Journal of Linguistic Geography (2016) **3**, 47–59. © Cambridge University Press 2016
doi:10.1017/jlg.2016.1

ORIGINAL RESEARCH

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

Earl K. Brown*

Kansas State University

The use of both production and perceptual data has the potential to provide a more complete picture of linguistic phenomena than would otherwise be the case, including when exploring regional linguistic variation. Utilizing the social networking platform Twitter and an online survey, this paper reports on a descriptive analysis of the geographic distribution of a less-commonly used syntactic form of the Spanish verb *gustar* 'to like, to please', referred to as experiential *gustar* (e.g., *cuando gustes* 'when you'd like'). The results from the analysis of 6,686 tweets together with the responses of 81 native Spanish-speaking participants in an online survey suggest that experiential *gustar* is produced and is perceived to be produced most often in Mexican Spanish, despite not being exclusive to that country. The paper contributes to the literature depicting the benefit of using both production and perceptual data in the study of dialectal variation, as well as to the literature documenting language variation in Spanish.

1. Introduction

One source of data for language researchers that is gaining in acceptance is the hugely popular social networking platform Twitter. In just the last few years a plethora of studies that take advantage of this medium have appeared. For example, Alvarez and Muñoz Muñoz (2012) perform a contrastive analysis of organizational and rhetorical devices in English and Spanish using data obtained from Twitter. Bamman et al. (2014) analyze lexical variation and gender identity as seen on this social network. Merkhofer (2013) studies the use of singular subject pronouns in English, including singular *they*, in a small corpus comprised of tweets, or messages sent on Twitter. Sang and Tjong (2011) go so far as to offer a guide of sorts for performing linguistic research using Twitter. With specific regard to regional variation, Ruiz Tinoco (2012) analyzes the subjunctive mood in Spanish in 22 different cities across the Spanish-speaking world, as seen on Twitter. In summary, an increasing number of language researchers, including dialectologists, now realize the potential Twitter holds for studying language use and variation.

While production data have been fruitful in the description and analysis of language use and variation, perceptual data have also been important in dialectology and have the potential to help paint a more complete picture of a given linguistic variable than would otherwise be the case. Commenting on the various methodologies utilized to study dialectal variation, Labov (1975:96, emphasis added) proposes that "most dialect studies use formal elicitation technique... others include the direct observation of speech.... But we can *also* study geographic and social dialects by eliciting judgments of acceptability and semantic interpretation." Haddican and Johnson (2012) study variants of particle verb alternation in Scotland and Southern England by using both production and perceptual data. The production data are pulled from Twitter, while the perceptual data come from an acceptability judgment experiment.

One kind of perceptual data that has been shown to be productive is that of folk knowledge of language. Preston and colleagues (e.g., Long & Preston, 2002; Niedzielski & Preston, 2000; Preston, 1989, 1993) have thoroughly demonstrated the usefulness of relying on what non-experts perceive about language use and regional variation. Preston (2011) presents an outline of the various methods that have been employed to study language through the knowledge of the folk. One technique that he discusses is that of studying regional variation through areal identification. Preston (1996, cited in Preston, 2011) describes a task in which recordings of speech produced by speakers in nine locations ranging from Central Michigan to Southern Alabama were played to laypeople. The listeners were given the task of identifying where the speakers might be from. The high level of accuracy of the informants is illustrative of the nuanced perceptual abilities native speakers of a language can often have regarding regional pronunciations.

In summary, both production data and perceptual data have been employed in the study of language use

*Address for correspondence: Earl K. Brown, 104 Eisenhower Hall, Manhattan, Kansas 66503, +1-785-532-6760, Email: ekbrown@ksu.edu

and regional variation. These two types of data hold the potential to present different but complementary perspectives on a given linguistic phenomenon.

1.1 Spanish gustar

One linguistic variable that could benefit from a descriptive study that documents its geographic distribution is the varied syntactic uses of the Spanish verb *gustar* 'to like, to please', as this variable usage is virtually absent from the literature detailing Spanish dialectology. Most commonly, the syntactic subject of *gustar* is the cause of the emotion while the person who experiences that emotion is encoded as a dative argument, as seen in example (1).[1]

(1)

"No	me	**gustan**	las	jevas	que	se	maquillan	mucho."
NEG	1S OBJ PRO	3P 'PLEASE'	FEM DEF ART	'CHICKS'	REL PRO	REFL PRO	'APPLY MAKEUP'	'A LOT'

'I don't like chicks who wear a lot of makeup.'[2] (Male, Caracas, Venezuela 2011-11-10)

In this paper, this canonical and most frequently-used form will be referred to as "causal *gustar*", owing to the fact that the cause of the emotion is the syntactic subject of the verb.

While less-commonly used, there is another syntactic usage of *gustar* in which the experiencer is the subject of the verb, as seen in example (2).

(2)

"Cuando	**gustes**,	me	envías	un	DM	y
'WHEN'	2S 'LIKE'	1S OBJ PRO	2S 'SEND'	MAS INDEF ART	'DM'	'AND'
te	doy	los	datos	del	chamán."	
2S OBJ PRO	1S 'GIVE'	MAS DEF ART	'INFORMATION'	PREP + DEF ART	'SHAMAN'	

'When you'd like, send me a DM [Direct Message] and I'll give you the information about the shaman' (Male, Lima, Peru 2011-11-19)

In this paper, this second, less common use of *gustar* will be labeled "experiential *gustar*", owing to the fact that the experiencer of the emotion is the syntactic subject of the verb.

The literature that identifies the geographic distribution of experiential *gustar* in the Spanish-speaking world is very limited in scope. Not only is it virtually absent from this literature, on the occasions that it is mentioned, the geographic areas in which it appears are not detailed. Further, no mention is made of the rates at which experiential *gustar* occurs in comparison to causal *gustar*. In fact, experiential *gustar* is absent from often-cited literature on Hispanic dialectology, such as Alonso (1967), García de Diego (1978), Sala et al. (1982), Zamora Vicente (1985), Cotton and Sharp (1988), Zamora Munné and Guitart (1988), Lipski (1994), and Alvar (1996).

An exception to this trend is seen in Kany (1951) and Boyd-Bowman (1960), who make reference to it, albeit in a fleeting manner. In his volume *American-Spanish Syntax*, Kany (1951:352) notes the omission of the preposition de 'of' in some phrasal prepositions and verbs, such as the change of *antes de que* 'before' to *antes que* and the change of *alegrarse de que* 'to be happy that' to *alegrarse que.* In this list of words, Kany includes *gustar de,* which represents an example of experiential *gustar*. No additional information is given specifically about experiential *gustar*. However, the fact that experiential *gustar* is included in a volume dedicated to Latin American Spanish syntax gives the impression that experiential *gustar* is used across Spanish-speaking Latin America, rather than being confined to a specific country or region, as may be the more accurate reality.

Conversely, in reference to a specific region of Spanish-speaking Latin America, that of Guanajuato, Mexico, Boyd-Bowman (1960:240) notes that, as a general rule, *de* is omitted after experiential *gustar*, citing Kany (1951), and then gives some examples: "*no gusta tomar una copita*" 'wouldn't you like to have a drink' and "*podemos hacer lo que usted guste*" 'we can do whatever you'd like'. This reference to experiential *gustar* is unique in that it ties experiential *gustar* to the Spanish of the specific region of Guanajuato, Mexico and possibly by extension, to Mexican Spanish more generally.

A few studies that describe contact varieties of Spanish and Spanish as a heritage language present experiential *gustar*, but always as an adstratum influence from the other language. For example, Silva-Corvalán (1994) documents the language of second and third generation Mexican migrant speakers of Spanish in Los Angeles and shows that these speakers produce *gustar* in a way that mirrors the syntactic behavior of English *to like*. With perceptual data, Pascual Cabo (2013) shows that heritage speakers of Cuban Spanish in Miami more readily accept features that approximate the syntactic behavior of *to like.* Finally, along the Brazilian-Uruguayan border, Klee and Lynch (2009) present further examples of *gustar* that diverge from the canonical, causal *gustar*. They attribute this usage to the adstratum influence of Brazilian Portuguese. In summary, while these few studies present what on a purely syntactic level are examples of experiential *gustar* in a specific area of the Spanish-speaking world, they do so to display the influence that the other language has on those varieties of Spanish and do not attempt to document where in the monolingual Spanish-speaking world experiential *gustar* may be used most commonly and at what rates in comparison to causal *gustar*.

Why is it important to study experiential *gustar*? The silence in the Spanish dialectal research on this verb form begs the question: Where, if anywhere, is this syntactic

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

form most common? Consequently, while exploratory and descriptive in nature, this paper helps to remedy the near absence of experiential *gustar* in the literature documenting Spanish dialectology. Further, this paper exemplifies the usefulness of studying language variation, including regional variation, with both production and perceptual data. Finally, this study also illustrates the innovative use of the social networking site Twitter for linguistic research and shows that this social networking platform is suitable for studying morphosyntactic regional variation across large geographic areas, including in the Spanish-speaking world and elsewhere.

The research questions motivating this paper are:

R1: What do production data from Twitter and perceptual data from two tasks show about the geographic distribution of experiential *gustar* in the Spanish-speaking world?

R2: Do the production and perceptual data concur with each other?

R3: How often does experiential *gustar* occur in comparison with causal *gustar*, the canonical form, in the places experiential *gustar* occurs?

2. Data and Methods

In order to use both production and perception data to document where in the Spanish-speaking world experiential *gustar* is used most often, if anywhere in particular, and to what extent it is used in comparison to causal *gustar*, two methods were employed. First, a large-scale search of Twitter was performed to extract the present tense indicative and subjunctive forms of this verb. Second, an online survey of native Spanish-speakers was administered.

2.1 Twitter as corpus

Twitter is a social networking platform that allows its users to write micro-blog entries of up to 140 characters in each message, referred to as a "tweet". When they open an account, users designate their account as either public or private. Messages written with public accounts can be seen by anyone, while messages written from private accounts can only be seen by their followers. To become a follower of a private account, the account owner must grant permission.

Utilizing Twitter to study linguistic phenomena, such as experiential *gustar*, that is not used often in running speech or writing is ideal, as hundreds of millions of tweets are written everyday. Twitter began in 2006 and five years later, by 2011, tweets were being written at the rate of 200 million per day.[3] Within another two years, by 2013, 500 million tweets were sent each day throughout the world. This trend shows no signs of stopping anytime soon (as of early 2015). One result of this large number of tweets being written each day is an increase in the likelihood that at least some of those tweets will contain infrequent forms. Additionally, Twitter is useful for descriptive studies, such as this one, that seek to measure where in the world infrequent forms may be used most. It would be virtually impossible to collect a corpus large enough from across the Spanish-speaking world, to study experiential *gustar* using traditional corpus-building techniques such as the sociolinguistic interview that must be subsequently transcribed by a researcher. In a sense, Twitter users with public accounts are offering themselves as research subjects to language researchers by making their language available in their tweets.

Despite its usefulness as a means of gathering huge amounts of language data from across the world in very little time, Twitter has some limitations that should be pointed out. Twitter gathers little information about its users; it only requires a user's name and location. As such, researchers who might like to investigate the role of sociolinguistic factors, such as the gender of users, socioeconomic status, educational status, language background, and other factors, will be hard-pressed to do so with data from Twitter. While it is true that researchers can make reasonable guesses about the gender of users based on their first name, this will not yield the most accurate results, particularly with names that tend to be more androgynous. Despite this limitation, some studies use innovative methodologies to analyze the use of the language in the tweets themselves to discuss the construction of gender identity on Twitter (cf. Bamman et al., 2014; Eisenstein et al., 2010).

With regards to the process of extracting tweets, while Twitter's website (twitter.com) offers a web interface to perform advanced searches, such as those that limit the search to a specific geographic location on the globe determined by latitude and longitude coordinates, performing a large number of searches in this manner would be onerous as well as error-prone for the large number of tweets needed to study infrequent forms. Manually searching each form in each city and then copying and pasting the data from a web browser, as well as likely having to page through several pages of results for some searches, would be difficult and time-consuming, at best.

Fortunately, Twitter offers an API (Application Programming Interface) that allows programmers to access tweets in an automated manner. Most computer scripting languages offer their users the ability to access this API and to retrieve tweets that meet certain criteria. These languages include Perl, Python, PHP, jQuery, Ruby, and R. For this paper, a script, or a sequence of computer commands, was written in the R programming language (Gentry, 2014; R Development Core

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

Team, 2014) to perform the searches. The script executed a separate search for each of the eleven different orthographic representations of the present indicative and present subjunctive forms of gustar in each of the twenty capital cities in Spanish-speaking Latin America and in Spain. The eleven forms were comprised of six present indicative forms and five subjunctive forms. The indicative forms are: *gusto* 1s, *gustas* 2s, *gusta* 3s, *gustamos* 1p, *gustáis* 2p, *gustan* 2/3p. The subjunctive forms are: *guste* 1/3s, *gustes* 2s, *gustemos* 1p, *gustéis* 2p, *gusten* 2/3p. As the objective of this paper is to analyze the contributions of production and perceptual data to the description of regional variation in language use, as well as to make an effort to document where experiential *gustar* is most common, the verbal tense was purposefully limited to present tense. Future studies should explore the possibility of a conditioning effect from verbal tense on the relative rates of occurrence of causal *gustar* and experiential *gustar* between dialects of Spanish.

The script pulled tweets written on either a computer or a hand-held device, such as a tablet or smart phone, that had Spanish set as the default language, as specified by the ISO (International Standards Organization) language code "es". The tweets were written on one of six days in November and December 2011 (November 10, 12, 17, 19 and December 1, 3) and were sent from within a 100-kilometer radius of the capital cities, which are: Asuncion, Paraguay; Bogota, Colombia; Buenos Aires, Argentina; Caracas, Venezuela; Guatemala City, Guatemala; Havana, Cuba; La Paz, Bolivia; Lima, Peru; Madrid, Spain; Managua, Nicaragua; Mexico City, Mexico; Montevideo, Uruguay; Panama City, Panama; Quito, Ecuador; San Jose, Costa Rica; San Juan, Puerto Rico; San Salvador, El Salvador; Santiago, Chile; Santo Domingo, Dominican Republic; and Tegucigalpa, Honduras.[4] Prior to issuing the searches, the latitude and longitude coordinates of a central point in each city were found using a simple online map,[5] and this information, along with the desired radius of 100 kilometers, was passed to the "searchTwitter" function of the R add-on package twitteR (Gentry, 2014). To illustrate, in order to retrieve the tweets sent in and near Mexico City, the coordinates of the central Zocalo plaza were retrieved and the geocode argument "19.432590, -99.133029, 100km" was passed to the "searchTwitter" function, which then issued the request to Twitter's API and retrieved the results. This type of search was performed eleven times in Mexico City, one time each for the eleven orthographic forms of gustar. As this procedure was repeated for the twenty cities, 220 unique search requests were executed by the R script. The decision to search for *gustar* in the capital city of each country was based on the fact that in most Spanish-speaking countries, if not all, the capital city and its surrounding metropolitan area is the most populous area of the country and thus promised to hold the largest number of tweets sent from public Twitter accounts, and therefore held the largest likelihood of containing examples of experiential *gustar*.[6]

While it is not presumed that all Twitter users in a given capital city grew up in the corresponding country, it is assumed that the influence of tweets sent by people from other countries, whether passersby or immigrants, will be mitigated by the large numbers of tokens analyzed. In other words, it is assumed that the majority of Twitter users in a given capital city are natives of the corresponding country and that the overall influence of Twitter users from other countries will be negligible. However, in an effort to minimize the possible influence of other dialects of Spanish, only cities from which at least 100 tweets were sent are analyzed in this study. Unfortunately for our purposes, as explained above, Twitter does not collect information about its users that would enable the exclusion of users who are not native to the country in which they send tweets.

Certain types of tweets were excluded from the analysis. "Retweets", or tweets that Twitter users receive and then forward to their followers, were excluded to avoid inflating the influence of any given token of *gustar*. Additionally, emulating the methodology of Eisenstein et al. (2010) and Kwak et al. (2010), tweets sent by Twitter users with 1,000 or more followers or 1,000 or more "friends" (people followed by the account holder) were excluded, as these Twitter users are usually marketers, celebrities with professional publicists, and news media sources. Finally, tweets that contained a URL were excluded, as they are often advertisements or updates on weather conditions posted by computer scripts, rather than humans.

Using this procedure, 9,279 tweets were extracted. Through a visual inspection of each tweet, 2,593 tweets were excluded for a variety of reasons. For example, many of the tweets with gusto were excluded because they represented the nominal usage rather than verbal usage of that word, that is 'pleasure' in English, (N = 1,466), or because they represented a preterit usage of the word (N = 735) that lacked a written accent, a common feature in Spanish internet orthography (cf. Myslín & Gries 2010). Other reasons for the exclusion of tweets were additional retweets (N = 114) that slipped through the initial filtering process performed with a computer script, the quoting of song lyrics (N = 53), references made to Facebook's (face book.com) "like" feature, which in Spanish is "*me gusta*" (N = 50), and advertisements sent from Twitter accounts with fewer than 1,000 followers or "friends" (N = 15). Finally, as mentioned above, only cities from

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

which at least 100 tweets were sent are analyzed in the results so as to mitigate the possibility of outsiders inflating the rate of usage of experiential *gustar* in that city. This excluded the combined 139 tweets from the five cities with fewer than 100 tweets each: Havana, Panama City, La Paz, Managua, and Santo Domingo. The exclusion of these 2,593 tweets left a total 6,686 tweets sent by 5,707 Twitter users for the analyses reported below.

2.2 Online survey

In addition to the production data retrieved from Twitter, perceptual data were gathered from an online survey of native Spanish-speakers. Drawing upon the principles of folk linguistics (cf. Niedzielski & Preston, 2000; Preston, 1993, 2011), the survey included an acceptability judgment task (cf. Labov, 1975) and a country identification task. Based on the results of the production data from Twitter, which show, as will be presented below, experiential *gustar* to be most common in Mexican Spanish, the purpose of the survey was to measure the extent to which Spanish-speakers perceive experiential *gustar* as being more a feature of Mexican Spanish than of other dialects.

The survey was open for ten days between November 25 and December 4, 2012, during which time 81 respondents from 14 different countries completed it. Participation in the survey was voluntary and the recruitment of participants was accomplished through the social networks of the author. The native Spanish-speaking friends, family members, and colleagues of the author and their native Spanish-speaking friends, family members, and colleagues were invited to complete the survey at their convenience within the time period that the survey was open. Appendix B gives the country of origin of the survey participants.

The survey was created by selecting five tweets with experiential *gustar* that represented common uses of the form. A nearly identical version of each tweet was then created, the only difference being the use of causal *gustar* instead of experiential *gustar*. This procedure resulted in five pairs of tweets: the original tweet with experiential *gustar* produced on Twitter, and a version of the tweet with causal *gustar*. The ten tweets were stored in a MySQL database and the webpage that displayed the survey was dynamically generated by a server-side PHP script each time it was accessed. The reason for storing the tweets in a database and using a dynamically generated webpage, rather than a static one, was to randomize for each participant the order of the five pairs of tweets as well as the order of the two sentences within each pair.[7] The decision to randomize the order of the tweets within the survey was driven by the assumption that uniformity of order of presentation of the tweets to the participants might have skewed the results.

Two tasks comprised the survey. First, an acceptability judgment task (cf. Labov, 1975) was carried out by asking the respondents to decide which tweet in each pair they preferred or sounded most natural in their country of origin. Secondly, based on the concepts of folk linguistics, the respondents were asked to write the name of the country where they would expect to hear the other tweet, the one that they did not choose. Finally, sociodemographic information was gathered, including, and most importantly, the country in which the respondent grew up. This latter information was used to correlate the respondents' choice of tweet with their country of origin in order to analyze any patterns of preference for the tweet with experiential gustar and certain countries. The text of the survey is included as an HTML file in Appendix C.

3. Results

Both the production data and the perceptual data suggest that experiential *gustar* is used more often in Mexican Spanish than in the Spanish of other countries.

3.1 Production

Unsurprisingly, causal *gustar* is used more frequently than experiential *gustar* in the tweets utilized for analysis. Of the 6,686 tokens of *gustar* analyzed, only 325 tokens, or 4.9%, contain examples of experiential *gustar*, leaving 6,361 tweets with causal *gustar*. This confirms the intuition that causal *gustar* is the more frequent, and canonical, syntactic form of this verb.

With respect to the rates of usage of experiential *gustar* between the cities, tweets sent from Mexico City contain this form more often, by far, than tweets sent from other cities. Of the 982 tweets sent from Mexico City, 205, or 20.9% contain experiential *gustar*. This rate is substantially larger, more than four times greater, than the rate in the city with the second highest usage, Buenos Aires, with 5.2% (36 of 691 tweets). See Table 1.

Map 1 presents a spatial representation of the information shown in Table 1, and highlights the fact that Mexico City stands out among the capital cities as having the highest rate, by far, of experiential *gustar*.

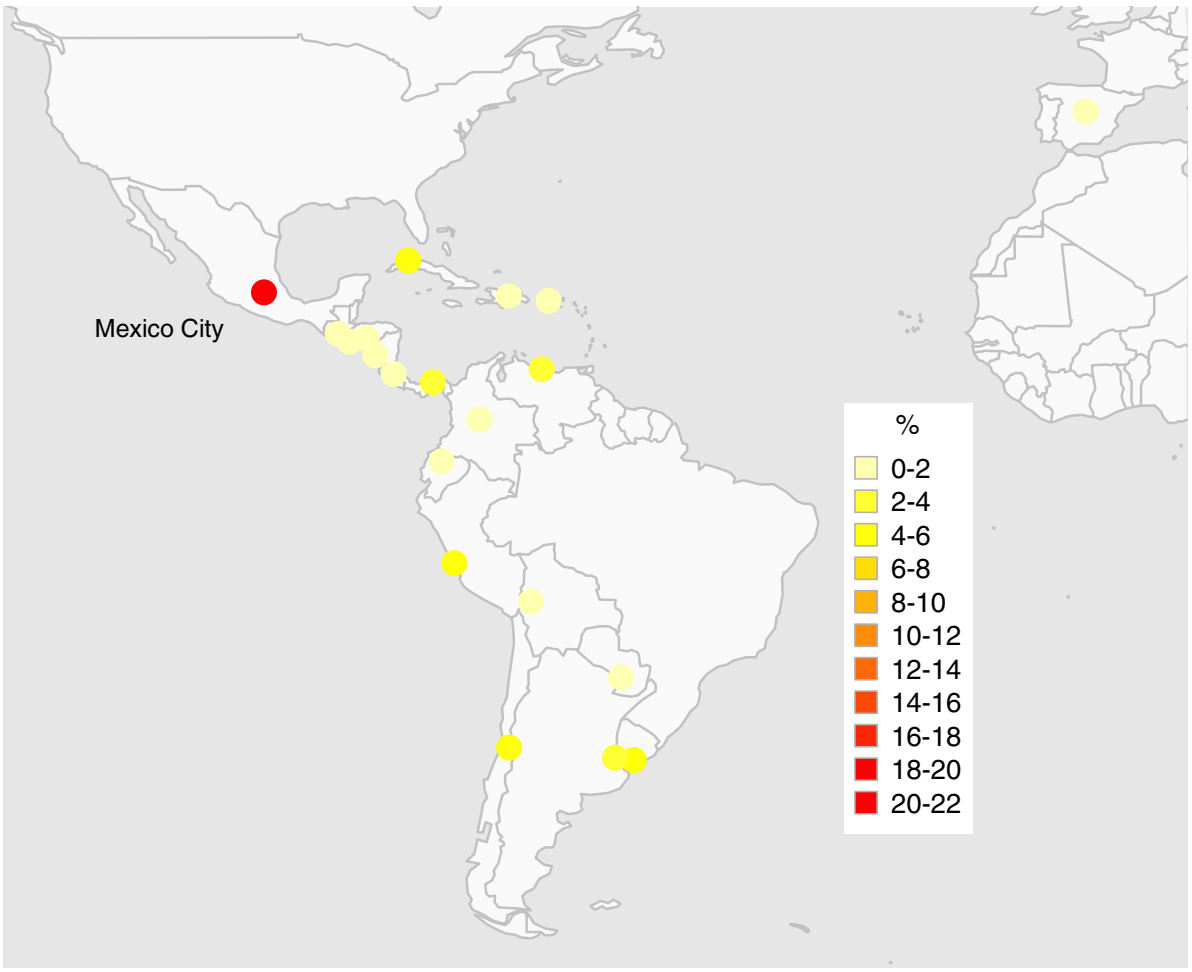
3.2 Frequent constructions with experiential gustar

An analysis of the 325 tokens of experiential *gustar* reveals several collocations and uses that experiential *gustar* frequently occurs in: in the subjunctive mood after the word *cuando* 'when', in the indicative mood after the word *si* 'if', and as a tag question. All of these constructions are used to extend an invitation, whether implicitly or explicitly. Of the 325 tokens of experiential

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

Table 1. *Distribution of* gustar *on Twitter in fifteen capital cities of Spanish-speaking countries in Latin America and in Spain*

City	Causal *gustar*	Experiential *gustar*	N	% exp. *gus.*
Mexico City	777	205	982	20.9%
Buenos Aires	655	36	691	5.2%
Lima	449	20	469	4.3%
Santiago	525	14	539	2.6%
Montevideo	212	5	217	2.3%
Caracas	698	15	713	2.1%
Tegucigalpa	104	2	106	1.9%
San Jose	112	2	114	1.8%
Asuncion	269	4	273	1.5%
Bogota	652	9	661	1.4%
Madrid	745	9	754	1.2%
Quito	253	3	256	1.2%
San Salvador	381	1	382	0.3%
Guatemala City	350	0	350	0.0%
San Juan	179	0	179	0.0%



Map 1. Distribution of *gustar* on Twitter in capital cities of Spanish-speaking countries in Latin America and in Spain.

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

gustar, 194 tokens, or 59.7%, occur in one of these three constructions. Specifically, 111 of the 325 tweets with experiential *gustar* occur after the word *cuando* 'when', 44 tokens after *si* 'if', and 39 tokens as a tag question. Example 2 above displays the use of experiential *gustar* after *cuando*, example 3 below shows its use after *si*, and example 4 shows it used as a tag question.

(3)

"te	*invito*	*a*	*una*	*merienda*	*esta*
2S OBJ PRO	1S 'INVITE'	PREP	FEM INDEF ART.	'SNACK'	DEM ADJ
tarde	*en*	*mi*	*casa,*	*si*	***gustas**."*
'AFTERNOON'	PREP	POS ADJ	'HOUSE'	'IF'	2S 'LIKE'

'I'd like to invite you over to my house this afternoon for a snack, if you'd like.' (Male, Madrid, Spain 2011-11-10)

(4)

"Vamos	*a*	*perisur?!*	***Gustan**?! :)*	*bueno*	*ahorita*	*los*	*veo*
1P 'GO'	PREP	'PERISUR'	2P 'LIKE'	'WELL'	'SOON'	2P OBJ PRO	1S 'SEE'
chiquitines:	*mientras*	*dejen*	*me*	*pongo*	*guapa*	*e"*	
'DARLINGS'	'WHILE'	2P 'RESPOND'	1S REF PRO	1S 'MAKE'	'BEAUTIFUL'	'OK'	

'Should we go to Perisur [Shopping Mall]?! Wanna go?! :) Well, I'll see you darlings in a minute: While you guys respond I'll make myself beautiful, ok.' (Female, Mexico City, Mexico 2011-11-10)

While not necessarily a frequent use of experiential *gustar*, a specific usage of this syntactic form of the verb that may be more geographically confined than others is the use of experiential *gustar* with the preposition *de* 'of' followed by a person. Examples 5 and 6 exemplify this usage in two tweets sent from Buenos Aires and Montevideo, respectively.

(5)

"Ahhh	*me*	*dice*	*q [que]*	*le*	*gusto!!!*	*Tenemos*
'AH'	1S OBJ PRO	3S 'SAY'	'THAT'	3S OBJ PRO	1S 'PLEASE'	1P 'HAVE'
9	*años???*	*Yo*	***gusto***	*de*	*vos...*	*Tomatelaaa"*
'NINE'	'YEAR'	1S SUB PRO	1S 'LIKE'	PREP	2S SUB PRO	'TAKE THAT'

'Ahhhh he tells me that he likes me!!! But, haven't we been together 9 years??? I like you... Take that!' (Female, Buenos Aires, Argentina 2011-11-12)

(6)

"para	*sentirme*	*una*	*escolar*	*yo*	*debería*	*decir*	*'yo*
PREP	INF 'FEEL' + 1S PRO	FEM INDEF ART	'SCHOOL GIRL'	1S SUB PRO	1S COND 'SHOULD'	INF 'SAY'	1S SUB PRO
gusto	*dél,*	*pero*	*él*	*no*	*gusta*	*mío'*	*#soydesalto*
1S 'LIKE'	PREP + 3S SUB PRO	CONJ	3S SUB PRO	NEG	3S 'LIKE'	POS PRO	'#SOYDESALTO'

'To feel like a school girl I should say "I like him, but he doesn't like me" #soydesalto [IAmFromSalto]' (Female, Montevideo, Uruguay 2011-11-10)

The results suggest that this use of experiential *gustar* may be a feature of Rioplatense Spanish, even though it is not exclusive to that region. While overall only 12.6% of the tweets with experiential *gustar* (41 out of 325 tweets) were written in Buenos Aires and Montevideo, 61% of the tweets containing experiential *gustar* with the preposition *de* 'of' followed by a person (11 out of 18 tweets) were written in those two cities. Of the capital cities outside of this region, there are two examples of this usage of experiential *gustar* in Asunción, and one each in Quito, Lima, Bogotá, San Salvador, and Mexico City. However, it should be noted that the limited number of tweets with this usage (N = 18) makes it difficult to conclude whether this type of experiential *gustar* is a peculiarity of Rioplatense Spanish more so than of other dialects. Future studies should investigate this possibility further.

In summary, an analysis of 6,686 tweets sent from within a 100-kilometer radius of fifteen capital cities in Spanish-speaking Latin America and in Spain show that,[8] while by no means exclusive to Mexico City, experiential *gustar* is utilized much more often in tweets sent from that city than in the tweets sent from other cities. Also, experiential *gustar* is frequently used in three constructions that extend invitations: after *cuando* 'when', after *si* 'if', and as a tag question. These three constructions account for nearly 60% of the 325 tokens of experiential *gustar* in the data. Additionally, the results suggest that one usage may be somewhat more geographically confined to Rioplatense Spanish, that of experiential *gustar* with the preposition *de* followed by a person. However, the small number of tokens of this particular usage of experiential *gustar* (N = 18) requires that future research take up this issue to confirm or disconfirm this possibility.

3.3 Online Survey

To the extent that the usage of *gustar* in the capital city of Spanish-speaking countries is generally representative of the usage of *gustar* in the rest of their corresponding countries,[9] the production data from Twitter suggest that experiential *gustar* is more prominent in Mexican Spanish than in the Spanish of other countries. This begs the question: Do native speakers of Spanish perceive experiential *gustar* as a feature of Mexican Spanish more so than as a feature of the Spanish of other countries? As explained above, the purpose of the online survey was to attempt to answer this question.

The results of both parts of the online survey, the acceptability judgment task and the country identification task, concur with the production data from Twitter by suggesting that, on average, experiential *gustar* seems to be a feature of Mexican Spanish more so than a feature of the Spanish of other countries. Of the 81

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

respondents of the survey, 19 grew up in Mexico, as seen in Appendix B. Interestingly, in the acceptability judgment task, a significantly higher percentage of these Mexican respondents preferred the tweets with experiential *gustar* than did non-Mexicans, as seen in Table 2.

As before, Map 2 shows that Mexico City stands out among the capital cities as having the highest rate of experiential *gustar* in the subjunctive mood.

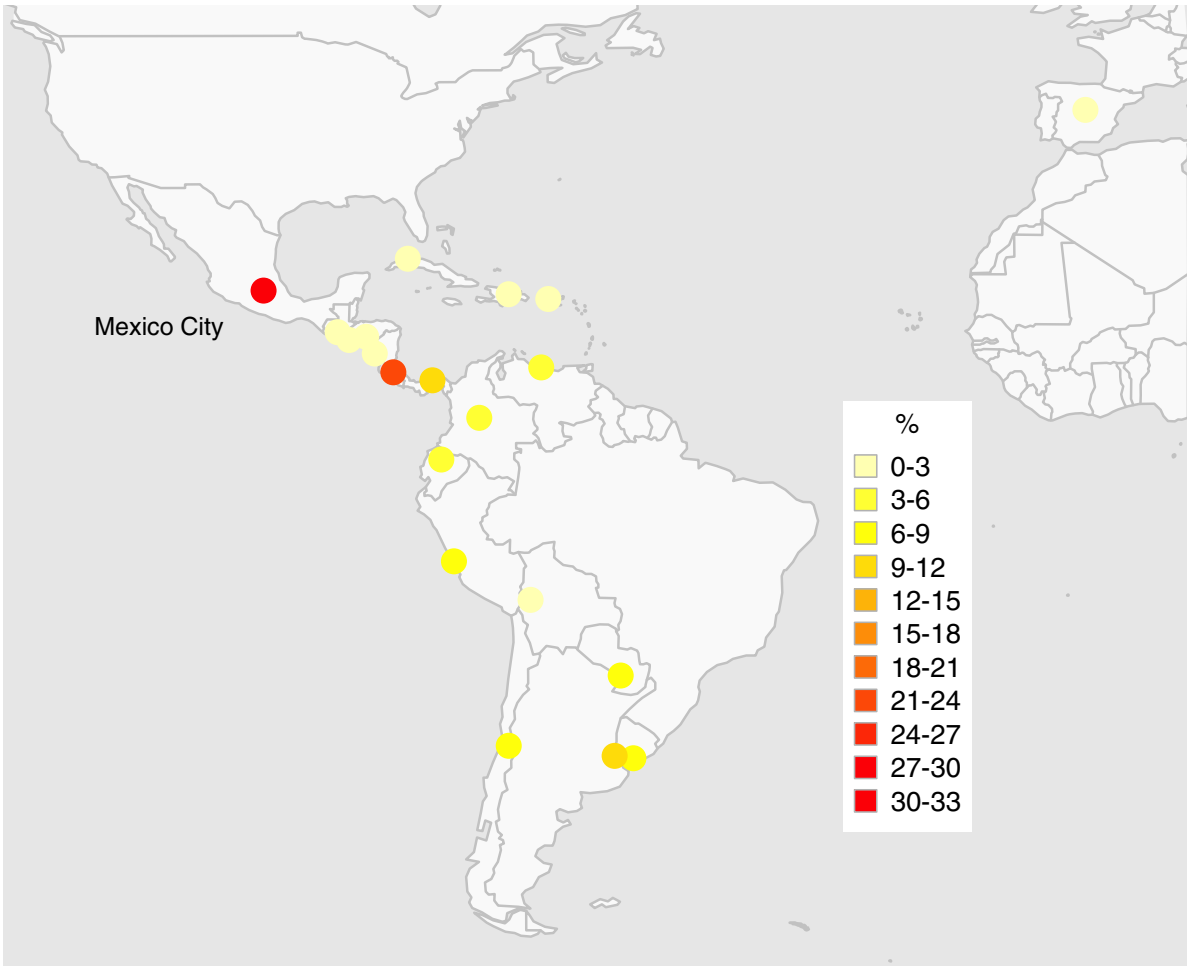
Table 2. *Distribution of preference for tweets by country of origin of online survey respondents*

Grew-up country	Cau. *gus.* tweets	Exp. *gus.* tweets	N	% exp. *gustar*	% data
Mexico	42	53	95	56%	23%
Other country	200	110	310	35%	77%
Total	242	163	405	40%	100%
Chi-square = 12.47, df = 1, p ≤ 0.001, phi = 0.18					

As presented above, one particular use of experiential *gustar*, that of *gustar* with the preposition *de* 'of' followed by a person, may be a peculiarity of Rioplatense Spanish, although it is likely not exclusive to that region. One of the five pairs of tweets featured in the survey included this presumed Rioplatense usage. When this pair of tweets was removed from the analysis, the difference between Mexicans and non-Mexicans in their preference for the tweets with experiential *gustar* increased. With the presumed Rioplatense use of experiential *gustar* excluded, the Mexican respondents chose the tweets with experiential *gustar* 70% of the time while the non-Mexican respondents chose those tweets only 44% of the time. See Table 3.

Map 3 again emphasizes the fact that the tweets sent from Mexico City stand out as having the highest rate of experiential *gustar* by far.

The results from the acceptability judgment task represent strong evidence that experiential *gustar*, in general, seems to be a feature of Mexican Spanish more so than a feature of the Spanish of other countries.



Map 2. Distribution of *gustar* in the subjunctive mood on Twitter in capital cites of Spanish-speaking countries in Latin America and in Spain.

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

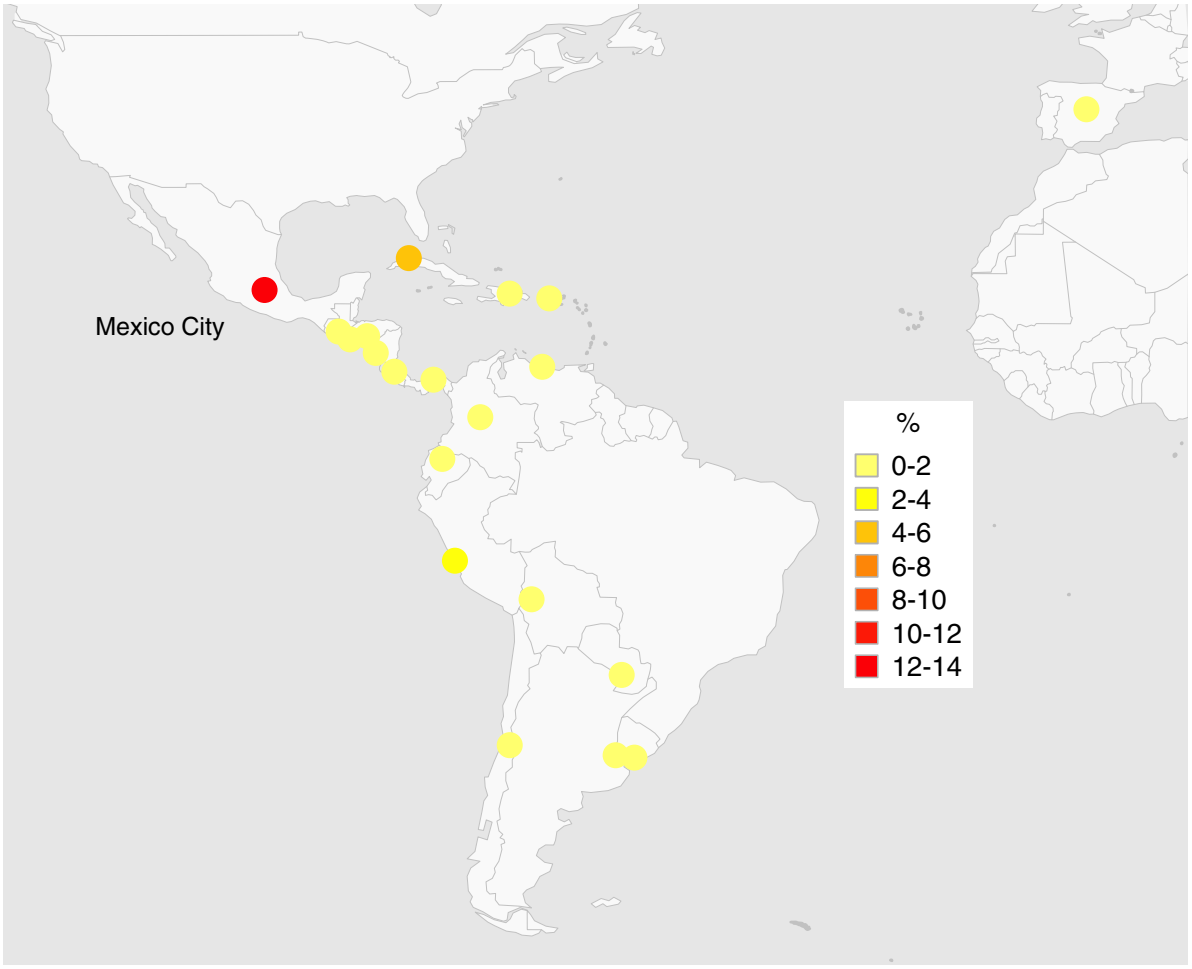
Further support for this conclusion is gleaned from the analysis of the countries in which experiential *gustar* is perceived to be used. As stated above, the respondents of the survey also participated in a country identification task. The participants were instructed to choose the tweet in each pair that they preferred or that sounded most natural in their country of origin (the acceptability judgment task), and then to write the name of the country where they would expect to hear the other tweet, the one that they did not choose (the country identification task). The results of this country identification task show that the non-Mexican respondents wrote in "Mexico" more often than other countries when the other, unchosen tweet contained experiential *gustar*.[10] See Table 4.

Table 3. *Distribution of preference for tweets by country of origin of online survey respondents, excluding the presumed Rioplatense use of experiential* gustar

Grew-up country	Cau. *gus.* tweets	Exp. *gus.* tweets	N	% exp *gustar*	% data
Mexico	23	53	76	70%	23%
Other country	140	108	248	44%	77%
Total	163	161	324	50%	100%
Chi-square = 15.96, df = 1, p ≤ 0.001, phi = 0.22					

Interestingly, when the tweet with the presumed Rioplatense use of experiential *gustar* and its corresponding version with causal *gustar* were removed from the analysis, "Mexico" was written in at an even higher rate as the perceived country in which respondents expected to hear the other, unchosen tweets when those tweets contained experiential *gustar*. See Table 5.

In summary, the results of the acceptability judgment task and the country identification task in the online survey provide perceptual data that concur with the results obtained from the production data extracted from Twitter; although not exclusive to Mexico, experiential *gustar* is associated more with Mexican



Map 3. Distribution of *gustar* in the indicative mood on Twitter in capital cites of Spanish-speaking countries in Latin America and in Spain.

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

Table 4. *Distribution of tweets by perceived country where the non-Mexican respondents expected to hear the unchosen tweets*

Perceived country	Cau. *gus.* tweets	Exp. *gus.* tweets	N	% exp *gustar*	% data
Mexico	15	52	67	78%	22%
Other country	95	148	243	61%	78%
Total	110	200	310	65%	100%
Chi-square = 6.4, df = 1, p = 0.01, phi = 0.14					

Table 5. *Distribution of tweets by perceived country where the non-Mexican respondents expected to hear the unchosen tweets, excluding the presumed Rioplatense use of experiential* gustar

Perceived country	Cau. *gus.* tweets	Exp. *gus.* tweets	N	% exp *gustar*	% data
Mexico	15	41	56	73%	23%
Other country	93	99	192	52%	77%
Total	108	140	248	56%	100%
Chi-square = 8.27, df = 1, p = 0.004, phi = 0.18					

Spanish than with the Spanish of other countries. The Mexican respondents of the survey preferred the tweets with experiential *gustar* more often than did the non-Mexican respondents. Likewise, among the non-Mexican respondents, the unchosen or dispreferred tweets that contained experiential *gustar* were more often expected to be heard in Mexico than in other countries.

4. Discussion

In answer to the first research question posed in Section 1, the production data from Twitter and the perceptual data from the online survey show that experiential *gustar* is used in many parts of the Spanish-speaking world, but that it seems to be a feature of Mexican Spanish more so than a feature of the Spanish of other countries. In response to the second research question, the production data and the perceptual data paint a similar picture with regard to the geographic distribution of experiential *gustar*. Finally, in response to the third research question, the production data show that experiential *gustar* occurs at a much lower rate than causal *gustar*, even in Mexican Spanish.

Several important implications are apparent from the methodology of this paper. First, the paper demonstrates the value of using both production and perceptual data in the study of language use and regional variation within language, thus allowing for a more complete picture of a given variable. The two types of data complement each other and can even feed into each other. For example, without perceptual data, the question remains as to whether native speakers perceive what the results from production show. In other words, when production data clearly point to one conclusion, as do the data from Twitter in this paper, the question naturally arises: "Do native speakers perceive X to be an important feature of the language of region Y?" Perceptual data can answer such a question. From the other point of view, if only perceptual data are utilized, researchers with an interesting finding may be left with the question: "Is this perception based on what speakers actually do in everyday language or is it simply a stereotyped perception?" By utilizing both production and perceptual data, researchers can gain a more complete understanding of linguistic phenomena.

This paper also contributes to the growing literature that demonstrates the usefulness of employing Twitter as a source of linguistic data. The fact that morpho-syntactic phenomena can so easily be retrieved from Twitter with virtually any computer scripting language makes it nearly effortless to retrieve thousands of data points from across the world in a matter of minutes. Granted, there are limitations on the information that can be captured about the users who write the tweets. Nevertheless, the ability to provide a snapshot of a given linguistic variable across huge language regions, such as the Spanish-speaking world, is promising. In summary, this paper demonstrates that Twitter can be employed as a geographically delimitable corpus of linguistic data.

Additionally, this paper takes a first step in documenting where experiential *gustar* in Spanish is used most and its rates of usage in comparison to causal *gustar*. As discussed above, virtually no works on Hispanic dialectology even mention its usage, let alone describe where it may be most common. Consequently, this paper represents a significant contribution to the literature that details dialectal variation in the Spanish-speaking world.

Further, as noted above, the results show that experiential *gustar* appears most frequently in only a few constructions used to extend an invitation. This finding holds an interesting implication for the teaching of Spanish as a second language. An analysis of ten beginning- and intermediate-level textbooks intended for second language learners of Spanish (Andradeet al., 2013; Blitt & Casas, 2011; Castells et al., 2009; Dorwick et al., 2006; Dorwick et al., 2011; Gonzalez-Aguilar & Rosso-O'Laughlin, 2011; Leeser et al., 2010; Murillo & Dawson, 2012; Ramos & Davis, 2008; Renjilian-Burgy et al., 2003) shows that these few uses of experiential *gustar* are not presented to the learners. Of course, the authors of these textbooks are primarily concerned with

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

helping their students gain a command of causal *gustar*, especially when the learners are native speakers of English and have a tendency to map the syntactic behavior of English *to like* to causal *gustar*, and thus make a transfer error. Nonetheless, the results of this paper seem to imply that teaching the few constructions in which the Twitter users most often utilize experiential *gustar* can help second language learners develop an idiomatic way of extending invitations. This is especially true for second language learners who are likely to have contact with Mexican and Mexican-American Spanish, which is a large proportion of second language learners of Spanish in the United States. Interestingly, the authors of one textbook intended for more advanced students at the high-intermediate and advanced levels (Salazar et al., 2013) recently added two examples of experiential *gustar* to the 2013 edition of their book. Curiously, the examples of experiential *gustar* are in a tag question and after the word *si* 'if', two of the three most frequent constructions in which experiential *gustar* occurs in the Twitter results in this paper. This seems to suggest that Twitter can also be used as a source of authentic language for second language learners, something more authors of language textbooks should strive to include in their works.

5. Conclusions

This paper has demonstrated the utility of combining production data and perceptual data to analyze language use and regional variation. Also, it has illustrated the suitability and usefulness of accessing the popular social networking platform Twitter to retrieve geographically delimited language data. Finally, the paper presents an exploratory attempt to document where experiential *gustar* is used most commonly in the Spanish-speaking world. The results show that this syntactic form of the verb is produced and is perceived to be produced most often in Mexican Spanish, despite not being exclusive to that country.

Acknowledgments

Thanks are expressed to Michael LeBaron for help writing the SQL code that randomized the order of the sentences in the online survey. Kristi Brown helped with word-smithing in parts of the paper and I appreciate that help. I also thank the anonymous reviewers for their feedback, which greatly improved the paper.

Notes

[1] These linguistic examples come from Twitter. Despite the fact that these Twitter users have public accounts, as a courtesy to them and the people referred to, the messages have been anonymized. The city name and date between parentheses indicate from where and when the tweet was sent. Additionally, explanatory notes are included between brackets, as internet orthography can be difficult to interpret (cf. Myslín & Gries 2010 for Spanish internet orthography) and slang can vary from country to country. Appendix A contains the explanation of the interlinear gloss codes.

[2] All translations of tweets are mine.

[3] According to the internet tracking website Internet Live Stats: http://www.internetlivestats.com/twitter-statistics/

[4] The search of tweets sent in the capital city of Equatorial Guinea, Malabo, failed to return results, presumably because the users' default language in their web browser or electronic device was not set to Spanish.

[5] http://www.gps.ie/gps-lat-long-finder.htm

[6] It should be noted that two capital cities, Guatemala City and San Salvador, are less than 200 kilometers from each other, and therefore had the potential of capturing the same tweets in a 25-kilometer wide area halfway between the two cities. However, a manual inspection of the tweets from these two cities verified that no duplicated tweets were captured by the R script.

[7] With the following SQL code: SELECT * FROM survey INNER JOIN (select pair, rand() AS first_rand FROM survey GROUP BY pair) AS temp ON survey.pair = temp.pair ORDER BY first_rand, rand().

[8] The reader will remember that five cities were excluded, as they had fewer than 100 tweets.

[9] Determining if, in fact, this is the case is outside of the scope of this paper and should be taken up elsewhere.

[10] The responses of the Mexican participants were excluded from this analysis in order to avoid skewing the results, as writing in "Mexico" was not an option for them on this question.

References

Alonso, Amado. 1967. *Estudios lingüísticos*. Madrid: Gredos.

Alvarez, Irina Arguelles & Alfonso Muñoz Muñoz. 2012. An insight into Twitter: a corpus based contrastive study in English and Spanish. *Revista de Linguistica y Lenguas Aplicadas* (RLLA) 7. 37-50.

Alvar, Manuel. 1996. *Manual de dialectología española: El español de España*. Barcelona: Ariel.

Andrade, Magdalena, Jeanne Egasse, Elías Miguel Muñoz & María Cabrera-Puche. 2013. *Tu mundo: español sin fronteras*. 1st edn. New York, NY: McGraw-Hill.

Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2).135-160.

Blitt, Mary Ann & Margarita Casas. 2011. *Exploraciones*. 1st edn. Heinle: Cengage Learning.

Boyd-Bowman, Peter. 1960. *El habla de Guanajuato*. Ciudad de México: Imprenta Universitaria.

Castells, Matilde Olivella, Elizabeth E. Guzmán, Paloma E. Lapuerta & Judith E. Liskin-Gasparro. 2009. *Mosaicos:*

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

Spanish as a world language. 5th edn. Upper Saddle River, NJ: Prentice Hall.
Cotton, Eleanor Greet & John M. Sharp. 1988. *Spanish in the Americas*. (Romance Languages and Linguistics Series). Washington, D.C.: Georgetown University Press.
Dorwick, Thalia, Ana Maria Perez Girones, Marty Knorre, William R. Glass & Hildebrando Villarreal. 2006. *¿Que tal?: An Introductory Course*. 7th edn. Boston: McGraw-Hill.
Dorwick, Thalia, Ana María Pérez-Gironés, Anne Becher, Casilde Isabelli & A. Raymond Elliott. 2011. *Puntos de partida: An Invitation to Spanish*. 9th edn. San Francisco, CA: McGraw-Hill.
Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2010. A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 1277-1287. Stroudsburg, PA, USA: Association for Computational Linguistics.
García de Diego, Vicente. 1978. *Dialectología española*. Madrid: Ediciones Cultura Hispánica del Centro Iberoamericano de Cooperación.
Gentry, Jeff. 2014. *twitteR: R based Twitter client*. https://github.com/geoffjentry/twitteR. (Accessed:).
González-Aguilar, María & Marta Rosso-O'Laughlin. 2011. *Atando cabos: Curso intermedio de español*. 4th edn. Upper Saddle River, N.J.: Prentice Hall.
Haddican, Bill & Daniel Ezra Johnson. 2012. Effects on the particle verb alternation across English dialects. *University of Pennsylvania Working Papers in Linguistics* 18(2). 31-40.
Kany, Charles E. 1951. *American-Spanish Syntax*. 2nd edn. Chicago: University of Chicago Press.
Klee, Carol & Andrew Lynch. 2009. *El español en contacto con otras lenguas*, (Georgetown Studies in Spanish Linguistics). Washington, D.C: Georgetown University Press.
Kwak, Haewoon, Changhyun Lee, Hosung Park & Sue Moon. 2010. What is Twitter, a social network or a news media? *WWW '10: Proceedings of the 19th International Conference on World Wide Web*. 591-600. New York: ACM.
Labov, William. 1975. Empirical foundations of linguistic theory. In Robert Austerlitz (ed.), *The scope of American linguistics*, 77-133. Ghent, Belgium: The Peter de Ridder Press.
Leeser, Michael, Bill VanPatten & Gregory D. Keating. 2010. *Asi lo veo: Gente, Perspectivas, Comunicación*. 1st edn. New York: McGraw-Hill.
Lipski, John. 1994. *Latin American Spanish*. New York: Longman.
Long, Daniel & Dennis R. Preston (eds). 2002. *Handbook of perceptual dialectology*, vol. 2. Philadelphia: John Benjamins.
Merkhofer, Elizabeth M. 2013. *She, he and they trending on Twitter: Polyvocal pronouns and more-public messages*. Washington, D.C.: Georgetown University thesis.
Murillo, Maria C. Lucas & Laila M. Dawson. 2012. *Con brio: Beginning Spanish*. 3rd edn. Hoboken, N.J.: Wiley.
Niedzielski, Nancy A. & Dennis R. Preston. 2000. *Folk linguistics*. Berlin: Mouton de Gruyter.
Pascual Cabo, Diego. 2013. *Agreement reflexes of emerging optionality in heritage speaker Spanish*. Gainesville, FL: University of Florida dissertation.
Preston, Dennis R. 1989. *Perceptual dialectology*. Dordrecht: Foris.
Preston, Dennis R. 1993. Folk dialectology. In Dennis R. Preston (ed.), *American dialect research*, 333-377. Philadelphia: John Benjamins.
Preston, Dennis R. 1996. Where the worst English is spoken. In E. Schneider (ed.), *Focus on the USA*, 297-360. (Varieties of English around the World). Amsterdam: John Benjamins.
Preston, Dennis R. 2011. Methods in (applied) folk linguistics. *AILA Review* 24(1). 15-39.
Ramos, Alicia & Robert L. Davis. 2008. *Portafolio*. 1st edn. New York, NY: McGraw-Hill.
R Development Core Team. 2014. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.
Renjilian-Burgy, Joy, Susan M. Mraz, Ana Beatriz Chiquito & Verónica De Darer. 2003. *Reflejos*. 1st edn. Boston: Cengage Learning.
Ruiz Tinoco, Antonio. 2012. Twitter as a corpus for Spanish geolinguistic studies. *Sophia Linguistica* 60. 147-163.
Sala, Marius, Dan Munteanu, Valeria Neagu Tudora & Sandru Olteanu. 1982. *El español de América: Léxico*. Bogotá: Instituto Caro y Cuervo.
Salazar, Carmen, Rafael Arias & Sara L. de la Vega. 2013. *Avanzando: Gramática española y lectura*. 7th edn. Hoboken, NJ: Wiley.
Sang, Kim & Erik Tjong. 2011. How to use Twitter for linguistic research. *Tabu* 39(1-2). 62-71.
Silva-Corvalán, Carmen. 1994. *Language contact and change: Spanish in Los Angeles*. (Oxford Studies in Language Contact). New York: Oxford University Press.
Zamora Munné, Juan C. & Jorge M. Guitart. 1988. *Dialectología hispanoamericana*. 2nd edn. Salamanca: Publicaciones del Colegio de España.
Zamora Vicente, Alonso. 1985. *Dialectología española*. 2nd edn. Madrid: Editorial Gredos.

Appendix A: Interlinear gloss codes

1S = first person singular
2S = second person singular
3S = third person singular
1P = first person plural
2P = second person plural
3P = third person plural
ADJ = adjective
ART = article
COND = conditional
CONJ = conjunction
DEF = definite
DEM = demonstrative
FEM = femenine
INDEF = indefinite
INF = infinitive
MAS = masculine
NEG = negation
OBJ = object
POS = possessive

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press

PREP = preposition
PRO = pronoun
REFL = reflexive
REL = relative
SUB = subject
+ indicates a compound word
'...' indicates an English translation

Appendix B: Number of online survey respondents by the country in which they grew up

Argentina	1
Bolivia	2
Chile	2
Colombia	16
Ecuador	1
El Salvador	7
Mexico	19
Paraguay	1
Peru	2
Puerto Rico	2
Spain	18
Uruguay	1
United States	2
Venezuela	7
Total	81

Appendix C: Online survey

Encuesta_sobre_gustar.html

https://doi.org/10.1017/jlg.2016.1 Published online by Cambridge University Press