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

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.
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, which is approximately 1% of all tweets that can be obtained freely.

b. Only geocoded tweets, 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, programmed in PHP by Adam Green has been used.

e. The tweets are arranged in a MySQL database, 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) 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’.

---

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/
Although the Corpus of Spanish is lemmatized, it does not distinguish non-standard forms as *hiciste* ‘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.

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

Figure 2. Frequencies of the form *hicistes* ‘you did’ in the Corpus of Spanish (Web-Dialects)
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](https://www.adminer.org/)

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.

**Figure 4.** Partial data of diste’s ‘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.

<table>
<thead>
<tr>
<th>VERB FORM</th>
<th>without -s</th>
<th>with -s</th>
<th>ratio %</th>
</tr>
</thead>
<tbody>
<tr>
<td>estuviste/s</td>
<td>1613</td>
<td>49</td>
<td>3.03</td>
</tr>
<tr>
<td>pudiste/s</td>
<td>1060</td>
<td>30</td>
<td>2.83</td>
</tr>
<tr>
<td>escribiste/s</td>
<td>255</td>
<td>8</td>
<td>3.13</td>
</tr>
<tr>
<td>pusiste/s</td>
<td>1290</td>
<td>39</td>
<td>3.02</td>
</tr>
<tr>
<td>llegaste/s</td>
<td>6500</td>
<td>229</td>
<td>3.52</td>
</tr>
</tbody>
</table>
The distribution of occurrences by country is shown in Table 2.

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

<table>
<thead>
<tr>
<th>VERB FORM</th>
<th>without -s</th>
<th>with -s</th>
<th>ratio %</th>
</tr>
</thead>
<tbody>
<tr>
<td>dejaste/s</td>
<td>3841</td>
<td>106</td>
<td>2.75</td>
</tr>
<tr>
<td>tuviste/s</td>
<td>1570</td>
<td>34</td>
<td>2.16</td>
</tr>
<tr>
<td>diste/s</td>
<td>3987</td>
<td>167</td>
<td>4.18</td>
</tr>
<tr>
<td>dijiste/s</td>
<td>4603</td>
<td>117</td>
<td>2.54</td>
</tr>
<tr>
<td>hiciste/s</td>
<td>9202</td>
<td>182</td>
<td>1.97</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>33921</strong></td>
<td><strong>961</strong></td>
<td><strong>2.83</strong></td>
</tr>
</tbody>
</table>

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.
6. Visualization of the data

For the visualization of the data, QGIS\textsuperscript{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.
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 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
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), the Habla Popular 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
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


