Dialectologia. Special issue, VII *(2017), 149-167.*
ISSN: 2013-2247

Received 20 September 2017.
Accepted 30 October 2017.

VARIATION OF THE SECOND PERSON SINGULAR OF THE SIMPLE PAST TENSE IN *TWITTER*: *HICISTE* VS. *HICISTES* 'YOU DID'[1]

Antonio RUIZ TINOCO
Universidad Sophia*
a-ruiz@sophia.ac.jp

Abstract

In this study, the geographical distribution of the *-s* added to the simple past tense of indicative in several high frequency verbs is analyzed, i.e. *hicistes* vs *hiciste*. For doing so, a geocorpus formed by geolocated tweets has been used. The data obtained suggest that the traditional explanation of addition by analogy with other verb tenses is still valid, especially in rural areas, although it is not the only factor. In urban areas of *voseo*, as is the case of the Río de la Plata variety, there is no great difference in the ratio of *-s* added with the surrounding rural areas, as had been considered in previous studies.

Keywords

morphological variation, geocorpus, social networks, GIS

LA VARIACIÓN DE LA SEGUNDA PERSONA DEL PLURAL DEL PRETÉRITO PERFECTO: *HICISTE* VS. *HICISTES*

Resumen

En el presente estudio se analiza la distribución geográfica de la *-s* añadida al pretérito simple de indicativo en varios verbos de alta frecuencia, como en *hicistes* vs *hiciste.* Para ello se ha utilizado un geocorpus formado por tuits geolocalizados. Los datos obtenidos sugieren que la explicación tradicional

[1] This study is a revised version of a communication in the XLIII Congress of the Japanese Association of Hispanists held at the University of Kanagawa, Japan, on October 7 and 8, 2017. This work has been developed under the FFI2013-41077-P project, funded by the Spanish Ministerio de Economía y Competitividad and by JSPS KAKENHI Grant Number 15K02527.
* Faculty of Foreign Studies Department of Hispanic Studies, 7-1 Kioicho, Chiyoda, Tokyo 102-0094.

de añadidura por analogía con otros tiempos verbales sigue siendo válida, especialmente en áreas rurales, aunque no es el único factor. En zonas urbanas de voseo como es el caso de la variedad rioplatense, no se aprecia una gran diferencia de ratio de *-s* añadidas con las zonas rurales circundantes, como se había considerado en estudios anteriores.

Palabras clave
variación morfológica, geocorpus, redes sociales, SIG

1. Introduction

The form of the 2nd person of the past tense ending in *-s* as *hicistes* 'you did' occurs frequently and is a phenomenon of variation on which there are abundant studies in almost all Spanish-speaking countries, such as the works by Charles E. Kany (1945) and John Lipski (1994). Lipski considers it the most acceptable non-standard form in the areas of *voseo*. The phenomenon has also been studied by María Vaquero (1998). Other more recent quantitative approaches are the works by Sonia Barnes (2012) and Claudia Parodi (2015) in Spanish-speaking areas of the United States, etc. Ralph Penny (2002) attributes *hicistes* to the analog extension of the final *-s* in the other 2nd person forms. Although this form is not academically accepted, it is frequently used in the informal style. On the other hand, according to Frago Gracia (2003), it is observed in low and middle social classes in both Spanish and Latin American Spanish. The RAE (2010) considers the non-standard form non-correct today although it mentions that it is frequent in historical texts. In general, it is a rural form, discredited and stigmatized. Many of the studies of this phenomenon have been based on spoken Spanish in different periods, including historical Spanish and limited to some geographical areas. Thus, it is difficult to obtain a global vision.

This work analyzes the geographical distribution of the *-s* added to the simple past tense of indicative in a sample of geolocated tweets containing forms of several high frequency verbs in Twitter microblogs. For doing so, a geocorpus of approximately 32 million tweets from all the Spanish-speaking areas and collected between January 2016 and October 2017 has been used. When comparing the occurrences of standard

and non-standard forms, differences in frequency of use depending on each verb have been found. These differences cannot be attributed solely to the speaker's perception of an analogy with the 2nd person of other verbal tenses. Moreover, relative frequencies in areas of *voseo* do not seem to favor their use especially.

In Catalan, there are also several cases of contrast between *-s* and ∅, in this case in relation to the 2nd person of the plural of the imperative of certain verbs. This person does not show the usual 2nd person morpheme *-s* in most verbs, although it appears in some (*dir* 'to say', *ser* 'to be', *estar* 'to be', *fer* 'to do', etc.). In some dialects there is a contrast between the presence and the deletion of *-s*, especially when the verb is followed by a clitic. Thus, "La flexió verbal en els dialectes catalans", by Antoni M. Alcover & Francesc de B. Moll, records the following forms for the 2nd person singular of the imperative of the verb *dir* 'to say': *digues* – *dis*, according to the dialects, but *digue'm* vs *dis-me*, *digue-li* vs *digues-li* – *dis-li*, when they are followed by a pronoun. The results of an ongoing research will show if the forms are maintained in the respective dialectal areas, if the action of the standard may have increased the solutions with *-s* and if there are possible points in common with Spanish.

The data has been inserted into a MySQL relational database. For the preparation of the respective distribution maps and the visualization of the results of the analysis, open source software QGIS has been used.

The resulting corpus allows us to compare the use of such forms in the same periods of time in distant areas, which would not have been possible through traditional methods of data collection. In this way the phenomenon can be observed from different perspectives.

2. Geocorpus

The data comes from a geocorpus of collected tweets that have the following characteristics:

a. The data has been obtained through direct connection to Streaming API 1.1 of Twitter, [2] which is approximately 1% of all tweets that can be obtained freely.

b. Only geocoded tweets,[3] which are a fraction of the total, have been collected, and forwarded tweets, retweets, have been excluded because they are duplicates and do not necessarily coincide with the place of origin.

c. From January 1, 2016 to October 1, 2017, 24 hours a day, approximately 32,400,000 tweets have been collected; so, a high degree of synchrony can be considered.

d. For the preparation of this geocorpus and according to our objectives, a modified version of *140dev Streaming API Framework*,[4] programmed in PHP by Adam Green has been used.

e. The tweets are arranged in a MySQL database,[5] accessible with a password over the Internet.

3. Data selection

For our analysis we have chosen ten verbs in the 2nd singular personal form of the simple past tense with high frequencies. We have used the corpus of Spanish (Web-Dialects) [6] by Mark Davies (2016) applying the search formula "(*ste.[VIS*])" with which we obtain the frequency of the standard verb forms of the simple past of tense indicative ending in *-ste*, ordered by frequency. The form *fuiste*, corresponding to the past tense of the verbs *ser* 'to be' and *ir* 'to go', has been eliminated. In Figure 1 the list of obtained verbs is shown, from which we choose the ten most frequent ones: *hiciste* 'you did', *dijiste* 'you said', *diste* 'you gave', *tuviste* 'you had', *dejaste* 'you left', *llegaste* 'you arrived', *pusiste* 'you put', *escribiste* 'you wrote', *pudiste '*you were able to' and *estuviste* 'you were'.

[2] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data
[3] That is, containing the fields of longitude and latitude of the place of emission.
[4] http://140dev.com/free-twitter-api-source-code-library/
[5] https://www.mysql.com/
[6] https://www.corpusdelespanol.org/web-dial/

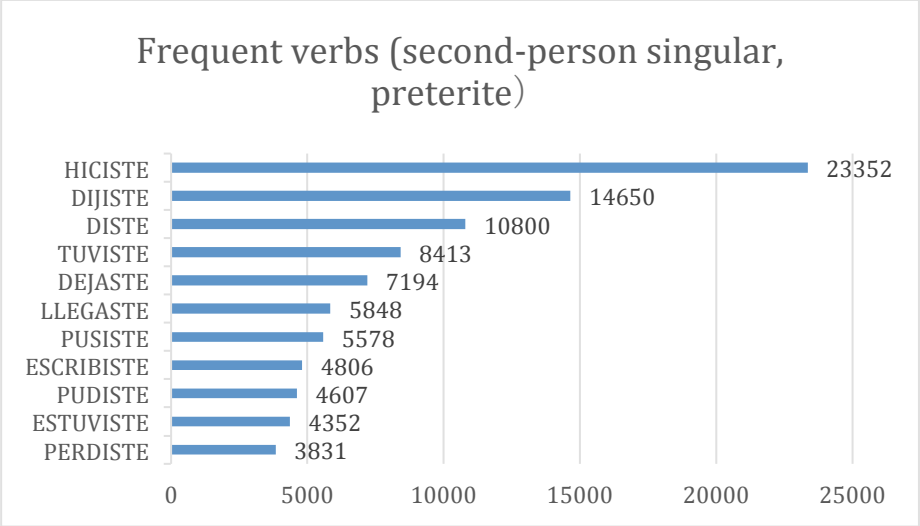
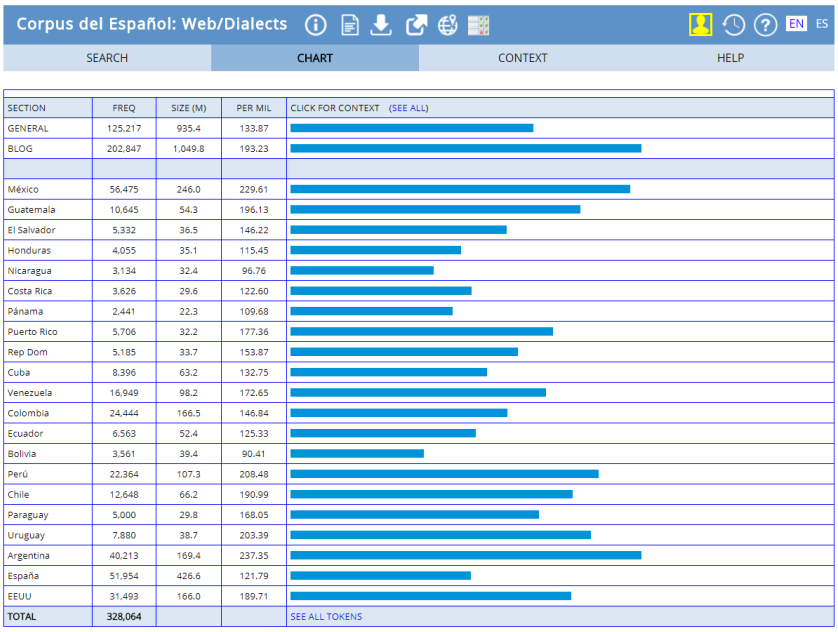


Figure 1. Frequent verbs in the 2nd person from the simple past

Although the Corpus of Spanish is lemmatized, it does not distinguish non-standard forms as *hicistes* 'you did', etc., so they have to be searched individually. Figure 2 shows the absolute and relative frequencies of *hicistes* in the Spanish-speaking countries.



SECTION	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
GENERAL	125,217	935.4	133.87	
BLOG	202,847	1,049.8	193.23	
México	56,475	246.0	229.61	
Guatemala	10,645	54.3	196.13	
El Salvador	5,332	36.5	146.22	
Honduras	4,055	35.1	115.45	
Nicaragua	3,134	32.4	96.76	
Costa Rica	3,626	29.6	122.60	
Panama	2,441	22.3	109.68	
Puerto Rico	5,706	32.2	177.36	
Rep Dom	5,185	33.7	153.87	
Cuba	8,396	63.2	132.75	
Venezuela	16,949	98.2	172.65	
Colombia	24,444	166.5	146.84	
Ecuador	6,563	52.4	125.33	
Bolivia	3,561	39.4	90.41	
Perú	22,364	107.3	208.48	
Chile	12,648	66.2	190.99	
Paraguay	5,000	29.8	168.05	
Uruguay	7,880	38.7	203.39	
Argentina	40,213	169.4	237.35	
España	51,954	426.6	121.79	
EEUU	31,493	166.0	189.71	
TOTAL	328,064			SEE ALL TOKENS

Figure 2. Frequencies of the form *hicistes* 'you did' in the Corpus of Spanish (Web-Dialects)

Dialectologia. Special issue, VII *(2017), 149-167.*
ISSN: 2013-2247

Moreover, Figure 2 shows that the frequency per million words varies between 0.03 p.m. from the Dominican Republic, although of a single occurrence, and 0.86 p.m. from Nicaragua. These data suggest that there is a considerable difference in the extent of the phenomenon among Spanish-speaking countries, but it is not possible to see more detailed distributions due to the methodology used for this corpus.

4. Interface

To search the geocorpus in the database, the Adminer[7] interface has been used, as shown in Figure 3.

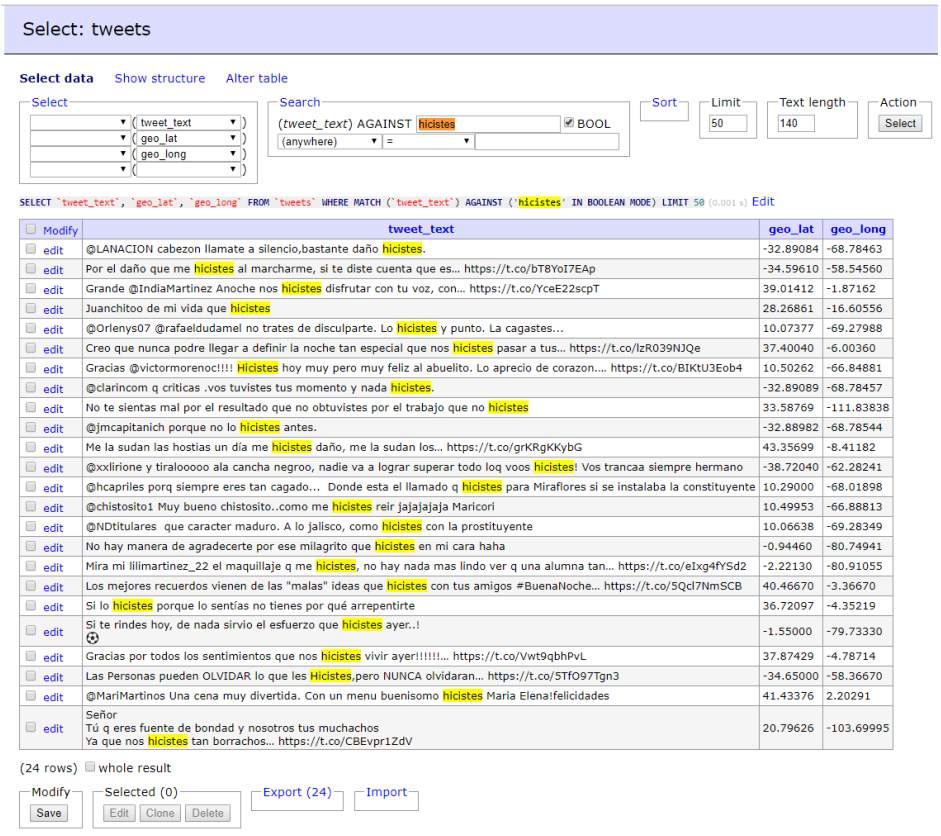
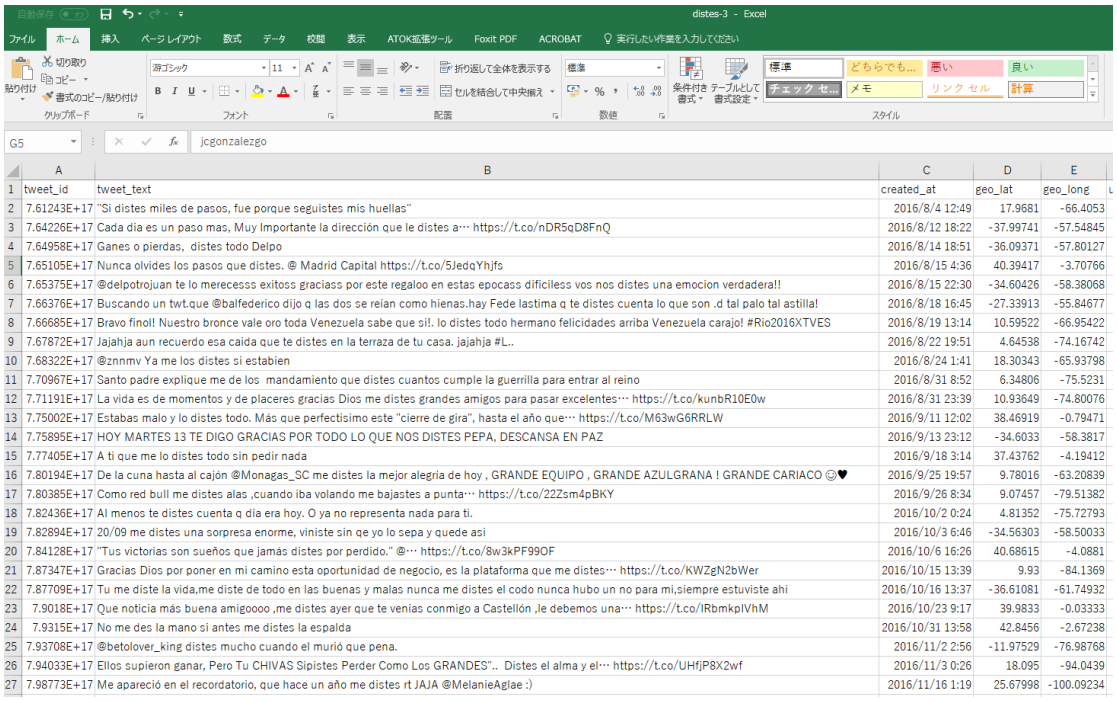


Figure 3. Adminer interface showing partial search results

Adminer has been used because it allows us to easily apply all the formulas available in MySQL, including regular expressions. It also allows to export the results to

[7] https://www.adminer.org/

various formats for further processing. In Figure 4 some data and several fields are shown in MS-Excel, including the text, the exact date and the coordinates of the place from where each tweet was sent.



	A	B	C	D	E
1	tweet_id	tweet_text	created_at	geo_lat	geo_long
2	7.61243E+17	"Si distes miles de pasos, fue porque seguistes mis huellas"	2016/8/4 12:49	17.9681	-66.4053
3	7.64226E+17	Cada dia es un paso mas, Muy Importante la dirección que le distes a··· https://t.co/nDR5qD8FnQ	2016/8/12 18:22	-37.99741	-57.54845
4	7.64958E+17	Ganes o pierdas, distes todo Delpo	2016/8/14 18:51	-36.09371	-57.80127
5	7.65105E+17	Nunca olvides los pasos que distes. @ Madrid Capital https://t.co/5JedqYhjfs	2016/8/15 4:36	40.39417	-3.70766
6	7.65375E+17	@delpotrojuan te lo merecessss exitoss graciass por este regaloo en estas epocass dificiless vos nos distes una emocion verdadera!!	2016/8/15 22:30	-34.60426	-58.38068
7	7.66376E+17	Buscando un twt.que @balfederico dijo q las dos se reían como hienas.hay Fede lastima q te distes cuenta lo que son .d tal palo tal astilla!	2016/8/18 16:45	-27.33913	-55.84677
8	7.66685E+17	Bravo finol! Nuestro bronce vale oro toda Venezuela sabe que sí!. lo distes todo hermano felicidades arriba Venezuela carajo! #Rio2016XTVES	2016/8/19 13:14	10.59522	-66.95422
9	7.67872E+17	Jajahja aun recuerdo esa caida que te distes en la terraza de tu casa. jajahja #L..	2016/8/22 19:51	4.64538	-74.16742
10	7.68322E+17	@znnmv Ya me los distes si estabien	2016/8/24 1:41	18.30343	-65.93798
11	7.70967E+17	Santo padre explique me de los mandamiento que distes cuantos cumple la guerrilla para entrar al reino	2016/8/31 8:52	6.34806	-75.5231
12	7.71191E+17	La vida es de momentos y de placeres gracias Dios me distes grandes amigos para pasar excelentes··· https://t.co/kunbR10E0w	2016/8/31 23:39	10.93649	-74.80076
13	7.75002E+17	Estabas malo y lo distes todo. Más que perfectisimo este "cierre de gira", hasta el año que··· https://t.co/M63wG6RRLW	2016/9/11 12:02	38.46919	-0.79471
14	7.75895E+17	HOY MARTES 13 TE DIGO GRACIAS POR TODO LO QUE NOS DISTES PEPA, DESCANSA EN PAZ	2016/9/13 23:12	-34.6033	-58.3817
15	7.77405E+17	A ti que me lo distes todo sin pedir nada	2016/9/18 3:14	37.43762	-4.19412
16	7.80194E+17	De la cuna hasta al cajón @Monagas_SC me distes la mejor alegria de hoy , GRANDE EQUIPO , GRANDE AZULGRANA ! GRANDE CARIACO ☺♥	2016/9/25 19:57	9.78016	-63.20839
17	7.80385E+17	Como red bull me distes alas ,cuando iba volando me bajastes a punta··· https://t.co/22Zsm4pBKY	2016/9/26 8:34	9.07457	-79.51382
18	7.82436E+17	Al menos te distes cuenta q día era hoy. O ya no representa nada para ti.	2016/10/2 0:24	4.81352	-75.72793
19	7.82894E+17	20/09 me distes una sorpresa enorme, viniste sin qe yo lo sepa y quede asi	2016/10/3 6:46	-34.56303	-58.50033
20	7.84128E+17	"Tus victorias son sueños que jamás distes por perdido." @··· https://t.co/8w3kPF99OF	2016/10/6 16:26	40.68615	-4.0881
21	7.87347E+17	Gracias Dios por poner en mi camino esta oportunidad de negocio, es la plataforma que me distes··· https://t.co/KWZgN2bWer	2016/10/15 13:39	9.93	-84.1369
22	7.87709E+17	Tu me diste la vida,me diste de todo en las buenas y malas nunca me distes el codo nunca hubo un no para mi,siempre estuviste ahi	2016/10/16 13:37	-36.61081	-61.74932
23	7.9018E+17	Que noticia más buena amigoooo ,me distes ayer que te venías conmigo a Castellón ,le debemos una··· https://t.co/IRbmkpIVhM	2016/10/23 9:17	39.9833	-0.03333
24	7.9315E+17	No me des la mano si antes me distes la espalda	2016/10/31 13:58	42.8456	-2.67238
25	7.93708E+17	@betolover_king distes mucho cuando el murió que pena.	2016/11/2 2:56	-11.97529	-76.98768
26	7.94033E+17	Ellos supieron ganar, Pero Tu CHIVAS Sipistes Perder Como Los GRANDES".. Distes el alma y el··· https://t.co/UHfjP8X2wf	2016/11/3 0:26	18.095	-94.0439
27	7.98773E+17	Me apareció en el recordatorio, que hace un año me distes rt JAJA @MelanieAglae :)	2016/11/16 1:19	25.67998	-100.09234

Figure 4. Partial data of *distes* 'you gave' in MS-Excel

5. Data obtained

The data obtained from all the Spanish-speaking countries of the ten selected verbs is shown in Table 1. There are 33,921 standard forms, 961 non-standard, and the ratio per verb varies between 1,97% for *hiciste/s* and 4,18% for *diste/s*.

VERB FORM	without *-s*	with *-s*	ratio %
estuviste/s	1613	49	3.03
pudiste/s	1060	30	2.83
escribiste/s	255	8	3.13
pusiste/s	1290	39	3.02
llegaste/s	6500	229	3.52

VERB FORM	without *-s*	with *-s*	ratio %
dejaste/s	3841	106	2.75
tuviste/s	1570	34	2.16
diste/s	3987	167	4.18
dijiste/s	4603	117	2.54
hiciste/s	9202	182	1.97
Total	33921	961	2.83

Table 1. Frequencies obtained from the forms with *-s* and without *-s*

The distribution of occurrences by country is shown in Table 2.

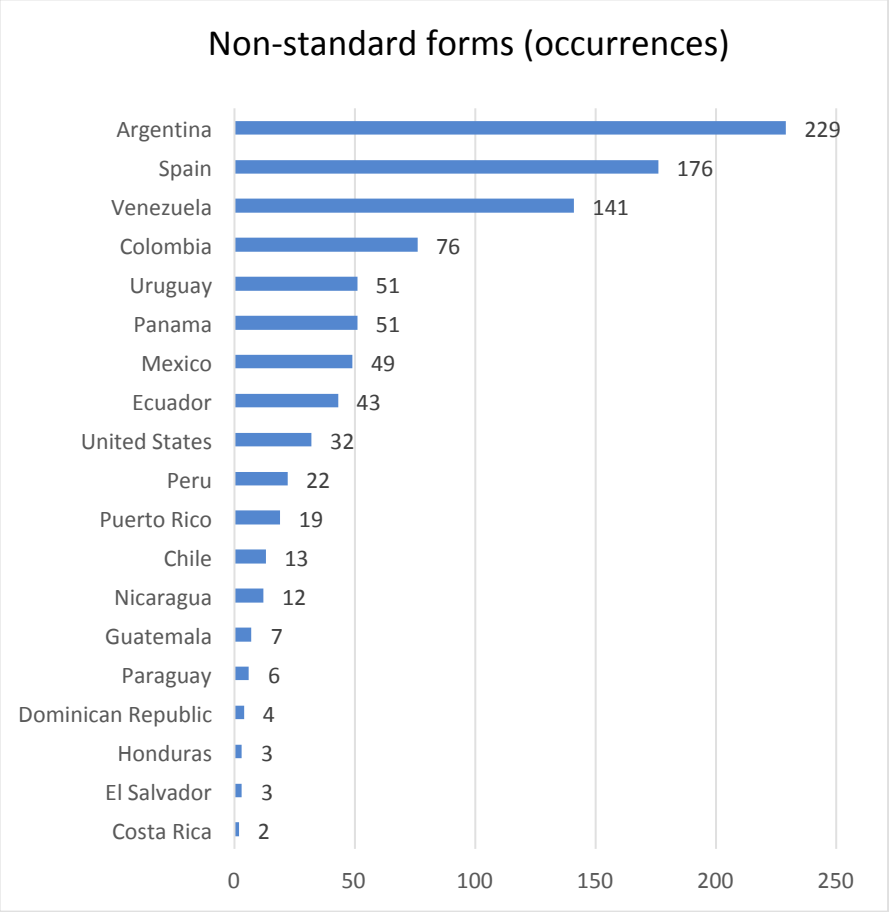


Table 2. Distribution of non-standard forms by country

Table 2 shows that Argentina, Spain and Venezuela have most occurrences; whereas, Costa Rica, El Salvador, etc. have fewest. For that reason, we need more data

so that these figures can be significant. Low figures are not only due to the lower total population, but also they may be due to the smaller population that uses Twitter.

Table 3 shows the proportion of use of the two forms, which gives us a new view of the phenomenon. Thus, the highest proportion is found in Puerto Rico (18.45%), Panama (14.74%), Nicaragua (13.33%), Venezuela (7.6%), etc. Countries such as Spain (5.31%) or Argentina (1.51%) have lower ratios as a whole, as shown in Figure 5.

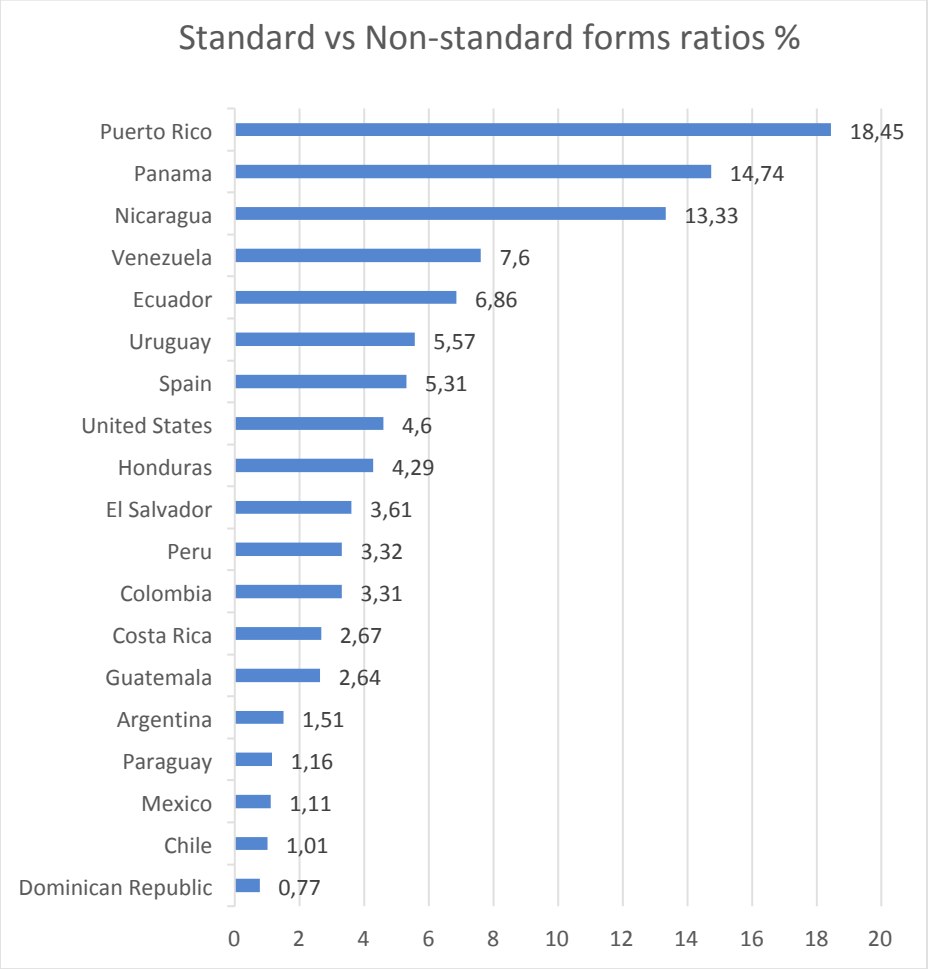


Table 3. Distribution by country of non-standard forms vs standard forms ratios

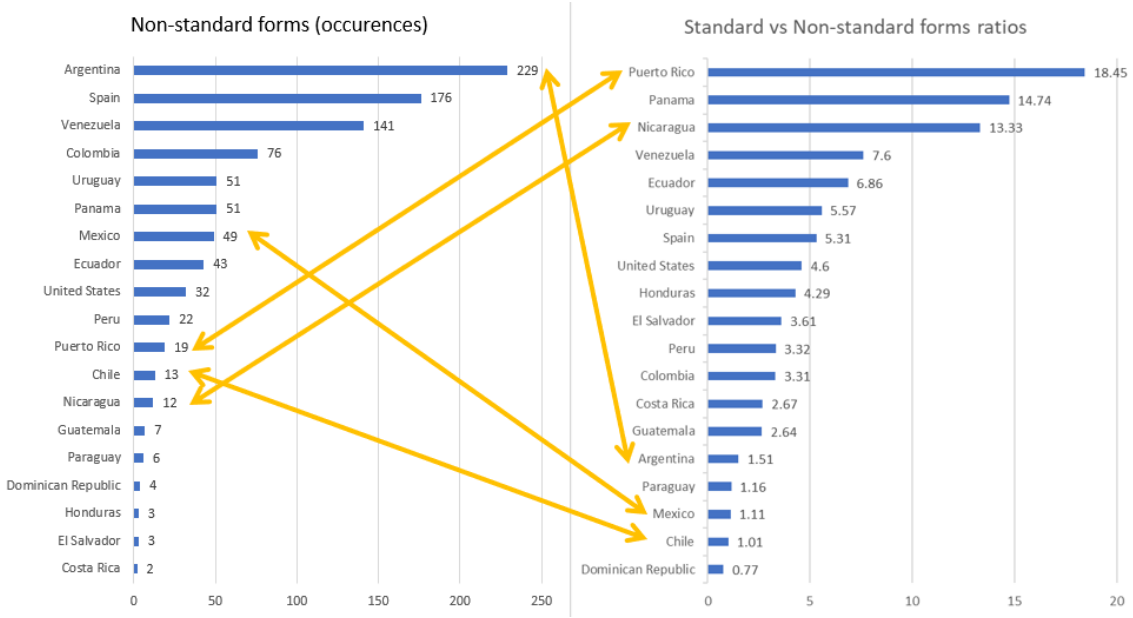


Figure 5. Absolute results (occurrences) vs. relative results (ratio)

6. Visualization of the data

For the visualization of the data, QGIS[8] ver. 2.18, a free and open source GIS application has been used. It contains a data management system as a database. This GIS system allows various forms of data selection and visualization. For example, Figures 6a and 6b show a small red dot for each occurrence of the standard (6a) and non-standard (6b) forms, proving that both occur in all Spanish-speaking areas, with higher absolute frequency in the most populated areas, such as Argentina, Spain, Venezuela, Colombia and other geographical areas.

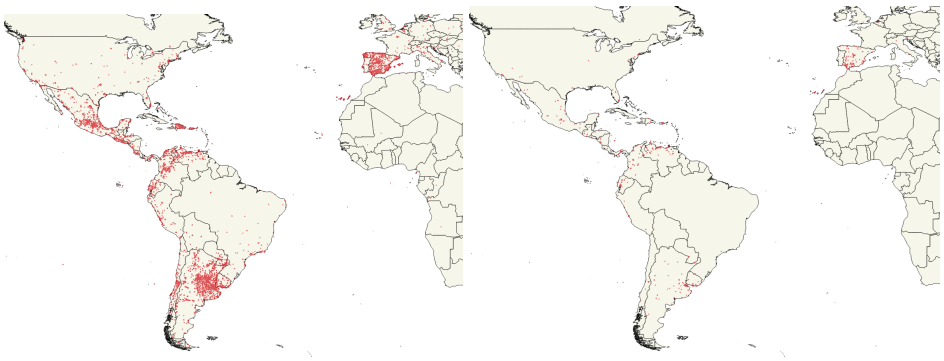


Figure 6a. Standard forms without *-s*

Figure 6b. Non-standard forms with *-s*

For a better visualization, the zones of higher absolute frequencies can be presented with heatmaps, as in Figures 7a and 7b. This method is a form of clustering, since it distinguishes the zones of higher absolute frequency by means of a scale of colors. The parameters for the color scale can be easily modified by varying the radius of influence (by meters, kilometers, etc.) of each of the occurrences. Figure 7a shows the heatmap corresponding to the distribution of the standard forms and Figure 7b shows the forms finishing with *-s*. In these maps we can see some areas (in red) where non-standard occurrences are concentrated, such as southern Spain, Río de la Plata, Venezuela, etc.

To better visualize the data with greater granularity we can reduce the map to a specific geographical area as shown in Figures 8a and 8b, of Spain and Argentina respectively, and for doing so we have modified the radius of influence.

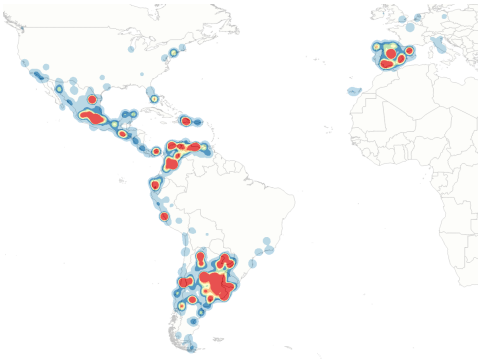


Figure 7a. Standard forms without *-s*

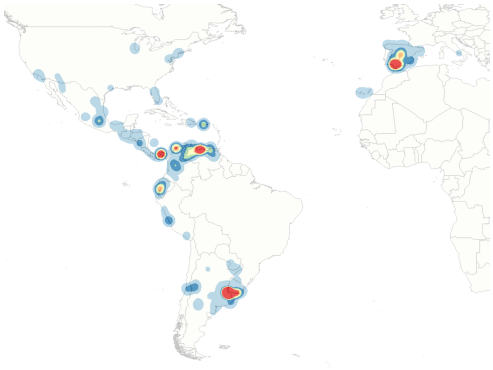


Figure 7b. Non-standard forms with *-s*

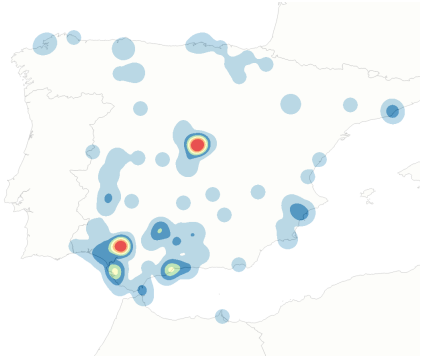


Figure 8a. Non-standard forms (Spain)

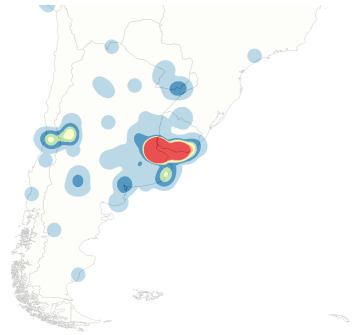


Figure 8b. Non-standard forms (Argentina)

Dialectologia. Special issue, VII *(2017), 149-167.*
ISSN: 2013-2247

Figure 8a shows that the geographical areas in Spain with most non-standard forms are Madrid and Seville, as well as the area around the Bay of Cadiz and the province of Malaga. Although Barcelona, Zaragoza and Valencia have a larger population than these southern areas, it is an indication of frequent use. In Catalonia, Asturias and Galicia there is no noticeable concentration of occurrences. Analogously, we can observe in Figure 8b that the area of Greater Buenos Aires (about 15 million inhabitants) and Montevideo concentrate many more occurrences than cities with also a large population such as Santiago de Chile (approximately 5 million inhabitants).

The absolute and relative data for each country can also be visualized using choropleth maps, with scales of colors like Figures 9 and 10. In Figure 9, Spain, Argentina, Mexico, Venezuela, Colombia, etc. are the countries with the darkest blue color, that is, with more occurrences than others. Figure 10 shows the countries with the highest ratio as shown in Table 3. In the choropleth maps, the same color or a graduated scale is usually used for each country (province, city, etc.) to mark regional, provincial differences, etc. Bolivia is white due to the lack of data.

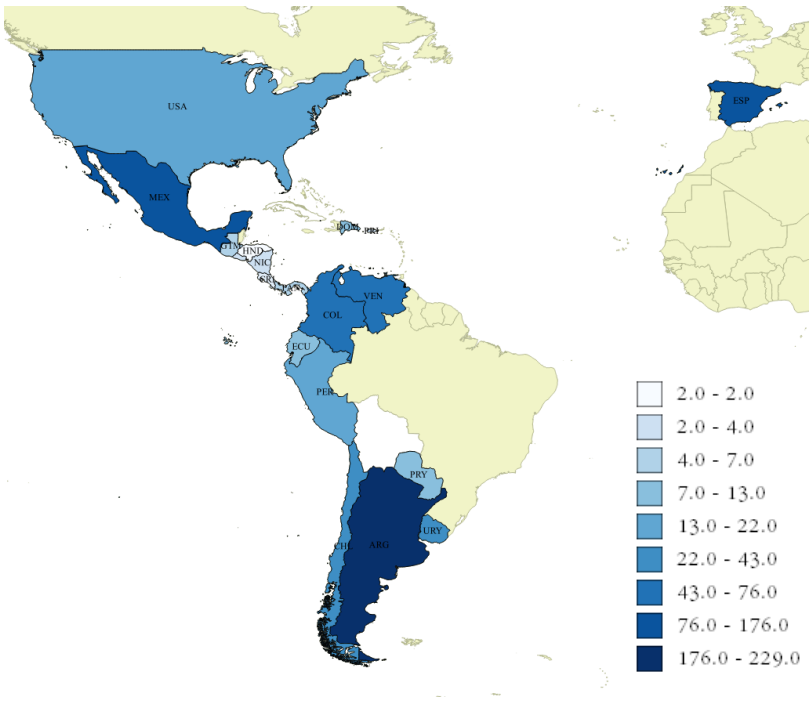


Figure 9. Choropleth maps of the occurrences of non-standard forms

Dialectologia. Special issue, VII *(2017), 149-167.*
ISSN: 2013-2247

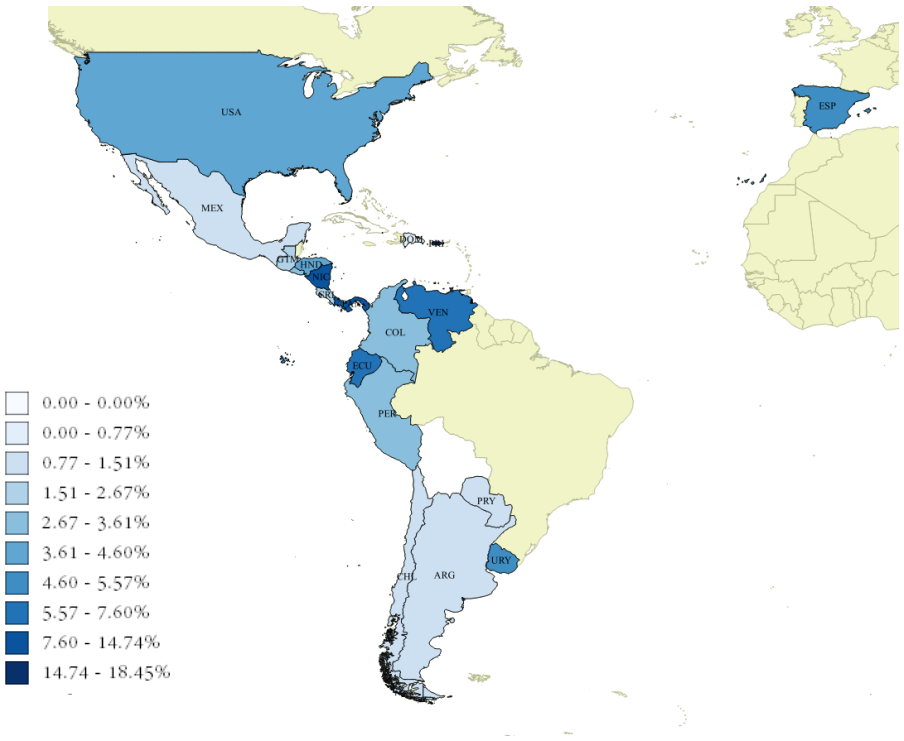


Figure 10. Choropleth map of the ratios of non-standard forms

Analogously to Figure 10, we can visualize the ratios by provinces, using QGIS with maps with less extensive administrative divisions, as shown in Figure 11 for Spain. This figure shows that the highest ratios are concentrated in the Southwest of the country, especially in the provinces of Cordoba, Huelva, Caceres and Cadiz as well as the Canary Islands, being virtually non-existent or with very low ratios in the rest of the country. The areas that include cities with a larger population such as Madrid, Barcelona, Valencia and Zaragoza are not found in zones that add the *-s*, except Seville, in the middle of an area that usually adds the *-s*. These areas coincide largely with southern varieties such as Andalusia and Extremadura, stigmatized in Spain.

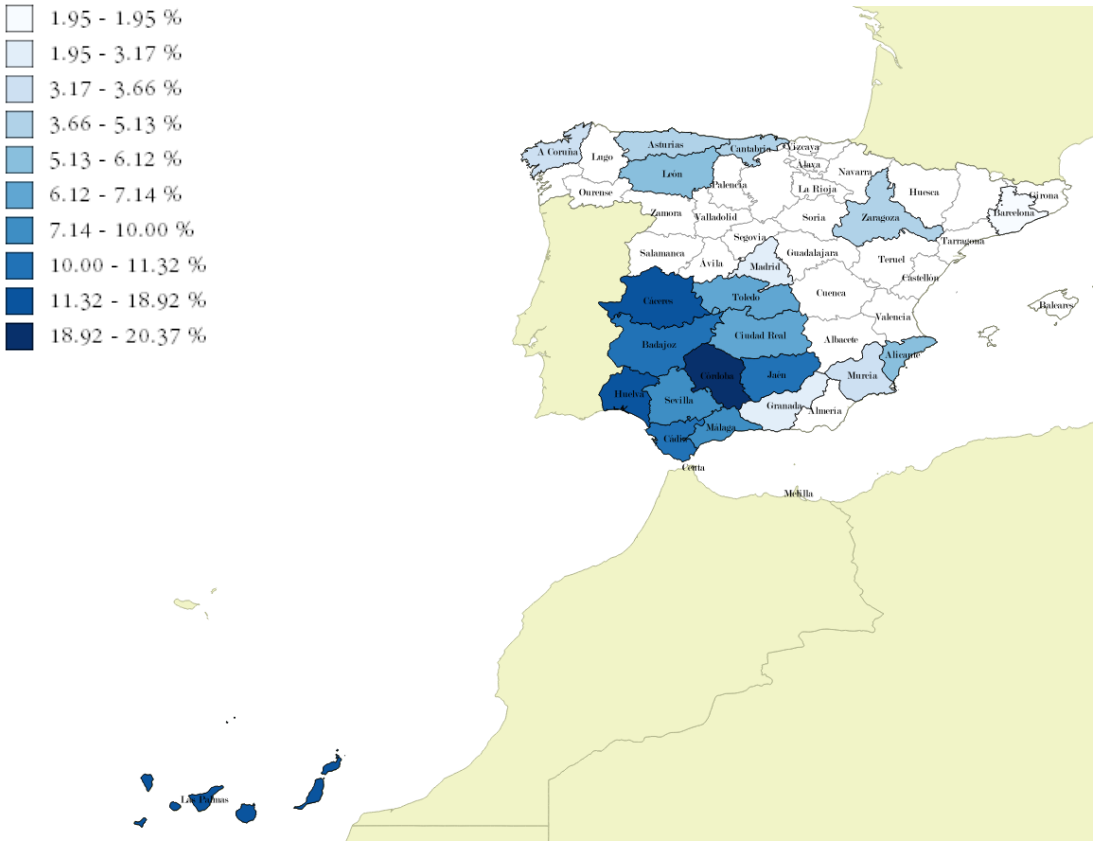


Figure 11. Choropleth map of the ratios of non-standard forms in the provinces of Spain

With the help of geographic data, often in shapefile[8] format, with details of the contours of any area, such as countries, cities, neighborhoods, etc., much more detailed maps can be prepared such as the following. In Figure 12 of the extensive province of Buenos Aires, the Federal District, in red on the map, contains the highest concentration of occurrences of non-standard forms. However, in the map of Figure 13 the surrounding areas offer the highest ratio of non-standard forms, as in Spain, these being more frequent in rural areas than in urban areas.

[8] https://en.wikipedia.org/wiki/Shapefile

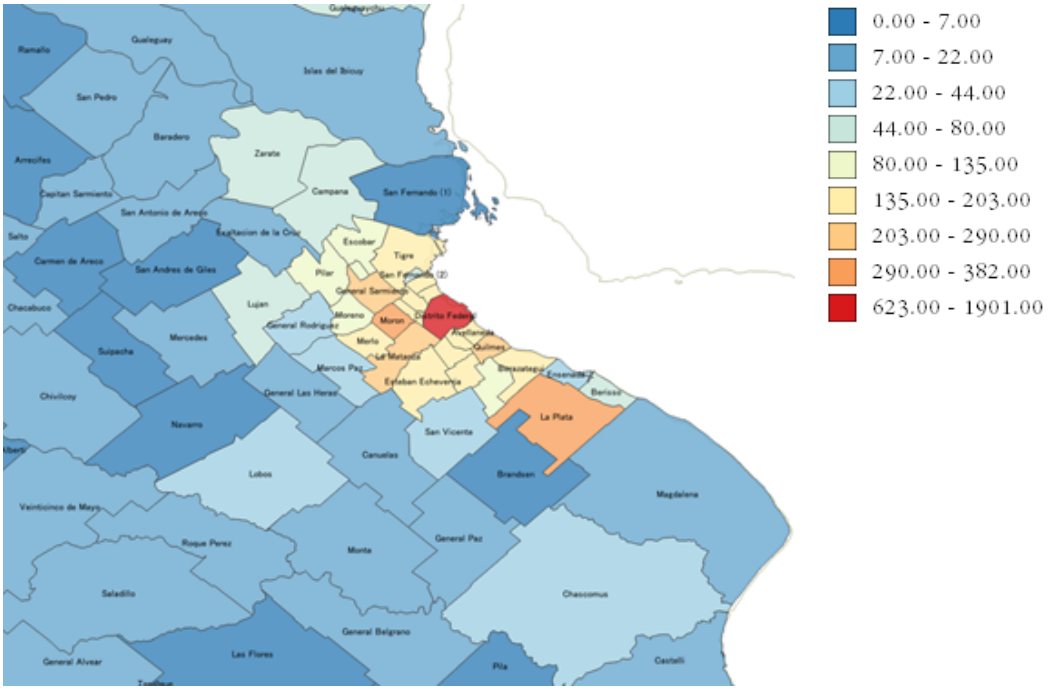


Figure 12. Thematic map of non-standard forms in the province of Buenos Aires

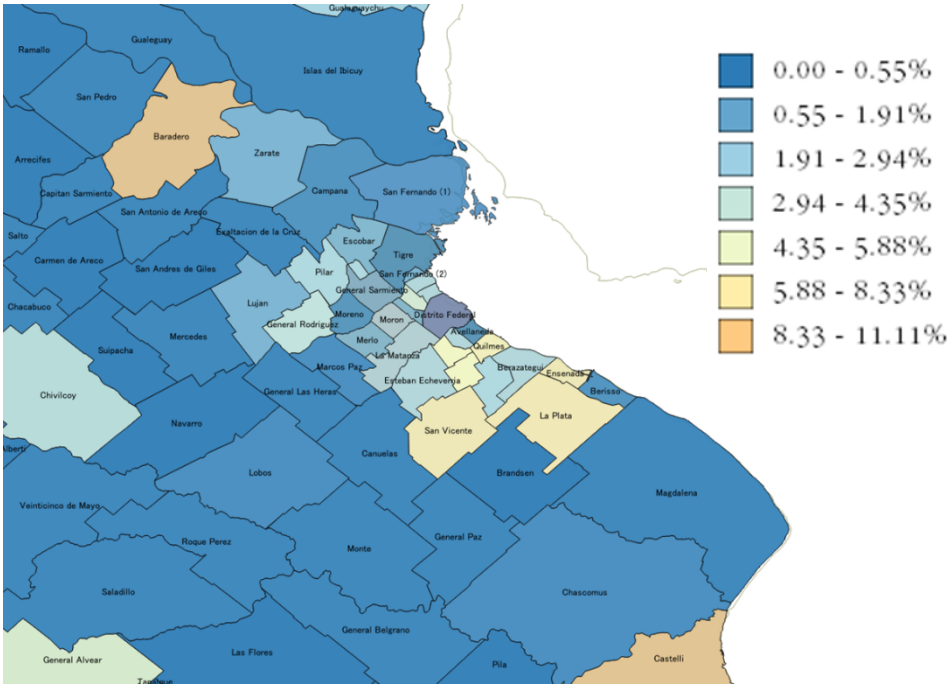


Figure 13. Thematic map of the ratios of non-standard forms in the province of Buenos Aires

Although Lipski (1994) said that the non-standard form was more acceptable, at present, with data of different nature such as tweets, such acceptance is not observed,

since the ratios are relatively low with exceptions such as Baradero, La Plata, San Vicente, etc. The data also suggests that the ratio is especially low in the Federal District, the most urban area of the province. In Barnes (2012), using data from oral transcription of the *Corpus de Referencia del Español Actual* (CREA), [9] the *Habla Popular*[10] by Lope Blanch (1976) and the *Corpus del Español* by Mark Davies, the ratios oscillate between 13% and 16%, whereas in the geocorpus, which is very different from this last study, the ratios vary between 0.77% in the Dominican Republic and 18.45% in Puerto Rico. It should not be forgotten that there are several areas in which data are scarce and therefore they are not significant.

7. Relationship of form, ratio and frequency

According to the theory of the addition of *-s* by analogy with other verb tenses in which a proportion of speakers uses the non-standard form, the relation between the frequency of the two forms and their ratios should be similar for each verb and for each geographical area. Figure 14 shows a graph with the linear regression of the standard form with the variant ending in *-s*. The points correspond to different verbs.

[9] http://corpus.rae.es/creanet.html
[10] http://www.iifilologicas.unam.mx/elhablamexico/index.php?page=habla-popular

Dialectologia. Special issue, VII *(2017), 149-167.*
ISSN: 2013-2247

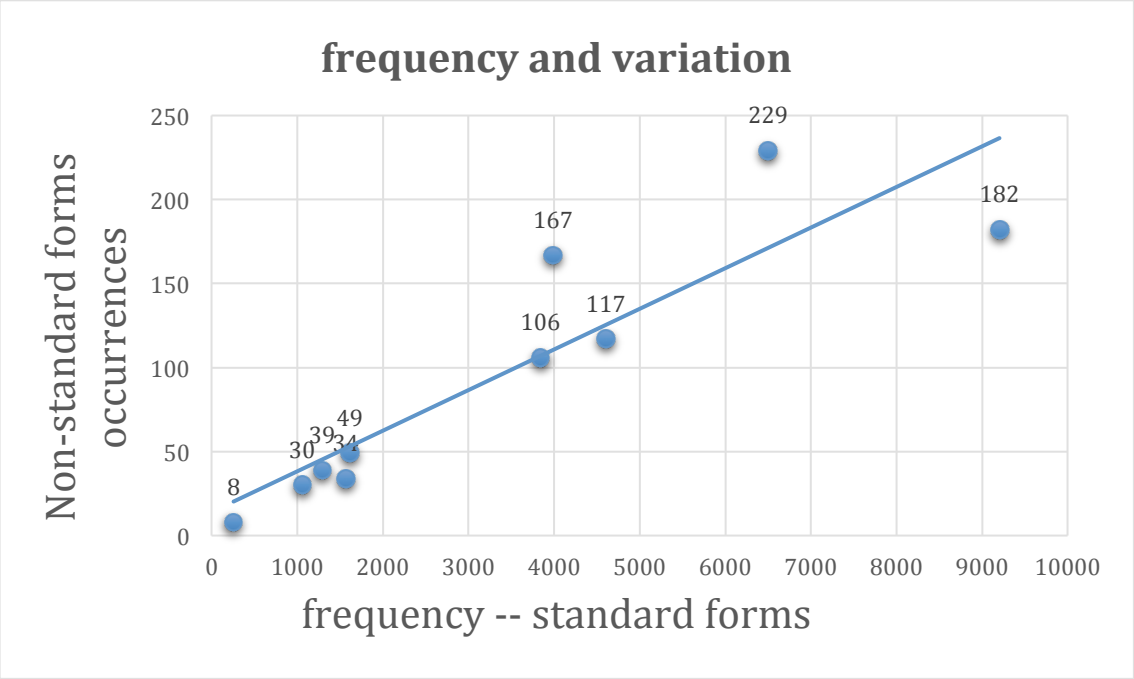


Figure 14. Linear regression between the frequency of the standard and non-standard forms

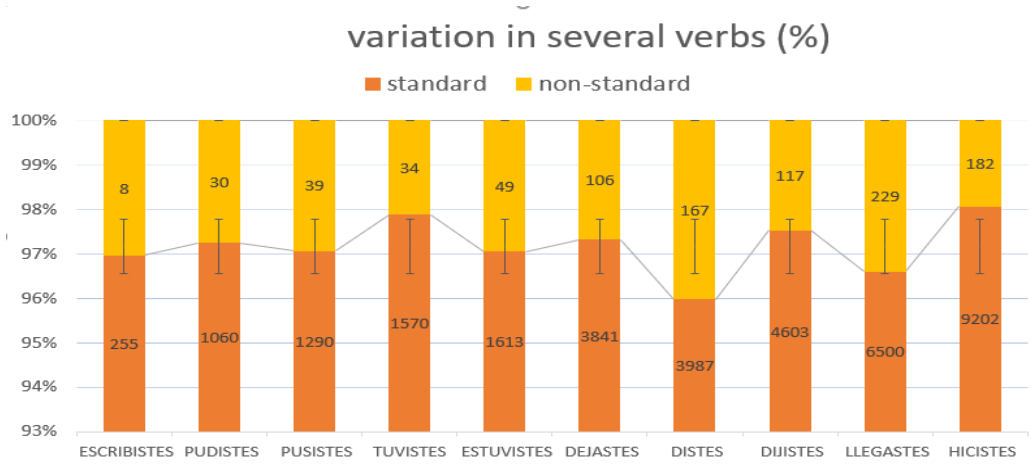


Figure 15. Variation of the ratios of each verb between the two forms

The graphs in Figures 14 and 15 suggest that, despite possible differences in regional varieties, the analogy hypothesis is reasonable to some extent, although some ratio differences also suggest that other unknown factors can intervene. They could be related to the coexistence of *voseo* forms, the degree of stigmatization in different areas, etc.

8. Conclusions

The *-s* added to the 2nd person of the singular of the perfect past tense is a phenomenon spread throughout the geography of Spanish with different ratios regarding the standard form. This phenomenon appears more frequently in rural areas, even in *voseo* areas. The ratios by geographic zones are different, but within limits that are not too extreme. In addition, the ratios of such forms are also variable according to the verb.

The use of geocorpora of tweets for the study of linguistic variation adds additional information to traditional studies based on surveys, etc. On the other hand, geocoded data allows the quick and accurate preparation of synchronous variants maps using GIS technology.

References

BARNES, Sonia (2012) "¿Qué dijistes?: A Variationist Reanalysis of Non-Standard -s on Second Singular Preterit Verb Forms in Spanish", *Selected Proceedings of the 2010 Hispanic Linguistics Symposium*, Somerville, MA: Cascadilla Press, 38-47.

DAVIES, Mark (2016-) *Corpus del Español: Two billion words, 21 countries*. Available online at http://www.corpusdelespanol.org/web-dial/. (Web / Dialects)

DÍAZ COLLAZOS, Ana María (2015) *Desarrollo sociolingüístico del voseo en la región andina de Colombia (1555-1976)*, Berlín: Mouton de Gruyter.

FRAGO GRACIA, Juan Antonio (2003) *El Español de América*, Cádiz: Universidad de Cádiz.

KANY, Charles E. (1945) *American-Spanish Syntax*, 2nd ed. Chicago: University of Chicago Press.

LIPSKI, John M. (1994) *Latin American Spanish*, London and New York: Longman.

PENNY, Ralph (2002) *A history of the Spanish language*, Cambridge: Cambridge University Press.

PARODI, Claudia (2015) "Dialectos del español de América: México y Centroamérica", in Javier Gutiérrez Rexach (coord.), *Enciclopedia de Lingüística Hispánica*, Vol. 2, London: Routledge, 375-386.

REAL ACADEMIA ESPAÑOLA (2010) *Nueva gramática de la lengua española. Manual*, Madrid: Espasa Libros, 2010.

VAQUERO DE RAMÍREZ, María (1998) *El español de América II: Morfosintaxis y léxico,* Madrid: Arco Libros.