					
	Number of observations (nobs) = 904					
	Log-likelihood = −1,321.93					
	AIC = 2,751.87			Random effects	Groups: speaker	
	Number of iterations (niter) = 11,597 (67,835)				Variance = 3.275	
	Maximum gradient = 9.70e−04				Std. Dev. = 1.81	
	Condition number of Hessian = 2.6e+04				Group Num. = 226	
		Estimate	**SE**	**OR|CI (95 %)**	***z* val**	**Pr (>|*z*|)**
Coef.	Colombia:cine	3.38512	1.23562	29.52|2.62–332.58	2.740	<0.01**
	Colombia:teatro	2.55947	1.22360	12.93|1.17–142.26	2.092	<0.05*
	Colombia:piscina	4.50955	1.26810	90.88|7.57–1091.14	3.556	<0.001***
	Argentina_Río.de.la.Plata:escuela	2.61998	1.00278	13.74|1.92–98.04	2.613	<0.01**
	Argentina_Río.de.la.Plata:iglesia	2.95555	0.99526	19.21|2.73–135.13	2.970	<0.01**
	Argentina_Río.de.la.Plata:peluquería	2.95417	1.00162	19.19|2.69–136.63	2.949	<0.01**

The results of the reduced mixed-effects ordinal regression model reported in Table 6 do not show any statistically significant effects for either 'country' or 'noun' as main effects. Statistically significant effects, however, emerged for the nouns *cine*, *teatro*, and *piscina* in Colombia, as well as for *escuela*, *iglesia*, and *peluquería* in Río de la Plata-Argentina. Compared to the reference level combination (*médico* in Chile), these noun-country combinations are significantly more likely to receive a higher rating. For instance, taking *cine* in Colombia as an example, the interaction between country and noun results in an odds ratio of approximately 29.52. In other words, the odds of receiving a rating of, for instance, '4' or higher for *cine* in Colombia are nearly 30 times greater than for *médico* in Chile. It is important to note, however, the relatively wide confidence intervals in all cases (e.g., for *cine* in Colombia: 2.62–332.58).[7] Other statistically significant results include *teatro* and *piscina* in Colombia,

7 Confidence intervals for the odds ratios indicate high variability in the data, as also indicated by the large standard errors of the log-odds estimates, potentially reflecting limited sample size, model complexity, or sparse data for certain predictor combinations. This sparsity may also contribute to the relatively low McFadden's R^2 value of 0.039 (cf. Appendix Section – Detailed model diagnostics).

with odds ratios of 12.93 and 90.88, respectively, and *iglesia*, *escuela*, and *peluquería* in Río de la Plata-Argentina, with odds ratios of 19.21, 13.74, and 19.19, respectively. It is also worth highlighting that the random effect of 'speaker' plays a crucial role in the model. With an estimated variance of 3.275 and a standard deviation of 1.81, this random effect indicates substantial individual variability across participants, as already noted descriptively throughout Section 4.2.1 in the context of high rating polarizations of many noun-country combinations.

As noted at the beginning of this section, besides the model reported in Table 6, we also managed to fit an alternative, more complex ordinal regression model (main effects: 'country', 'noun', and 'age'; two-way interactions: 'country*noun' and 'age*-country'). Although this model proved problematic and unreliable in several respects, as previously mentioned, we still consider it worthwhile to briefly comment on some aspects related to it. First, the results from the alternative model were largely consistent with those of the reduced model shown in Table 6: the six noun-country combinations that were significant in the reduced model remained significant, with slightly lower coefficients for the Argentinian Río de la Plata combinations and slightly higher ones for the Colombian combinations, maintaining the same significance levels. No additional noun-country combinations reached significance, nor did 'country' or 'noun' as main effects. The only additional significant results that emerged were, first, that independently of the country, the youngest age group exhibited a significantly higher probability of assigning higher ratings to the *ir a*-BCS items compared to the oldest age group from Chile, which served as the reference category (log-odds: 2.13; odds ratio: 8.42; $p = 0.014$). Secondly, in contrast to this overall trend, the youngest age group from Colombia displayed an opposing tendency, being more likely to assign lower ratings than the reference group (log-odds: –2.75; odds ratio: 0.06; $p = 0.029$). However, we emphasize that these results should be interpreted with extreme caution due to the issues identified with this more complex alternative model.

5 Discussion: using Twitter for detecting theoretically predicted but previously undocumented morphosyntactic variants?

The research question addressed in this paper is whether social media platforms such as Twitter/X can be used to detect and geographically map theoretically predicted but empirically previously undocumented morphosyntactic variation. To this end, Section 3 presented a Twitter-based study on Spanish BCS with *ir a* + eight different nouns (lit. 'to go to + [noun]'; $n = 6{,}206$). Given the lack of transparency regarding the Twitter search algorithm mentioned in Section 1, Section 4 sought to

Table 7: Comparison of statistically significant *ir a*-BCS experimental acceptability results and Twitter data-based results

	Exp. mean|median	Percentage of BCS-variant in Twitter study
Colombia: *cine*	5.44|7	99.7 % (top BCS-variant country)
Colombia: *teatro*	4.78|5	94.6 % (top BCS-variant country)
Colombia: *piscina*	5.89|7	96.5 % (top BCS-variant country)
Argentina (R. Pl.): *escuela*	3.75|1.5	50.4 % (Argentina top BCS-variant country)
Argentina (R. Pl.): *iglesia*	3.8|2	24.3 % (Argentina 3rd for BCS-variant)
Argentina (R. Pl.): *peluquería*	4.65|7	71.9 % (Argentina top BCS-variant country)

validate part of these Twitter-based results through a pilot acceptability judgment experiment involving 226 Spanish speakers from six different countries and regions of the Spanish-speaking world. With the aim of answering the research question, the present section comparatively discusses the results from both empirical studies.

The pilot acceptability judgment experiment identified six noun-country combinations for which the likelihood of receiving higher ratings for a given value than the reference category (*médico* in Chile) reached statistical significance. These combinations are summarized in Table 7, along with their experimental means and medians and the corresponding Twitter results.

Table 7 shows that all statistically significant experimental results align with the Twitter data: Colombian experimental participants were significantly more likely to assign higher ratings to *cine*, *teatro*, and *piscina*, which corresponds to Colombia being the top BCS-variant country for these nouns in the analyzed Twitter data. The same is true for *escuela* and *peluquería* in Río de la Plata Argentina. The sixth noun-country combination to receive significantly higher acceptability ratings in the experiment, *iglesia* in Río de la Plata Spanish, also exhibited a substantial percentage of the BCS variant in Argentinian Twitter data. In addition, Table 7 shows that the differences between the statistically significant acceptability ratings closely mirror those between the different BCS-variant percentages in the Twitter data: The Colombian *ir a*-BCSs exhibit the highest experimental acceptability means and medians and the highest BCS percentages in the Twitter data, whereas the Argentinian Río de la Plata *ir a*-BCSs display lower experimental acceptability means and medians and lower Twitter BCS percentages. Simplifying over this latter issue for the moment (but see below), the findings in Table 7 thus already support an affirmative answer to the research question posed in this paper. They show that social media platforms such as Twitter can indeed be successfully used to detect and geographically map theoretically predicted but previously undocumented morphosyntactic variants. Drawing on Section 2.2, Colombia had not previously been associated in the literature with the use of diatopically marked Spanish BCSs of any kind. The same

applies to *ir a escuela*, *ir a iglesia*, and *ir a peluquería*, which have not previously been documented as grammatical BCSs in any variety of Spanish, including Argentinian Río de la Plata Spanish. That our findings seem to indicate the acceptability of these BCSs in Argentinian Río de la Plata Spanish is, however, less surprising than what we observed in the Colombian cases: On a more general level, the Río de la Plata region is known to be a BCS hotspot in the Spanish-speaking world (Section 2.2).

The statistically significant experimental acceptability ratings from Table 7 do not, however, include all top *ir a*-BCS-variant noun-country combinations reported in Table 3 of Section 3.2.1. This suggests that the overall affirmative answer to the research question may need to be nuanced. To this end, Table 8 lists all 21 noun-country combinations that did not achieve significantly higher acceptability ratings in the experiment but also exhibited *ir a*-BCS-variant percentages above 20 % in the Twitter data.[8] Countries with fewer than 40 combined observations for the BCS and definite variants in the Twitter data are marked with *; *R. Pl.* stands for 'Río de la Plata'.

The question that arises in view of Table 8 is whether these noun-country combinations represent cases in which the analyzed Twitter data yield false positive results concerning *ir a*-BCS acceptability. This could, for instance, be due to undisclosed pre-selection processes in Twitter's search algorithm returning a disproportionately high number of BCS data explicitly queried for, which, in reality, represent infrequent performance errors. While this may well be true for some cases in Table 8, an across-the-board conclusion in this sense may also be overly simplistic.

A first possible non-algorithm-related explanation for high Twitter *ir a*-BCS-variant percentages that are inconsistent with the lack of statistical significance of higher experimental ratings, such as those in Table 8, could be that during the time span considered in the analyzed data, Twitter imposed a limitation of a maximum of 140 characters per post (increased to 280 characters since September 2017). Given this technical constraint, it would not be too surprising if Twitter users omitted definite articles to save space and thus 'accidentally' produced BCSs, which then, in experimental acceptability judgment tasks like the one reported in Section 4 (where the medium of the utterance was not specified), received low acceptability ratings. In such cases, the problematic instances from Table 8 would represent omissions of functional elements similar to what happens with bare nouns in newspaper headlines (for Spanish: Sáez Rivera 2013; for Dutch: Oosterhof and Rawoens 2017; for German: Reich 2017; for English: Weir 2009).[9] While this may well be true for some

8 Note that Table 8 excludes both Chile and *médico*, as these served as the reference levels for the ordinal regression model presented in Section 4.2.2 (but see note 10).

9 But see Gerards and Kabatek (2018) for morphosyntactic phenomena that emerged in certain discourse traditions and text genres and were then generalized to more text types of a language.

Table 8: Noun-country combinations with *ir a*-BCS-variant >20 % in the Twitter data and no significantly higher experimental acceptability ratings.

Noun	Country	Twitter: BCS percentage	Comments: experiment
escuela	Argentina	50.4 % (*n* = 243/482) [Arg. global]	No significance: –R. Pl.
'school'	Colombia*	77.9 % (*n* = 7/9)	
	Spain*	70 % (*n* = 14/20)	
	Venezuela*	88.9 % (*n* = 8/9)	
colegio	Argentina	45.6 % (*n* = 342/750) [Arg. global]	No significance: +/–R. Pl.
'high school'	Colombia*	70 % (*n* = 14/20)	
	Spain*	75 % (*n* = 12/16)	
	Venezuela*	100 % (*n* = 5/5)	
cine 'cinema'	Venezuela*	22.2 % (*n* = 6/27)	
piscina 'pool'	Argentina*	40 % (*n* = 2/5) [Arg. global]	No significance: +/–R. Pl.
teatro	Argentina	55.8 % (*n* = 201/360) [Arg. global]	No significance: +/–R. Pl.
'theater'	Spain	27.9 % (*n* = 12/43)	
	Venezuela*	25 % (*n* = 5/20)	
peluquería 'hairdresser'	Argentina	28.1 % (*n* = 113/402) [Arg. global]	No significance: –R. Pl.
	Venezuela*	43.5 % (*n* = 10/23)	
iglesia	Argentina	24.3 % (*n* = 44/181) [Arg. global]	No significance: –R. Pl.
'church'	Colombia	29.3 % (*n* = 17/58)	
	Venezuela	50 % (*n* = 23/46)	

individual data points, we do not, however, believe that this explanation accounts for the entire picture. Specifically, it does not seem plausible to us that such an explanation provides a meaningful way to interpret the large differences in BCS-variant percentages even within Table 8. Similarly, such an approach could not meaningfully account for noun-country combinations whose Twitter BCS-variant percentages are well below 20 % (e.g., *cine* in Argentina; Twitter BCS-variant percentage: 2.4 % [*n* = 7/291]; cf. Table 2). Thus, rather than being due to Twitter as a medium and its character limitations, we believe it is more likely that the potential false Twitter positives could, at least in part, be due to other factors. Some of these are illustrated below through two noun-country combinations from Table 8, *ir a colegio* and *ir a iglesia* in Venezuela, before being placed into a broader perspective in line with the research question of this paper.

Section 4.2.1 demonstrated that in the pilot acceptability judgment experiment, the ratings for *ir a colegio* and *ir a iglesia* in Venezuela were strongly polarized toward both extreme ends of the rating scale. For *ir a colegio*, 8 out of 20 participants assigned an extremely high rating of '6' or '7', while another 8 out of 20 assigned an extremely low rating of '1' or '2'. For *ir a iglesia*, the experimental ratings were even more polarized: 12 out of 22 participants assigned a rating of '1' or '2'; 9 out of 22 assigned a rating of '6' or '7'. As a reviewer rightly points out, the substantial individual variability

observed across experimental participants – including, but not limited to, responses to *ir a colegio* and *ir a iglesia* in Venezuela –[10] may, in part, be attributable to the experimental design, in which the *ir a*-BCS items were embedded in short sentences that lacked pragmatic and discourse cotext and context. In other words, there is a possibility that the BCS variant may be perceived as more natural when accompanied by specific cotextual elements or contextual cues that were not present in the experimental stimuli. To the best of our knowledge, however, there is very little prior research on such potential factors. One relevant study is Pires de Oliveira and Rothstein (2013), who argue that Brazilian Portuguese bare singular count objects are facilitated by modifiers expressing habituality. A second is Wall (2022), who claims that Brazilian Portuguese pseudo-incoporated singulars headed by indefinite articles are characteristic of informal, diaphasically low-register contexts. Clearly, further research is needed on such factors, which could plausibly account for some of the observed variability in acceptability judgments: Some participants may have made additional co(n)textual assumptions – e.g., regarding habituality or register – and accordingly rated the BCS variant more favorably, while others did not and thus assigned lower ratings. If this were the case, then at least some of the potential false BCS-positives from Twitter included in Table 8 might instead be attributable to the experimental design, specifically its lack of cotextual elements and contextual cues.[11]

In addition to such potential design-related factors, there may also be other, as yet unidentified, variables contributing to the strong polarizations observed in the acceptability data. One such potential additional factor – already briefly mentioned in Section 4.2.2 in the context of an alternative ordinal regression model that was not further elaborated due to optimization issues and other challenges – could be 'age'.

10 Note that another, though somewhat less dramatic case of polarization, not included in Table 8 for the reasons explained in note 8, is *médico* in Chile (BCS-variant percentage: 93.8 %, n = 225/240; experimental mean and median: '3.05' and '3'). For this noun-country combination, 8 out of 19 participants assigned a rating of '1' or '2', while 3 out of 19 gave a rating of '6' or '7'. What is particularly interesting with regard to the experimental rating polarization of this noun-country combination is that one experimental participant from Chile, whose response was excluded from the data analysis due to the participant's having clearly identified the tested variable (cf. Section 4.1), assigned a maximum rating of '7' to *ir a médico*.

11 The same reviewer who highlights potential cotextual and contextual factors also suggests that the noun itself might have influenced the experimental ratings. While it is true that certain nouns are more conducive to pseudo-incorporation (as is the case of BCSs; cf. Section 2.1), the eight nouns examined in this study were carefully selected based on the criteria outlined in Section 3.1: (i) correspondence to SWDs in Spanish, (ii) documented cross-linguistic variation between SWDs and BCSs, i.e. inclusion of a stereotypical telic component in their lexical entry (following Pustejovsky 1995), and (iii) a promising number of BCS hits in exploratory Twitter searches. It is therefore not expected that the nouns themselves had a significant impact on the experimental ratings reported in Section 4.2.1, or on the BCS variant percentages observed in the Twitter data (Section 3.2.1).

This alternative model suggested that, with the exception of Colombia, younger speakers from the countries covered by the experiment tended to assign significantly higher ratings to the *ir a*-BCSs tested compared to older speakers. If corroborated by future research, this finding would challenge Kany's (1969) position that Latin American Spanish BCSs are diachronically decreasing (cf. Section 2.2). Given that the Twitter population is skewed toward younger users (cf. Section 1), while this may not necessarily be the case for the experimental participants, the factor of 'age' could, at least in part, also explain the large number of potential false Twitter positives included in Table 8.

Another potentially relevant factor could be that 'country' is too coarse a geographical reference category, and that small-scale diatopic varieties within the countries included in the experiment differ in their acceptance of the *ir a*-BCSs tested. Even though a first exploration of the polarized experimental ratings for Venezuelan *ir a iglesia* and *ir a colegio* based on the current place of residence of the Venezuelan experimental participants did not reveal any clear micro-diatopic factors behind the distribution of low versus high ratings,[12] it would not be surprising if future research on Venezuelan *ir a iglesia* and *ir a colegio* based on larger data sets would unveil such more fine-grained geographic patterns: It is well known that even varieties of the same language make differing use of BCSs and SWDs with the same noun (Sections 2.1 and 2.2). This fact is also reflected by the experimental data from Argentina, where Río de la Plata participants but not non-Río de la Plata ones showed a statistically significant likelihood of higher ratings with three *ir a*-BCSs. Again, such micro-diatopic differences could explain some of the potential false Twitter positives included in Table 8.

In sum, if the relevance of cotextual factors (e.g., habitual modifiers), demographic variables (e.g. speaker age), micro-diatopic variation, diaphasically low-register situations, or even a combination of these, were to be corroborated by future research, then high Twitter BCS-variant percentages – such as the 100 % and 50 % figures observed in the Venezuelan *ir a colegio* and *ir a iglesia* cases – without a statistically significant increase in experimental acceptability ratings could simply be the result of experimental item design and/or differing speaker representations across the two data sources particularly with regard to features positively associated with *ir a*-BCS acceptance.

---

12 Among the eight Venezuelan participants from Eastern coastal states, three assigned a rating of '7' to *ir a iglesia*, while four assigned a rating of '1' to the same *ir a*-BCSs. Similarly, among the five participants from Western coastal states, two rated *ir a iglesia* '7', while three assigned a rating of '1' or '2.' In the case of *ir a colegio*, three of the six participants from Western coastal states assigned a rating of '2', while two rated the same *ir a*-BCS '7'. Among the six participants from the Capital District, two assigned a rating of '1' or '2', while three assigned a rating of '6' or '7'.

Relatedly, an approach investigating additional cotextual, demographic and diasystematic factors in a more fine-grained manner could also provide meaningful explanations for two additional observations. First, the strong experimental polarization of some *ir a*-BCS noun-country combinations that *did* reach statistical significance in the experiment (*teatro* in Colombia and *escuela*, *iglesia*, and *peluquería* in Argentinian Río de la Plata Spanish; cf. Figures 4a, 4b, 4d and Table 7) could also reflect the influence of such additional factors. Second, demographic and diasystematic factors, in combination with their different prevalence in the participant groups, could explain why the statistically significant *ir a*-BCSs from Colombia in Table 7 displayed higher experimental ratings and higher Twitter BCS-variant percentages than the statistically significant ones from Río de la Plata-Argentina: The three Colombian *ir a*-BCSs could be more generalized diatopically, diastratically, and/or diaphasically than the three Argentinian Río de la Plata ones. This interpretation aligns well with the previous finding reported in Section 2.2 that some other, already documented Río de la Plata BCSs exhibit greater productivity in Uruguay than in Argentina and with the fact that two of the three statistically significant Argentinian Río de la Plata BCSs (*ir a iglesia* and *ir a peluquería*) were among those for which the Twitter study failed to document the maximum limit of $n = 500$ (Table 1).[13]

Unfortunately, the methodology used in the Twitter study outlined in Section 3.1 only allowed for control of the factor 'country of origin' and a full annotation of all 6,206 occurrences for the presence or absence of habitual modifiers exceeded the financial resources available for this study.[14] Similarly, the restricted sample size of the experimental data does not enable us to further pursue any of the possible microdiatopic or demographic lines of investigation sketched above. In other words, we can neither definitively clarify whether co(n)textual, demographic and/or diasystematic factors explain the differences between the Colombian and Argentinian Río de la Plata *ir a*-BCSs in Table 7, nor can we determine which of the *ir a*-BCSs in

13 Varying degrees of generalization in the use of *ir a*-BCSs are also noted by a reviewer, who suggests that certain specific combinations may have served as models or prototypes for the spread of the BCS pattern. The reviewer further proposes investigating the diatopically oriented data from a diachronic perspective in order to trace the emergence and development of the phenomenon – an idea that is, in principle, both compelling and intriguing, but must be left to future research. This is particularly the case because such an undertaking would first need to address the question of whether the occurrences of BCSs in the different countries examined in this paper are historically related or represent independent instances of polygenesis.

14 Note, however, that this might be a worthwhile endeavor, given that the BCS example in (30a) includes a habitual modifier – *qué rico* 'how nice [to…]' (an individual-level predicate) – whereas the minimal pair in (30b), which features the definite article, does not. Instead, (30b) contains the clearly episodic expression *ahora quiero* 'now I want to'.

Table 8 – if any – reflect a more fine-grained data pattern of some kind and which ones are false Twitter positives.

Summing up this latter point along with all other aspects of this discussion, our research question can be answered as follows:

1. Data from social media platforms can indeed be successfully used to detect and geographically map theoretically predicted but previously undocumented morphosyntactic variants, at least in mid- and large-sized countries where production reaches a numerically critical share of the total global Twitter output in a given language. In the present case, this applies to the Spanish BCSs *ir a cine* 'go to the cinema', *ir a piscina* 'go to the pool', and *ir a teatro* 'go to the theater' for Colombian Spanish, as well as *ir a iglesia* 'go to church', *ir a peluquería* 'go to the hairdresser', and *ir a escuela* 'go to school' for Argentinian Río de la Plata Spanish.
2. At the same time, data from social media platforms may pose the risk of yielding false positive conclusions regarding the acceptability of theoretically predicted but previously undocumented morphosyntactic variants. However, it is extremely difficult to distinguish false positives from social media platforms (potentially caused by undisclosed and skewed search algorithms and/or technical restrictions, such as character limits) from data that do not constitute false positives across the board but are instead too coarse to capture more fine-grained patterns of diatopic, diastratic, diaphasic, demographic, and/or pragmatic cotextually-induced variation.
3. When the primary interest exclusively lies in identifying a potentially more fine-grained geographic pattern of a theoretically predicted but previously undocumented morphosyntactic variant, a more detailed annotation of social media platform data may, to some extent, help address the issue outlined in point 2. Such an approach could, for instance, integrate a combination of a more detailed annotation of location field input, GPS-based geotagging, and other methods that infer a user's geographical origin based on the form, content, and/or meta-analysis of their posts and social network activity on the investigated platform.
4. Micro-diatopic and non-diatopic factors influencing the acceptability of theoretically predicted but previously undocumented morphosyntactic variants, however, may be significantly more challenging to control for within data sets extracted from social media platforms. Therefore, the investigation of such variants using social media platform data – particularly in the case of low-frequency phenomena – should be complemented by additional methods. One such method could involve carefully designed linguistic experiments that control for micro-diatopic, diastratic, diaphasic, demographic, and pragmatic cotextual factors.

With 1–4 in mind, the following section provides a brief conclusion of the paper.

6 Summarizing conclusion

The present paper addressed the research question of whether data from social media platforms, such as Twitter/X, can serve as a viable resource for identifying and geographically mapping theoretically predicted but previously undocumented morphosyntactic variants. To this end, Section 1 first provided a general overview of previous work on the investigation of diatopic variation based on linguistic data from Twitter. Section 2 then introduced the variable exemplarily chosen to investigate the research question, namely bare count singulars. Both from a cross-linguistic perspective and from that of Spanish, the language chosen for the empirical part of the paper, it motivated the choice of bare count singulars since, alongside short weak definites, these represent one of two morphosyntactic surface variants of the single underlying phenomenon of semantic pseudo-incorporation. Section 3 then presented a Twitter-based study on eight Spanish *ir a* + bare singular count noun combinations (lit. 'to go to + [noun]'; n = 6,206). Subsequently, part of the results of this study was validated in Section 4 by means of a pilot acceptability judgment experiment involving 226 Spanish speakers from six different countries and regions of the Spanish-speaking world. This validation revealed that data from social media platforms can indeed be successfully used to detect and geographically map theoretically predicted but previously undocumented morphosyntactic variants. However, it also demonstrated that relying solely on social media data is risky, as such data do not allow us to confidently exclude false positive conclusions regarding the availability of theoretically predicted but previously undocumented morphosyntactic variants. The reason for this is that we typically lack knowledge of the search algorithms used by social media platforms, and that social media data do not (easily) provide enough explicit metadata to detect more fine-grained diatopic variation beyond the country level. While this problem – as well as that of potential cotextual pragmatic factors (dis)favoring a given phenomenon – could possibly be mitigated through more sophisticated data annotation techniques, social media data may prove to be genuinely deficient when additional factors of variation (diastratic, diaphasic, and/or demographic) come into play. We therefore argued that the safest approach to conducting social media data-based investigations of theoretically predicted but previously undocumented morphosyntactic variants is to combine such data sets with others obtained through additional methods that allow for easier control of diatopic, diastratic, diaphasic, demographic, and cotextual pragmatic factors. One such method could be carefully designed linguistic experiments.

We hope that the insights provided in this paper, in the spirit of Nguyen (2021: 213), serve as a valuable starting point for further social media data-based "bottom-up discovery" of linguistic phenomena of all types, including, but ideally extending beyond morphosyntactic features. It would be particularly interesting, for example, to examine whether a methodology similar to the one employed in this study could also be successfully applied to linguistic domains that have traditionally been considered difficult to investigate using (written) social media data, such as phonetics and phonology (cf. Section 1). Similarly, it would be worthwhile to explore morphosyntactic phenomena that are less amenable to linear string-based search techniques of the kind used here. Furthermore, extending this line of research to include variables with non-binary variant outcomes – adapting statistical methods as necessary – would represent a promising avenue for future work. Finally, we hope that the paper serves as an incentive for a deeper exploration of varietally marked Spanish bare singular count nouns, an area of Spanish grammar that clearly requires more in-depth investigation.

Research ethics: The local Institutional Review Board deemed the study exempt from review.
Informed consent: Informed consent was obtained from all individuals included in this study, or their legal guardians or wards.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission. A1 is responsible for writing all sections of the paper. A2 is responsible for data collection and analysis of the study reported in Section 3. A3 supported data collection and analysis of the study reported in Section 4.
Use of Large Language Models, AI and Machine Learning Tools: ChatGPT was applied in specific instances for debugging R code, improve language of text previously written by Author 1, and enhance existing figures.
Conflict of interest: The authors state no conflict of interest.
Research funding: Part of this research was funded by University of Leipzig.
Data availability: The raw data can be obtained on request from the corresponding author.

Appendices

See Tables A.1 and A.2.

Table A.1: Twitter study – annotation and decision-making process for determining geographical data origin.

Case	Procedure and decision
(i) Location field: single specific country name and/or single globally unique place name (e.g.: *Buenos Aires*, (*Argentina*))	Validity of corresponding country assignment assumed
(ii) Location field: blank, punctuation mark, special character(s), reference to pop culture, fictional location, place outside Earth, etc. (e.g.: *around the corner*, *Mars*, ☺, *the World Trade Center*, *!!!*, *on stage*)	Twitter profile and, if applicable, other social medial profiles of user searched for cues of geographic user origin deemed valid according to Table 2. → If found: corresponding country assignment → If not: data point dismissed
(iii) Location field: abbreviations (country, place name, institution) (e.g.: *ARG*, *CABA* [Ciudad Autónoma de Buenos Aires], *UNRC* [Universidad Nacional de Río Cuarto])	Spanish *Wikipedia* searched for abbreviation → If found and unequivocal: corresponding country assignment → If not: data point dismissed
(iv) Location field: single country flag (e.g.: AR)	Validity of corresponding country assignment assumed
(v) Location field: non-Spanish-speaking country or place name according to cases (i) – (iv), (vi) (e.g.: *Tanzania*, *Genève*)	Cf. procedure and decision (ii) N.B.: USA considered Spanish-speaking
(vi) Location field: more than one country or place name according to cases (i) – (v), (vii) (e.g.: *Madrid/Buenos Aires*, AR/*Mexico*, *Rusia/Cuba*)	→ If procedure (ii) successful and unequivocal: corresponding country assignment N.B.: if one of the two non-Spanish speaking, only Spanish-speaking one considered N.B.: if *USA* + *Puerto Rico*, origin considered as *Puerto Rico*
(vii) Location field: place name existing in more than one Spanish-speaking country according to cases (i) – (iv), (vi) (e.g.: *Rosario* [Argentina] vs. *Rosario* [Mexico])	→ If procedure (ii) successful and unequivocal: corresponding country assignment → If not: Spanish *Wikipedia* searched and number of inhabitants for different homophone places determined → If largest number of inhabitants ≥ three times that of next largest number of inhabitants: corresponding country assignment; if not: data point dismissed
(viii): Location field no unequivocal result according to cases (i) – (vii)	→ If GPS geotagging provided and coordinates corresponding to Spanish-speaking country: corresponding country assignment → If not: data point dismissed

Table A.2: Experimental pilot study – filler items.

Fully ungrammatical	Medium acceptability	Fully grammatical
Lola perdió a un cuadro.	*Martina busca su padre.*	*Jaime habla bien español.*
Daniel toma a un vaso.	*Hugo oyó el vecino.*	*Pedro sabe muchas cosas.*
Lucas habló españoles.	*Alex andó a menudo.*	*Rafael habla mal francés.*
Martín dijo muchos cosas.	*Gonzalo vio el amigo.*	*Elena pudo con todo.*
David habló mala ruso.	*Antonio compró al maíz.*	*Valeria llegó muy tarde.*
Jimena pode cantar bien.	*Ana las ve, a las mesas.*	*Alejandra habló mucho.*
Lara llega mucho tarde.	*Carlota observa Marco.*	*Marc necesita a Ana.*
Sara saber hablar bien.	*Alba no se da de cuenta.*	*Carmen miró a su madre.*

Detailed model diagnostics (Section 4.2.2)

The cumulative link mixed model (CLMM) was fitted using the clmm function in R with adjusted optimization settings (maxIter = 2,000, gradTol = 1e-5, method = "nlminb"). A random intercept for 'speaker' was included to account for clustering in the data. The model successfully converged, achieving a maximum gradient of 0.0007, well below the convergence threshold. The condition number of the Hessian matrix (cond.H = 2.6e+04; cf. Table 6) was relatively high, suggesting potential numerical sensitivity. However, multicollinearity diagnostics using Generalized Variance Inflation Factor (GVIF), adjusted for the degrees of freedom (GVIF^(1/(2*Df))), confirmed no problematic collinearity among the predictors, with values below '5' (Country = 2.89, Noun = 2.35, Country:Noun interaction = 1.35). The threshold coefficients of the CLMM (−0.442, 0.715, 1.155, 1.501, 1.856, and 2.303; not included in Table 6 for the sake of conciseness) were progressively increasing, with no irregularities or overlaps, supporting the validity of the proportional odds assumption. A test for nominal effects, conducted using a cumulative logistic model without random effects (clm), showed no violations of the proportionality assumption, as all *p*-values were above 0.05, suggesting that the predictor effects remain constant across all thresholds. Residual diagnostics, including a residual plot against predicted probabilities and a histogram of residuals, indicated no systematic bias and confirmed that residuals were symmetrically distributed around zero, with no significant outliers. A likelihood ratio test was conducted to compare the model with interaction effects to a model without interaction effects. The results indicated that the model with interaction provided a significantly better fit (χ^2 (35) = 75.25, p < 0.001; model without interaction: AIC = 2757.1, log-likelihood = –1,359.6). Model fit was further evaluated using McFadden's R^2, which was calculated as 0.039. The full model achieved a log-likelihood of –1,321.93 (cf. Table 6), compared to –1,375.89 for the null model.

References

Abitbol, Jacob Levy, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot & Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in Twitter: A multivariate analysis. In Pierre-Antoine Champin, Fabein Gandon, Lionel Médini, Mounia Lalmas & Panagiotis G. Ipeirotis (eds.), *Proceedings of WWW 2018: The 2018 web conference, April 23–27, 2018 (WWW 2018), Lyon, France. ACM, New York, NY, USA*, 1125–1134. International World Wide Web Conferences Steering Committee Republic and Canton of Geneva Switzerland.

Aguilar-Guevara, Ana & Carolina Oggiani. 2023. Weak definite nominals. *Language and Linguistics Compass* 17(6). e12503.

Aguilar-Guevara, Ana & Joost Zwarts. 2010. Weak definites and reference to kinds. In Nan Li & David Lutz (eds.), *Proceedings from SALT 20*, 179–196. Ithaca NY: Cornell University.

Aguilar-Guevara, Ana & Joost Zwarts. 2013. Weak definites refer to kinds. *Recherches Linguistiques de Vincennes* 42. 33–60.

Álvarez Martínez, María. 1986. *El artículo como entidad funcional en el español de hoy*. Madrid: Gredos.

Ambrosio Aguilar, Agustin D., Everardo Bárcenas, Guillermo Molero Castillo & Rocío Aldeco Pérez. 2021. Geolocation of tweets in Spanish with transformer encoders. In Reyes Juárez-Ramírez, Carlos Fernández y Fernández, Samantha Jiménez, Alan Ramírez-Noriega, César Guerra-García, Raúl A. Aguilar Vera & Guillermo Licea Sandova (eds.), *Proceedings of the 9th international conference in software engineering research and innovation (CONISOFT)*, 227–231. Los Alamitos/California, Washington & Tokyo: IEEE Computer Society.

Anderson, Monica, Michelle Faverio & Jeffrey Gottfried. 2023. *Teens, social media and technology 2023*. Pew Reseacrch Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2023/12/PI_2023.12.11-Teens-Social-Media-Tech_FINAL.pdf (Accessed 12 January 2025).

Bland, Justin & Terrell A. Morgan. 2020. Geographic variation of *voseo* on Spanish Twitter. In Diego Pascual y Cabo & Idioia Elola (eds.), *Current theoretical and applied perspectives on Hispanic and Lusophone linguistics*, 7–38. Amsterdam: John Benjamins.

Blaxter, Tamsin & David Britain. 2021. Hands off the metadata!: Comparing the use of explicit and background metadata in crowdsourced dialectology. *Linguistics Vanguard* 7(s1). 20190029.

Borik, Olga & Berit Gehrke. 2015. An introduction to the syntax and semantics of pseudo-incorporation. In Olga Borik & Berit Gehrke (eds.), *The syntax and semantics of pseudo-incorporation*, 1–43. Leiden: Brill.

Borthen, Kaja. 2003. *Norwegian bare singulars*. Trondheim: Norwegian University of Science and Technology dissertation.

Bosque, Ignacio. 1996. Por qué determinados sustantivos no son sustantivos determinados. Repaso y balance. In Ignacio Bosque (ed.), *El sustantivo sin determinación. La ausencia de determinante en la lengua española*, 13–119. Madrid: Visor Libros.

Bosque, Ignacio. 2021. La gramática de construcciones. Una mirada externa. *Borealis: An International Journal of Hispanic Linguistics* 10(1). 1–41.

Brown, Earl K. 2016. On the utility of combining production data and perceptual data to investigate regional linguistic variation: The case of Spanish experiential gustar 'to like, to please' on Twitter and in an online survey. *Journal of Linguistic Geography* 3(2). 47–59.

Carlson, Greg. 2006. The meaningful bounds of incorporation. In Svetlana Vogeleer & Liliane Tasmowski (eds.), *Non-definiteness and plurality*, 35–60. Amsterdam: John Benjamins.

Carlson, Greg & Rachel Sussman. 2005. Seemingly indefinite definites. In Stephen Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 71–86. Berlin/Boston: de Gruyter.

Carlson, Greg, Rachel Sussman, Natalie Klein & Michael Tanenhaus. 2006. Weak definite noun phrases. In Christopher Davis, Amy Rose Deal & Youri Zabbal (eds.), *Proceedings of NELS 36*, 179–196. Amherst, MA: GLSA.

Casanova, Vanessa. 2020. El uso del complemento posesivo verbal por el complemento de régimen preposicional en español actual. *Moderna Språk* 114(3). 264–301.

Christophersen, Paul. 1939. *The articles. A study of their theory and use in English*. Copenhagen: Munksgaard.

Chung, Sandra & William A. Ladusaw. 2004. *Restriction and saturation*. Cambridge, Mass.: MIT Press.

Claes, Jeroen. 2017. La pluralización de haber presentacional en el español peninsular: datos de Twitter. *Sociolinguistic Studies* 11(1). 41–64.

Dayal, Veneeta. 2011. Hindi pseudo-incorporation. *Natural Language and Linguistic Theory* 29(1). 123–167.

De Benito Moreno, Carlota. 2022. Uso de los medios digitales de comunicación como corpus de español. In Giovanni Parodi, Pascual Cantos-Gómez & Chad Howe (eds.), *Lingüística de corpus en español/The Routledge handbook of Spanish corpus linguistics*, 481–493. London: Routledge.

De Benito Moreno, Carlota & Ana Estrada Arráez. 2018. Aproximación metodológica al estudio de la variación lingüística en las interacciones digitales. *Revista de Estudios del Discurso Digital (REDD)* 1. 74–122.

Dijkstra, Jelske, Wilbert Heeringa, Lysbeth Jongbloed-Faber & Hans Van de Velde. 2021. Using Twitter data for the study of language change in low-resource languages. A panel study of relative pronouns in Frisian. *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.644554.

Dobrovie-Sorin, Carmen, Tonia Bleam & M.-Teresa Espinal. 2006. Bare nouns, number and types of incorporation. In Liliane Tasmowski & Svetlana Vogeleer (eds.), *Non-definiteness and plurality*, 51–79. Amsterdam: John Benjamins.

Donoso, Gonzalo & David Sánchez. 2017. Dialectometric analysis of language variation in Twitter. In Preslav Nakov, Marcos Zampieri, Nikola Ljubešić, Jörg Tiedemann, Shevin Malmasi & Ahmed Ali (eds.), *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects*, 16–25. Valencia: Association for Computational Linguistics.

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In Shuly Wintner, Sharon Goldwater & Stefan Riezler (eds.), *Proceedings of the 14th conference of the European chapter of the Association for computational linguistics*, 98–106. Gothenburg: Association for Computational Linguistics.

Duggan, Maeve & Joanna Brenner. 2013. *The demographics of social media users, 2012*, vol. 14. Washington, DC: Pew Research Center's Internet & American Life Project.

Eisenstein, Jacob. 2017. Identifying regional dialects in on-line social media. In Charles Boberg, John Nerbonne & Dominic Watt (eds.), *The handbook of dialectology*, 363–383. Hoboken: Wiley Blackwell.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In Hang Li & Lluís Màrquez (eds.), *Proceedings of the 2010 conference on empirical methods in natural language processing (EMNLP)*, 1277–1287. Cambridge, MA: Association for Computational Linguistics.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(11). e113114.

Ennis, Juan A. 2015. Italian-Spanish contact in early 20th century Argentina. *Journal of Language Contact* 8(1). 112–145.

Espinal, Ma Teresa & Sónia Cyrino. 2017a. On weak definites and their contribution to event kinds. In Olga Fernández Soriano, Elena Castroviejo Miró & Isabel Pérez Jiménez (eds.), *Boundaries, phases and interfaces: Case studies in honor of Violeta Demonte*, 130–150. Amsterdam: John Benjamins.

Espinal, Ma Teresa & Sónia Cyrino. 2017b. The definite article in Romance expletives and long weak definites. *Glossa* 2(1). 23. 1–26.

Espinal, Ma Teresa & Louise McNally. 2011. Bare nominals and incorporating verbs in Spanish and Catalan. *Journal of Linguistics* 47(1). 87–128.

Estrada Arráez, Ana & Carlota de Benito Moreno. 2016. Variación en las redes sociales: datos twilectales. *Revista Internacional de Lingüística Iberoamericana* 14(2(28)). 77–111.

Farkas, Donka F. & Henriëtte de Swart. 2003. *The semantics of incorporation: From argument structure to discourse transparency*. Stanford: CSLI Publications.

Fontanella de Weinberg, María B. 1987. *El Español Bonaerense. Cuatro Siglos de Evolución Lingüística (1580–1980)*. Buenos Aires: Hachette.

Gerards, David P. 2022. Clitics in informal written sources of Angolan Portuguese and their similarity to informal Brazilian Portuguese. In Anja Hennemann & Benjamin Meisnitzer (eds.), *Linguistic hybridity. Contact-induced and cognitively motivated grammaticalization and lexicalization processes in Romance languages*, 15–46. Heidelberg: Winter.

Gerards, David P. & Johannes Kabatek. 2018. Grammaticalization and discourse traditions: The case of Portuguese *caso*. In Oscar Loureda Lamas & Salvador Pons Bordería (eds.), *Beyond grammaticalization and discourse markers: New issues in the study of language change*, 115–159. Leiden/Boston: Brill.

Gerards, David P. & Elisabeth Stark. 2022. Non-maximal definites in Romance. In Marco Bril, Martine Coene, Tabea Ihsane, Petra Sleeman & Thom Westveer (eds.), *RLLT19, special issue of isogloss. Open Journal of Romance Linguistics*, vol. 8(5)/5, 1–32.

Gonçalves, Bruno & David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PLoS ONE* 9(11). e112074.

Gonçalves, Bruno & David Sánchez. 2016. Learning about Spanish dialects through Twitter. *Revista Internacional de Lingüística Iberoamericana* 14(2). 65–75.

Graham, Mark, Scott A. Hale & Devin Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer* 66(4). 568–578.

Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami & Dinsheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2(11). https://doi.org/10.3389/frai.2019.00011.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46(4). 293–319.

Haddican, Bill & Daniel E. Johnson. 2012. Effects on the particle verb alter-nation across English dialects. *University of Pennsylvania Working Papers in Linguistics* 18(2). 31–40.

Hawkins, John A. 1978. *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. London: Croom Helm.

Hecht, Brent, Lichan Hong, Bongwon Suh & Ed. H. Chi. 2011. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In Desney Tan, Geraldine Fitzpatrick, Carl Gutwin, Bo Begole & Wendy A Kellogg (eds.), *Proceedings of the SIGCHI conference on human factors in computing systems*, 237–246. New York: Association for Computing Machinery.

Hecht, Brent & Monica Stephens. 2014. A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 197–205. Washington: Association for the Advancement of Artificial Intelligence.

Heim, Irene. 1982. *The semantics of definite and indefinite noun phrases*. Amherst: University of Massachusetts Amherst dissertation.

Hoff, Mark. 2020. *Cerca mío/a* or *cerca de mí* ? A variationist analysis of Spanish locative + possessive on Twitter. *Studies in Hispanic and Lusophone Linguistics* 13(1). 51–78.

Hong, Lichan, Gregorio Convertino & Ed Chi. 2011. Language matters in Twitter: A large scale study. In Nicolas Nicolov, James G. Shanahan, Lada Adamic, Ricardo Baeza-Yates & Scott Counts (eds.), *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5(1), 518–521. Washington: Association for the Advancement of Artificial Intelligence.

Hu, Xuke, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan & Friederike Klan. 2023. Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys* 56(5). 1–37.

Huang, Yuan, Diansheng Guo, Alice Kasakoff & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59. 244–255.

Ihsane, Tabea, David Paul Gerards & Elisabeth Stark. 2025. Indefinite determiners: Why DE can be enough. Insights from Francoprovençal. *Journal of Linguistics*. https://doi.org/10.1017/S0022226725100832.

Jones, Taylor. 2015. Toward a description of African American vernacular English dialect regions using "Black Twitter". *American Speech* 90(4). 403–440.

Jørgensen, Anna Katrine, Dirk Hovy & Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In Wei Xu, Bo Han & Alan Ritter (eds.), *Proceedings of the workshop on noisy user-generated text*, 9–18. Beijing: Association for Computational Linguistics.

Julie, Thiombiano, Malo Sadouanouan & Traore Yaya. 2023. A geolocation approach for tweets not explicitly georeferenced based on machine learning. In S. Ossowski, P. Sitek, C. Analide, G. Marreiros, P. Chamoso & S. Rodríguez (eds.), *Distributed computing and artificial intelligence, 20th international conference*, 223–231. Cham: Springer.

Kallulli, Dalina. 1999. *The comparative syntax of Albanian: On the contribution of syntactic types to propositional interpretation*. Durham: University of Durham dissertation.

Kany, Charles. 1969 [1945]. *American-Spanish Syntax*. Madrid: Gredos.

Kellert, Olga. 2024. Geographic variation of *voseo* and *tuteo* on X (Twitter) with a consideration of mixing cases (*vos puedes*). *Revue Romane* (online first). https://doi.org/10.1075/rro.23006.kel.

Kuguel, Inés & Carolina Oggiani. 2016. La Interpretación de sintagmas preposicionales escuetos introducidos por la preposición *en*. *Cuadernos de Lingüística de El Colegio de México* 3(2). 5–34.

Kulkarni, Vivek, Bryan Perozzi & Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of internet language. In Markus Strohmaier & Krishna P. Gummadi (eds.), *Proceedings of the tenth international AAAI conference on web and social media (ICWSM 2016)*, 615–618. Washington: Association for the Advancement of Artificial Intelligence.

Laca, Brenda. 1999. Presencia y ausencia de determinante. In Ignacio Bosque & Violeta Demonte (eds.), *Gramática descriptiva de la lengua española*, vol. 1, *1: Sintaxis básica de las clases de palabras*, 891–928. Madrid: Espasa Calpe.

Lamsal, Rabindra, Aaron Harwood & Maria Rodriguez Read. 2022. Where did you tweet from? Inferring the origin locations of tweets based on contextual information. In Shusaku Tsumoto, Yukio Ohsawa, Lei Chen, Dirk Van den Poel, Xiaohua Hu, Yoichi Motomura, Takuya Takagi, Lingfei Wu, Ying Xie, Akihiro Abe & Vijay Raghavan (eds.), *Proceedings of 2022 IEEE International Conference on Big Data (Big Data)*, 3935–3944. Piscataway: IEEE.

Leonetti, Manuel. 2019. On weak readings of definite DPs. In Natascha Pomino (ed.), *Proceedings of the IX NEREUS International Workshop "Morphosyntactic and semantic aspects of the DP in Romance and beyond"*, 1–25. Fachbereich Linguistik: University of Constance.

Levshina, Natalia. 2015. *How to do linguistics with R. Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Lipski, John. 2008. *Afro-Bolivian Spanish*. Madrid & Frankfurt: Iberoamericana-Vervuert.

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Mahajan, Rhea & Vibhakar Mansotra. 2021. Predicting geolocation of tweets: Using combination of CNN and BiLSTM. *Data Science and Engineering* 6. 402–410.

Marttinen Larsson, Matti & Miriam Bouzouita. 2018. *Encima de mí* vs. *encima mío*: un análisis variacionista de las construcciones adverbiales locativas con complementos preposicionales y posesivos en Twitter. *Moderna sprak* 112(1). 1–39.

Massam, Diane. 2009. Noun incorporation: Essentials and extensions. *Language and Linguistic Compass* 3/4. 1076–1096.

Melo, Fernando & Bruno Martins. 2016. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS* 21(1). 3–38.

Mithun, Marianne. 2010. Constraints on compounding and incorporation. In Irene Vogel & Sergio Scaliese (eds.), *Compounding*, 37–56. Amsterdam: John Benjamins.

Munn, Alan & Cristina Schmitt. 2002. Bare nouns and the morphosyntax of number. In Teresa Satterfield, Christina Tortora & Diana Cresti (eds.), *Current issues in Romance languages*, 225–239. Amsterdam: John Benjamins.

Nguyen, Dong. 2021. Dialect variation on social media. In Marcos Zampieri & Preslav Nakov (eds.), *Similar languages, varieties, and dialects: A computational perspective*, 204–218. Cambrige: Cambridge University Press.

Nguyen, Dong, Dolf Trieschnigg & Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. In Daniele Quercia, Bernie Hogan, Meeyoung Cha, Cecilia Mascolo & Christian Sandvig (eds.), *Proceedings of the on web and social media*, vol. 9(1), 666–69. Washington: Association for the Advancement of Artificial Intelligence.

Oggiani, Carolina. 2021a. Una aproximación a las expresiones nominales definidas débiles en el español del Río de la Plata. In Cecilia Bértola, Carolina Oggiani & Ana Clara Polakoff (eds.), *Estudios de lengua y gramática*, 87–95. Montevideo: Universidad de la República.

Oggiani, Carolina. 2021b. "Escribir artículo": nombres singulares escuetos en posición de objeto en español rioplatense. *Borealis – An International Journal of Hispanic Linguistics* 10(2). 313–333.

Oggiani, Carolina. 2022. Los escuetos definidos débiles en español rioplatense. *Revista de Estudos da Linguagem* 30(1). 239–268.

Oosterhof, Albert & Gudrun Rawoens. 2017. Register variation and distributional patterns in article omission in Dutch headlines. *Linguistic Variation* 17(2). 205–228.

Pato, Enrique & Carlora de Benito Moreno. 2017. Tráenolos para comérnolos o la 'transposición' del clítico en español actual. *Philologica Jassyensia* 13(1). 121–136.

Pavalanathan, Umashanthi & Jacob Eisenstein. 2015. Confounds and consequences in geotagged Twitter data. In Lluís Màrquez, Chris Callison-Burch & Jian Su (eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2138–2148. Lisbon: Association for Computational Linguistics.

Perea, María Pilar & Antonio Ruiz Tinoco. 2016. Análisis del uso y distribución de formas léxicas dialectales del catalán en Twitter. *Revista Internacional de Lingüística Iberoamericana* 14(28). 49–64.

Pires De Oliveira, Roberta & Susan Rothstein. 2013. Bare singular arguments in Brazilian Portuguese: Perfectivity, telicity, and kinds. In Johannes Kabatek & Albert Wall (eds.), *New perspectives on bare noun phrases in Romance and beyond*, 189–222. Amsterdam: John Benjamins.

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, MA: MIT Press.

RAE/ASALE 2009 = Real Academia Española/Asociación de Academias de la Lenga Española. 2009. *Nueva gramática de la lengua española*. Madrid: Espasa.

Rahimi, Afshin, Trevor Cohn & Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. In Regina Barzilay & Min-Yen Kan (eds.), *Proceedings of the 55th annual meeting of the*

Association for computational linguistics, vol. 2, 209–216. Vancouver: Association for Computational Linguistics.

Reich, Ingo. 2017. On the omission of articles and copulae in German newspaper headlines. *Linguistic Variation* 17(2). 186–204.

Russ, Robert B. 2013. *Examining regional variation through online geotagged corpora*. Columbus: Ohio State University MA thesis.

Russell, Bertrand. 1905. On denoting. *Mind* 14. 479–493.

Sáez Rivera, Daniel M. S. 2013. Bare nominals in American-Spanish headlines. In Johannes Kabatek & Albert Wall (eds.), *New perspectives on bare noun phrases in Romance and beyond*, 157–188. Amsterdam: John Benjamins.

Scheffler, Tatjana, Johannes Gontrum, Matthias Wegel & Steve Wendler. 2014. Mapping German tweets to geographic regions. In Josef Ruppenhoffer & Gertrud Faaß (eds.), *Proceedings of the NLP4CMC workshop at KONVENS*, 26–33. Hildesheim: KONVENS.

Schütze, Carson & Jon Sprouse. 2014. Judgment data. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 27–50. Cambridge: Cambridge University Press.

Severo, Ohanna Teixeira Barchi. 2019. An experimental study on the interpretation of bare singulars in Mexican Spanish. *Revista de Estudos da Linguagem* 27(2). 575–601.

Squires, Lauren. 2015. Twitter. Design, discourse, and the implications of public text. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), *The Routledge handbook of language and digital communication*, 239–256. New York: Routledge.

Strelluf, Christopher. 2022. Regional variation and syntactic derivation of low-frequency need-passives on Twitter. *Journal of English Linguistics* 50(1). 39–71.

Stvan, Laurel Smith. 2009. Semantic incorporation as an account for some bare singular count noun zses in English. *Lingua* 119. 314–333.

Tinoco, Antonio Ruiz. 2013. Twitter como corpus para estudios de geolingüística del español. *Sophia Linguistica: Working Papers in Linguistics* 60. 147–163.

van Geenhoven, Veerle. 1998. *Semantic incorporation and indefinite descriptions: Semantic and syntactic aspects of noun incorporation in West Greenlandic*. Stanford: CSLI.

van Halteren, Hans, Roeland Van Hout & Romy Roumans. 2018. Tweet geography: Tweet based mapping of dialect features in Dutch Limburg. *Computational Linguistics in the Netherlands Journal* 8. 138–162.

Wall, Albert. 2017. *Bare nominals in Brazilian Portuguese: An integral approach*. Amsterdam: John Benjamins.

Wall, Albert. 2022. Number-neutral indefinite objects in Brazilian Portuguese as a case of semantic incorporation. *Journal of Portuguese Linguistics* 21. 1–29.

Weir, Andrew. 2009. *Article drop in English headlinese*. London: University College London MA thesis.

Willis, David. 2020. Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. *Glossa* 5(1). 103.

Zheng, Xin, Jialong Han & Aixin Sun. 2018. A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering* 30(9). 1652–1671.

Zola, Paola, Constantino Ragno & Paulo Cortez. 2020. A Google Trends spatial clustering approach for a worldwide Twitter user geolocation. *Information Processing & Management* 57(6). 102312.

Zwarts, Joost. 2014. Functional frames in the interpretation of weak definites. In Ana Aguilar-Guevara, Bert Le Bruyn & Joost Zwarts (eds.), *Advances in weak referentiality*, 265–285. Amsterdam: John Benjamins.