Preprint version. Please cite original version:

Vaamonde, Gael (2025): «Using "Small and Tidy" Historical Corpora to Explore Linguistic Variation in Early Modern Spanish: New Possibilities in the Paradigm of Digital Humanities». In Roberto J. González Zalacain and Gael Vaamonde (Eds.): *Digital Humanities in Medieval and Early Modern Spanish Texts: Current Perspectives and Approaches*, Routledge, pp. 23-51. DOI: https://doi.org/10.4324/9781003393771-3.

Using "small and tidy" historical corpora to explore linguistic variation in Early Modern Spanish: New possibilities in the paradigm of Digital Humanities

Gael Vaamonde
University of Granada
https://orcid.org/0000-0001-8360-2805

Abstract

This work presents the main characteristics of two specialised historical corpora: the P.S. Post Scriptum corpus, composed of private letters, and the Oralia Diacrónica del Español (ODE), comprising inventories of goods, witness depositions and medical certificates. In particular, it is shown that the decisions made in the different stages of the corpus design and compilation mean they are especially appropriate digital resources for the study of linguistic variation in Early Modern Spanish. To this end, three case studies are analysed that address dialectical variation phenomena belonging to different levels of language analysis: *laísmo* (i.e. the use of *la* and *las* as indirect object), as a morphosyntactic phenomenon; the use of the suffixes *-ico* and *-ito* as a morphological phenomenon, and *seseo* as a phonetic phenomenon. The ultimate aim of this work is to draw attention to the possibilities offered by these small and tidy corpora for research into historical Spanish linguistics.

Keywords

corpus linguistics, historical linguistics, digital humanities, seseo, laísmo, diminutives

1. Introduction

Although corpus linguistics predates the arrival of computers, its current development is inconceivable without the existence of the electronic format. The relevance of the digital world in the development of this discipline is revealed in the terminology used by corpus linguists themselves when tracing its history. Mayer (2008) speaks of *pre-electronic corpora* in order to refer to the work involved in the compiling of texts prior to the creation of the first computerised corpus in the 1960s; Francis (1992) had already proposed the ingenious expression of *Language corpora B.C.* (i.e. *Before Computers*) to refer to this same era; and Leech (1992: 106) dubs the period initiated from this point as *computer corpus linguistics*. This dependence on computers is also revealed in the definition of corpus we tend to find in manuals, given that one of its essential features is precisely the fact it is stored on a digital device (McEnery *et al*. 2006: 5, Gries 2009: 7; Rojo 2021: 1). In short, and as stated by Jenssen (2014: 115), "if there is one discipline within the humanities that has embraced the digital nearly since the inception of modern computer science, it must be corpus linguistics".

Computer technology has allowed us to create large corpora and afforded us the necessary tools to process and extract their content rapidly and effectively. Nevertheless, beyond the common denominator that is the electronic format, there are different meanings as regards the way in which linguistic corpora are constructed and used in the present day. For a number of years now the process of constructing a corpus has moved between two clearly opposing extremes. At one end are those who prefer a quantitative perspective and, as a result, design large linguistic corpora, generally taking advantage of the enormous amount of information available online. At the other extreme we can place those corpora that are reduced in size, which normally involves a creation process that prioritises a qualitative perspective resulting in meticulous treatment of the compiled data.

The advantages and drawbacks presented by both perspectives were summarised by Mair (2006) in two expressions frequently used when speaking about linguistic corpus design: *big and messy* and *small and tidy*. Thus, the advantage of big and messy corpora is their large size; however, being products based solely on data taken from the web they show problems of representativeness and lack specific metadata relevant for language research. In contrast, the reduced size of small and tidy corpora normally translates into sparse typological variety and difficulty in obtaining examples of infrequent linguistic phenomena; however, this type of corpora generally offers high codification as well as a greater control of metadata, facilitating highly selective searches.

In the case of contemporary Spanish, examples of big and messy corpora include esTenTen18[1], with close to 20 billion words, *NOW*[2], comprising 7 billion, and CdEweb[3], which stands at 2 billion. In contrast, we can cite e-resources such as BDS[4], ADESSE[5], SenSem[6], and AnCora-ES[7]. These corpora, with sizes varying between half a million and a million and a half words, have been manually annotated with detailed syntactic and semantic data, and permit the retrieval of a variety of information on Spanish verbs, syntactic constructions and valency alternations. Lastly, at an intermediate point between these two extremes we have the so-called reference corpora, such as CREA[8] —160 million words— and CORPES XXI[9] —350 million words. Due to their very nature, reference corpora cannot reach the degree of detail and revision of small corpora, but entail a much greater size; nor do they reach the volume of data of massive corpora, but are designed to guarantee a number of standards relating to representativeness and balance lacked by the latter (Rojo 2021: 81).

If we turn to the sphere of historical linguistics, and without moving away from the Spanish corpora, a number of important points should be made as regards the aforementioned situation. The first is that no massive historical corpora exist in the case of Spanish; in other words, there are no examples similar in size and design to those big and messy corpora cited above for contemporary Spanish. Therefore, the opposition is reduced to that established by small versus reference corpora[10]. The second important point to make is working with historical texts and manuscripts puts compilers in an additional quandary: they must choose between resorting to available editions or creating their own from original sources. This decision, related to the choice between quantity or quality, opens another important difference between reference and small and tidy corpora to be established: only the latter can include their own palaeographic

[1] This is a one of the many corpora available in Sketch Engine (https://www.sketchengine.eu/).
[2] Available in https://www.corpusdelespanol.org/now/.
[3] Available in https://www.corpusdelespanol.org/web-dial/.
[4] Available in https://www.bds.usc.es/.
[5] Available in http://adesse.uvigo.es/.
[6] Available in http://grial.edu.es/sensem/corpus/main.
[7] Available in https://clic.ub.edu/corpus/es.
[8] Available in https://www.rae.es/banco-de-datos/crea.
[9] Available in https://www.rae.es/banco-de-datos/corpes-xxi.
[10] Of course, it is always possible to use the internet as a data source for studying diachronic linguistic phenomena. By way of example, see Octavio de Toledo y Huerta (2016).

editions based on unique criteria, whereas the former are mainly compiled from modern editions published online, despite this not being the best option for linguistic enquiry (Dollinger 2004: 6-7, Claridge 2008: 250-251, Honkapohja *et al*, 2009: 456-460).

We currently dispose of at least three historical reference corpora for Spanish, which are CORDE[11] —250 million words—, CDH[12] —350 million— and CdEhist[13] —100 million—. Along with these, in recent years a number of specialised historical corpora have been published that complement the foregoing, and which boast transcripts with greater philological detail and greater homogeneity, among other advantages deriving from their reduced size. Worthy of mention, for example, are Biblia Medieval[14], CODEA+ 2022[15], CORDIAM[16], CORDICan[17], CorLexIn[18], OSTA[19] and Panépica Digital[20], amongst others. The size of these corpora varies considerably, oscillating between the 50,000 words of Panépica Digital and the over 30 million forms contained in OSTA.

The aim of this chapter is to show the advantages of these latest resources for linguistic research by examining two small and tidy corpora of Early Modern Spanish: P. S. Post Scriptum. A Digital Archive of Ordinary Writing[21] (henceforth, PS) and Oralia Diacrónica del Español[22] (henceforth, ODE). The chapter is divided into two main parts. Firstly, the main features of these two corpora will be briefly described. Secondly, their potentialities as data sources for linguistic variation and historical dialectology will be illustrated through the analysis of three well-known phenomena of Spanish variation: (i) case variation in unstressed third person pronouns and, particularly, the distribution of laísmo, (ii) the use of diminutive suffixes and, particularly, the distribution of *-ito* and *-ico*, and (iii) the phonological opposition between /θ/ and /s/ and, particularly, the distribution of seseo. These three cases studies cover morphosyntactic, morphological and phonological issues, respectively. Generally speaking, the aim of this contribution is to prove how PS and ODE, being small and tidy corpora, are particularly useful to explore non-standard historical text material —marginally represented in reference corpora—, to identify diachronic regional varieties and to combine quantitative methods with close contextual analysis.

2. PS and ODE corpora. General description and distinguishing features

The PS corpus (see Vaamonde 2018) was created in the Center of Linguistics of the University of Lisbon between 2012 and 2017. PS comprises around 5000 private letters written in Spanish and Portuguese during the early modern period. The Spanish part of the corpus, which this work uses, is made up of 2447 letters, which is the equivalent of approximately one million words. The dates of the Spanish subcorpus range from 1510 to 1833.

The majority of these letters are unpublished and were conserved up to the present day archived within legal proceedings, given that they were used as evidence in deciding on the guilt or

[11] Available in https://www.rae.es/banco-de-datos/corde.
[12] Available in https://www.rae.es/banco-de-datos/cdh.
[13] Available in https://www.corpusdelespanol.org/hist-gen/.
[14] Available in https://www.bibliamedieval.es/BM/index.php.
[15] Available in https://www.corpuscodea.es/.
[16] Available in https://www.cordiam.org/.
[17] Available in https://www.ull.es/corpora/cordican.
[18] Available in https://corlexin.unileon.es/.
[19] Available in http://www.hispanicseminary.org/osta-es.htm.
[20] Available in http://corpora.ugr.es/cid/.
[21] Available in http://ps.clul.ul.pt.
[22] Available in http://corpora.ugr.es/ode.

innocence of an accused party being tried for a specific crime. Their authors, therefore, are individuals from very different social backgrounds. They are quite frequently authors with poor literacy or "mãos inábeis" (Marquilhas 2000: 235), who tend to put everyday matters into writing through a poorly elaborated discourse, produced relatively spontaneously, and in which what is said is normally more important than how it is said. This particular feature makes PS an especially appropriate digital resource for researching non-standardised linguistic phenomena from a diachronic point of view.

The ODE corpus (see Calderón Campos & Vaamonde 2020) is being developed by the DiLEs group (Diacronía de la Lengua Española) at the University of Granada. At the time of writing, ODE comprises 527 documents produced during the early modern period, in a time frame spanning 1509 to 1896. The size of the corpus currently stands at 700,000 words, although this figure increases regularly.

ODE comprises three text types that share the peculiarity of including linguistic material belonging to the oral plane and the sphere of the everyday: witness depositions made over the course of legal proceedings, inventories in which lists of objects are registered for different purposes, and medical certificates which, inserted into criminal proceedings, describe in great detail injuries suffered by victims of assault. The interest in the depositions is related, above all, to the fragments in direct style that appear sporadically throughout the text, when it is necessary to literally transcribe the words pronounced by a witness. Furthermore, inventories of goods and medical certificates constitute historical sources of great value for analysing the lexicon in the past, particularly in reference to everyday objects and anatomical designations. The three text types represented in ODE involve the transfer to paper of oral statements of witnesses, evaluators and surgeons, to which it is unsurprising for linguistic traits of the vernacular of the declarants or scribe to find their way into the documents.

	PS (Spanish)	**ODE**
Documents	2447	527
Words	1,000,000	700,000
Typology	private letters	witness depositions inventories of goods medical certificates
Range of dates	1510 - 1833	1509 - 1896
Geographical area	practically entire peninsula	mainly centre and south of Spain

Table 1. Main characteristics of PS and ODE

Both corpora compile, ultimately, texts characterised by their communicative immediacy (Koch & Oesterreicher 2007), that is, discursive practices which reflect oral features and vernacular uses. It is worth remembering that these types of historical sources do not usually fit into historical reference corpora, given they are not easy to find: their compiling requires a painstaking search and selection effort in different archival fonds. And, despite this, these sources are especially important for the diachronic study of the language, which must be inevitably based on written texts:

> La inmensa mayoría de los textos escritos con que el lingüista histórico ha de enfrentarse son claramente unidimensionales en este sentido: provienen de un sector de la comunidad, el 'superior', y suelen manifestar un lenguaje cuidado, elaborado, en el que, ciertamente, se desarrollan mucho más que en otros ámbitos las potencialidades del idioma, pero en el que suele predominar también una actitud conservadora, reacia a las modificaciones (Cano Aguilar 1996: 375).

Along with the communicative immediacy of their sources, PS and ODE share other no less important characteristics for the diachronic study of language and which derive from their condition of small and tidy corpora. In this work it is relevant to highlight at least three: firstly, the possibility of including metadata that include useful information for the researcher; secondly, the creation of digital editions that respect both the original spelling of the text and diverse palaeographic aspects of the document in its entirety; and thirdly, the methodology applied for coding, processing and retrieving text data from historical handwritten sources.

In regards to the first aspect, the types of sources to which these corpora pay attention permits certain metadata of undoubted linguistic interest to be obtained in quite a reliable manner, amongst which attention should be drawn to dating and geographical information. In the case of PS, the missives themselves generally include date and place of writing. It is convenient to remember, furthermore, that these letters have not been located as isolated pieces; rather, they were archived within the legal proceedings they accompanied as evidence, and have been conserved as such up to our times. Also, many of these legal proceedings include the questioning of different witnesses and defendants made over the course of the action, and their consultation has often permitted information regarding the geographical origins of the authors of the missives to be incorporated into the corpus. In other words, we have access to the place the letter was created and the place of origin of its author, which may or may not coincide. The second piece of information is, in any event, essential for designing historical maps of a specific language phenomenon[23].

In the case of ODE, all of the documents are dated and generated in a specific place, so this information is clearly recorded on the manuscript. However, in contrast to the epistolary documentation, which generally guarantees a direct relationship between content and metadata, the documentation stored on ODE makes it necessary establish a relationship that is indirect, but one that is equally valid in terms of dialectal exploitation. Depositions, inventories and medical certificates are the result of a very similar transfer of information: the declarant —or evaluator— speaks and the scribe takes down what is heard in writing. Although it is true that we do not have access to the biographical information of any of these actors, the geographical origin of the declarants and the scribes who produced these texts had to be very close to that of the places where they were written. We can put forward two arguments that clearly support this idea. On the one hand, and from a linguistic perspective, the results obtained on different phenomena from this type of documentation are on the whole coherent with the contemporary dialectal panorama, which is an indicator of the relationship between the place the document was created and the place of origin of the scribe (Calderón Campos 2019: 119). On the other hand, and from a historical point of view, the influence of family relationships in the transfer of public notary positions has been confirmed, in the sense that it was possible for them to be handed down from father to son (Mendoza García 2011), which reinforces this theory of dialectical belonging.

In regards to the second aspect, these electronic resources are not limited to offering a linguistic corpus; rather, they also include scholarly digital editions of their documents, along with the corresponding facsimile images thereof. They therefore combine a linguistic interest with a philological interest. Given in all cases it involves cases of unpublished documentation, these editions are created from scratch and based on consistent editorial principles. The starting

[23] Regarding this relationship between the letter and the diatopic variation of its author, there are the problematic cases of copied letters, that is non-original ones —marginals in PS— and delegated writing, which a mental author dictates for another hand to write. Both subtypes are duly marked, so these texts can be easily excluded from results in the search interface.

premise for the construction of both corpora is the creation of palaeographic editions; in other words, transcriptions designed to achieve a faithful representation of the source: the original spelling is retained and various particular details on layout and appearance of the text in the manuscript are reflected, such as additions, deletions, abbreviations or omissions. From the technical perspective, and in accordance with common practices in the field of digital humanities, transcriptions have been carried out in XML, adopting the standard proposed by the TEI (Text Encoding Initiative) consortium for the representation of texts in digital form.

There is no doubt that the adoption of this philological dimension, which brings with it the practice of a scrupulous transcription of the texts in XML-TEI along with the inclusion of the facsimile, requires considerable time and effort. Accordingly, it is only feasible in the building of small corpora such as those that concern us here. In contrast, it should be pointed out that this task has a clear benefit for the linguist. Access to the facsimile publication permits any aspect of the transcript to be contrasted with the original document and gives the researcher a guarantee of the authenticity of the information. Homogeneity in editorial principles ensures a greater quality thereof, as it avoids the appearance of discrepancies in the results of a query. Remaining faithful to the original spelling increases the possibilities of exploitation of the corpus, as it is only in this way that it can be queried effectively on phonetic phenomena. Lastly, retaining how the source text is rendered or presented provides the corpus philological rigour, increases search options and avoids linguistic misinterpretations (Janssen & Vaamonde 2020: 277-280).

Regarding the third aspect, both corpora make use of the TEITOK web-based platform (Janssen 2016). This online resource incorporates different natural language processing tools, which allow the compiler to automate corpus annotation tasks. In this manner, PS and ODE do not just rely on the palaeographic transcriptions of the texts; they also offer different levels of annotation together with the original word forms, such as regularised spelling forms, grammatical tags and lemmas (Vaamonde 2018). By grammatical (or part-of-speech) tagging we mean the task of assigning tags that contain morphosyntactic information on each word in the corpus. In PS and ODE, the tagset used is based on the proposal by the EAGLES group for the morphosyntactic annotating of lexicons and corpora for all European languages (Leech & Wilson 1996). According to the EAGLES recommendations, each tag comprises a sequence of letters and numbers, where each letter or number represents a specific morphosyntactic feature depending on its position within the sequence. Thus, for example, the form *aborrecen* is assigned the tag VMIP3P0, which stands for verb (V), main (M), indicative (I), present (P), third person (3), plural (P); and the word *vecinas* is assigned the tag NCFP000, which indicates noun (N), common (C), feminine (F), plural (P). The output of the automatic tagger has been manually revised for both corpora by their corresponding team of linguists, which ensures a very low error rate.

In regards to information retrieval, the TEITOK platform permits these corpora to be queried using the CQP (Corpus Query Processor) syntax (Evert & Hardie 2011). This powerful language, which admits the use of regular expressions, is designed for carrying out sophisticated searches on the different levels of annotation presented by a corpus, including possible metadata. It also admits both queries relating to individual words and those regarding lexical and grammatical patterns of greater complexity.

In short, PS and ODE comprise texts close to conceptional orality and the vernacular varieties of the early modern period. This thorough search for orality within written documents marks a difference compared to historical reference corpora in Spanish, which are known to be characterised by a scarce representation of spoken language (Rodríguez Puente 2016). Moreover, both corpora use handwritten sources that are unequivocally dated and that, in the majority of cases, can be assigned to a specific geographical location. The consideration of these metadata opens up the possibility of carrying out dialectical approaches on determined phenomena in different centuries and outline historical-linguistic maps, as we shall see in this work. In relation to the technical aspect, both corpora coincide in combining methods and standards that apply within digital humanities and corpus linguistics: they use XML-TEI for the digital edition of primary sources, the EAGLES standard for the grammatical tagging of the texts and the CQP syntax for data retrieval. The integration of these languages via the TEITOK platform allows the user to carry out selective searches of differing complexity, both in the digital edition and the annotated corpus.

In order to show the usefulness of these two corpora, we have selected three linguistic phenomena of different significance: *laísmo* (use of *la* as an indirect object), as a morphosyntactic phenomenon; the use of the diminutives *-ito* and *-ico*, as a morphological phenomenon, and *seseo* (use of the phoneme /s/ for graphemes <ce>,<ci>, <z> and <s>), as a phonetic phenomenon. We should point out that any one of these three phenomena reveal more complex dialectal panorama than it is possible to address in this study. This work, necessarily superficial, is aimed at highlighting the advantages of PS and ODE for the diachronic research of Spanish and, in particular, for the analysis of phenomena specific to the vernacular and subject to dialectical variation, such as those we are concerned with. We dedicate the following paragraphs to their description, data filtering and discussion. The reader may, however, find a specialised bibliography that will permit each of the case studies selected to be looked at in-depth.

3. Laísmo

3.1. Description of the phenomenon

Studies on grammatical variations of Spanish have paid special attention to the unstressed series of third person personal pronouns *le(s), la(s), lo(s)*. This paradigm leads to a difference between so called canonical uses —also known as etymological or distinguishing uses— and innovative uses —also known as referential or confounding uses— of pronouns. The former are those employments adjusted to the inherited Latin canon: accusative forms *la(s), lo(s)* for the direct object and dative forms *le(s)* for the indirect object. The latter are those uses that do not follow the syntactic function of the model and give rise to phenomena of leísmo (2a), laísmo (2b) and loísmo (2c) (NGLE 2009: §16.8). The following examples, taken from PS, serve to illustrate these usage variants[24]:

(1) **Canonical uses**:
a. vuelvo a pedir que le dé a ese hombre el dinero (PSCR5407)
b. a ella la han encontrado con la capa de su hijo arrebujada (PS5055)
c. y Bartolomé Romero, vuestro tío, que lo mató un macho de una coz (PS6020)

[24] Save where the opposite is indicated, the examples are transcribed with expanded abbreviations and normalised spelling. We accompany each example of the corresponding code to the document from which it has been taken, which may be used to retrieve the example from the search interface of each corpus. The codes used in PS always begin with PS or PSCR and four random digits. The codes used in ODE are composed of eleven characters that provide information on the place of creation, year, text type and four random digits; for example, AL1704I0001 means Almería, 1704, Inventory, 0001.

(2) **Innovative uses:**
a. y le mataré como a un perro (PSCR6636)
b. y si tuviera el honor de verla la daría un abrazo (PSCR6885)
c. yo en persona me veré con ustedes o los escribiré un papel (PS5019)

In this work we will limit ourselves to analysing laísmo, that is, the use of *la(s)* instead of canonical *le(s),* paying special attention to its geographical distribution during the early modern period. For a general overview of the variation, see Fernández Ordóñez (1999) and Gómez Seibane (2012). Amongst the studies that have addressed unstressed pronouns from a historical perspective, mention should be made of the seminal works by Lapesa (2000 [1968]) and Echenique Elizondo (1981). More recent studies can be found in Fernández Ordóñez (2001), Flores Cervantes (2006), Sáez Rivera (2008), Gómez Seibane (2013), Vaamonde (2015) and Sánchez-Prieto Borja & Vázquez Balonga (2019).

3.2. Data collection

The retrieval of the total group of occurrences of laísmo documented in a corpus is not an easy task, as it demands certain starting conditions. This obviously implies that the corpus used be tagged with some type of morphological annotation; otherwise, our query will be ineffective, given that the search for the *la* or *las* character sequence will return both those cases of the personal pronoun *la(s)* and those of the definite article *la(s)*. Annotated corpora permit, in principle, this type of information to be disambiguated. Nevertheless, it is known that automatically tagged corpora contain errors and must be post-edited by linguists, but reference corpora, due to their size, cannot be manually revised. The case of the sequence *la(s)* is particularly noteworthy in this regard, given that both forms —determiner and pronoun— are extremely frequent, which complicates their disambiguation. If we search for the pronoun *la* in CdEhist or CDH between the 16th and 19th centuries, we will obtain an overwhelming group of results (at the moment, almost 50,000 cases in CdEhist and almost 4,000,000 cases in CDH), but also a high number of false positives, in other words, of examples corresponding to the definite article *la*. The problem, in any event, is even more complicated and goes beyond categorial disambiguation, given that the cases we really need are only those in which the pronoun *la(s)* has the syntactic function of indirect object.

This is where small and tidy corpora are useful, as their annotation is normally revised and, where appropriate, widened to cover specific linguistic phenomena. In PS and ODE, which use the EAGLES standard, the definite article *la* is assigned the tag DA0FS0 (Determiner, Article, Feminine, Singular), whereas the pronoun *la* corresponds to the tag PP3FSA00 (Pronoun, Personal, 3rd person, Feminine, Singular, Accusative). In both corpora, furthermore, the sixth position of the tag corresponding to the pronoun is reserved for differentiating the employment of *la(s)* as a direct object (A for Accusative) from its employment as an indirect object (D for dative). We are thus able to retrieve all cases of laísmo documented in these corpora with a single query. The CQP syntax necessary to retrieve these cases is that which we offer below:

```
In PS:        [pos="PP3F(S|P)D00" & nform="las?"] :: match.text_lang = "ES"
In ODE:       [pos="PP3F(S|P)D00" & nform="las?"]
```

Given that PS includes two subcorpora, Spanish and Portuguese, it is necessary to specify in which of them the query is to be applied. For everything else, the syntax is identical in both cases. The *pos* (part of speech) attribute is intended to define of the morphological tag, which in the previous query has been specified with the value of 3rd person feminine personal pronoun, singular or plural, and in regards to indirect object. The attribute *nform* (normalized form) is used to specify the form with normalised spelling: *las* in the above query, with the optional final *s* to guarantee to ourselves that we will obtain both the singular and plural cases.

The results we obtained, and which are comparable from the search interfaces of these corpora, are as follows: 274 cases in PS and 10 cases in ODE. As expected, PS is considerably more productive for obtaining examples of a pronominal variety. The nature of their sources, especially suitable for documenting morphosyntactic phenomena, and their greater geographical range, especially in the centre and north of the peninsular, explain this disparity. Thus, we shall start out from the data offered by PS to analyse the geographical distribution of laísmo in the early modern period.

As we have already mentioned above, the Spanish PS corpus comprises 2447 letters. Nonetheless, we have excluded from this group those copies and letters that are the result of delegated writing, in order to leave ourselves solely with missives that constitute original documents written by the hand of the signee. In this way, we ensure that we have a greater reliability in the sources to be analysed from the dialectal point of view. Our starting point, on applying this filter, is a corpus comprising 2380 letters. As we have pointed out, 274 occurrences of laísmo have been identified in this group, distributed amongst 131 missives and 89 different authors. The complete figures from which we start are summarised in Table 2:

	Ocurrences	**Letters**	**Authors**
Non laísmo	960688	2249	978
Laísmo	274	131	89
Total	960962	2380	1067

Table 2. Distribution of laísmo in Post Scriptum

We should draw attention to the fact that PS does not provide information regarding the geographical origin of all of the authors that make up the corpus, but it does do so for a considerable majority. Specifically, this type of information is included for a total of 873 authors. PS catalogues this information in two ways: place of birth and place of residence. For the purpose of clarity, in our analysis we have selected a single geographical origin per author, proceeding as follows to do so: if both places are known for a specific author, place of birth takes priority over place of residence; if only one is known, it is this which has been taken into account. Following this procedure, we have been able to connect 77 of the 89 laísta authors in our corpus to a specific territory in the peninsular. Finally, we have associated these 77 places with their corresponding geographical coordinates and imported the data to the QGIS 3.10 geographic information system in order to project them onto a map. The result is shown in Figure 1, which we discuss in the following section.

3.3. Results and analysis

The current situation regarding the employment of the pronouns *le(s)*, *la(s)* and *lo(s)* permits two clearly differentiated areas to be established: a distinguising zone, that is, subject to etymological uses, versus an confounding zone, in which non-etymological uses appear —including laísmo— and "que comprende los territorios del occidente y centro de Castilla situados al sur de la cordillera cantábrica hasta alcanzar La Mancha" (Fernández Ordóñez 1999: 1363). Basing ourselves on dialectical studies on contemporary data, today all or almost all of the territories included in the provinces of Ávila, Burgos, Madrid, Palencia, Toledo and Valladolid would be laísta areas. The outermost part of this isogloss reveals zones of transition where different solutions compete, although cases of laísmo are equally documented in bordering zones of Leon, Zamora, Salamanca, Caceres, Ciudad Real and La Rioja (Fernández Ordóñez 1999: 1364, Gómez Seibane 2012: 30). We shall take this information as a reference in order to contrast it with the historical data extracted from PS.

The map we show in Figure 1 georeferences the 77 authors of known geographical origin from which one or more cases of laísmo have been documented in the corpus. Note that in the map only 59 points out of 77 are visualised, due to the fact that the geographic information of some authors is identical and therefore overlaps: we document five authors from Toledo, five from Valladolid, five from Pontevedra, three from Madrid and two from Cuenca, Seville, Algete (Madrid), Madridejos (Toledo), Novés (Toledo) and Talavera de la Reina (Toledo). The places that coincide, approximately, with areas that are today laísta appear marked in black, whereas the places that are removed from the current confounding zone are represented in red.

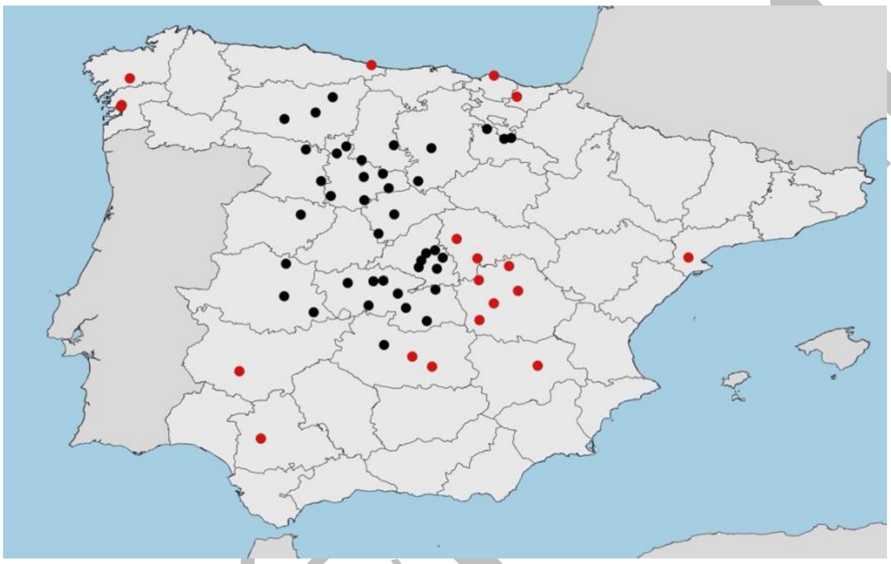

Figure 1. Geographical distribution of laísta authors in PS.

The results we have obtained reveal at least three interesting aspects, which we shall briefly discuss. Firstly, attention should be drawn to the fact that a large section of laísta authors documented in PS originate from territories that are laísta in the present day. Specifically, 54 of the 77 georeferenced authors originate from the current laísta zone. In other words, the dialectical situation of laísmo that our data from the early modern period reveal coincide to a significant extent with the current distribution of this phenomenon in peninsular territory. The provinces of Madrid, Toledo and Valladolid stand out from the others, given there are 35 laísta authors between the three.

	Laísta zone	Non laísta zone	Total
Laísta authors	54	23	77
Non laísta authors	206	590	796
Total	260	613	873

Table 3. Laísta and non-laísta authors in PS according to geographical zone.
$\chi2$ = 63.64, df = 1, p < 0.001, phi = 0.269

Secondly, we identify a not inconsiderable number of laísta authors —up to 11— originating from Castilla la Nueva, and mainly distributed throughout the provinces of Cuenca and Guadalajara. These examples of laísmo are documented in a time frame spanning 1689 to 1804;

that is, they are produced around or during the 18th century. It is important to draw attention to this datum, given that this is the century when laísmo was clearly at its peak simply due to imitation of the prestigious linguistic norm of Madrid (Lapesa 2000 [1968]: 304; Sáez Rivera 2008: 1090). Thus, these cases of laísmo we have documented in authors from La Mancha could be explained by their closeness to Madrid, the central disseminator of innovative usages throughout the century in question. We set out some examples below:

(3) a. En fin, madre ha estado muy enfadada e inquieta de eso y no quería ir a confesar con don Diego. Y la disuadí y la hice que fuera (PSCR5734, 1721).
b. Pues te digo que eres una tonta porque te apesadumbras de poco. Lo que sentiré se haya llevado las cosillas que la has dado, que lo demás es una friolera (PSCR5748, 1738)
c A la hora poco más de haber confesado a una mujer, llegamos los dos a tomar agua bendita. La fui acompañando por la calle y la empecé a decir palabras equívocas (PSCR5738, 1791).

Thirdly, and as reflected in the previous map (red dots), there are cases of laísmo in various authors —up to 12— whose place of origin is considerably removed from the current confounding zone: four from Pontevedra, two from Seville and one each from Llanes (Asturias), Vergara (Guipúzcoa), Baquio (Vizcaya), Santiago de Compostela (La Coruña), Zafra (Badajoz) and Tortosa (Tarragona). A number of reasons may be put forward for giving account of these cases.

On the one hand, particular mention should again be made of the influence the Madrid norm may have had on some of these authors, above all in the eighteen century, although also on others who produced their missives in later eras. Examples of the spread of laísmo beyond the *meseta central* have been indicated in different diachronic studies on pronominal variation: cases of laísmo have been documented in authors from Andalusia (Pérez Teijón 1985: 87-95), Aragon (Sáez Rivera 2008: 1089) and Galicia (Octavio de Toledo 2019: 100). Our data corroborate this dissemination of laísmo beyond its vernacular borders. See the following by way of example:

(4) a. Y haz tú el concierto con nuestra patrona y pon todo lo que la dieres a cambio (PSCR6546, 1701, author from Asturias).
b. Y avísenme VMs cuántas son [las barricas] en todas a punto fijo para que, cuando salga el decreto, las ponga el número (PSCR6688, 1732, author from Seville)
c. No sé que tu hermana tenga novedad desde que murió tu madre. Llegué a tu casa de paseo con el reverendo padre fray Layón con el designio de darla algunos consejos, los cuales recibió con gusto (PSCR7800, 1833, author from Pontevedra)

On the other hand, some cases of laísmo beyond the area where it is supposedly expected are explained upon reviewing the biographical profiles of their respective authors. We thus know that the author originating from Baquio (Biscay) also lived in Valladolid, which is a clearly laísta zone. Also, a number of the laísmos associated with Seville are documented in the letters of a countess who had residences in Toro (Zamora) and in Leon. In both cases, therefore, place of residence permits the explanation of what place of birth fails to clarify, at the same time illustrating the relevance of having historical corpora that include this type of metadata.

Finally, we should take into account the possibility that some of these cases do not constitute true laísmos, rather, that they are the result of a transcription error or a questionable reading due to the poor condition of the manuscript. Let us not forget that, for the case we are interested in, the simple changing of an *e* to *a* is sufficient reason to introduce a false positive in the analysis. This latter circumstance is in our opinion applicable to the case of the author originating from Zafra (Badajoz), whose sole example is that which we give below. We accompany the example of the corresponding fragment in the original source[25]:

(5) VM me haga merced darme noticia de mi señora doña María López de Castro, a quien escribí de Amberes, que estimaré la vaya bien (PS6044, 1687, author from Badajoz)

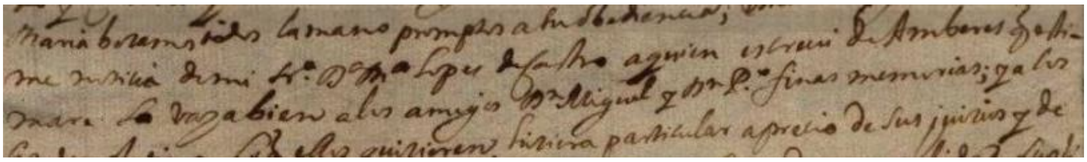

Figure 2. Fragment of letter PS6044. Year 1687.

Note that the manuscript reveals a corrosion in the ink on the vowel of the clitic, at the beginning of the third line of Figure 2, which creates doubt regarding the true form used by this author: *la* or *le*? Beyond the solution that has been adopted in the transcription, this case serves to demonstrate another advantage of small and tidy historical corpora: the possibility of including the facsimile editions together with the text transcriptions. Access to the facsimile affords researchers the possibility of checking the occurrence of a specific phenomenon against the original, and thus being able to make a decision on whether or not to include it in their analysis.

4. The diminutives *-ito* and *-ico*

4.1. Description of the phenomenon

The Spanish boasts a considerable inventory of appreciative suffixes, amongst which diminutives comprise a particularly interesting series for linguistic research. The employment of the diminutive forms in contemporary Spanish reveals a rich and complex dialectical panorama. The *-ito/-ita* form is the most widespread. Together with it, and limiting ourselves solely to peninsular Spanish, we find the following variations, which though not exclusive are predominant in their respective territories: the suffix *-ete/-eta* in Aragon, Levante and Catalonia; the suffix *-illo/-illa* in Andalusia; the suffix *-ico/-ica* in Navarre, Aragon, Murcia and Granada; the suffix *-ín/-ina* in Asturias and Leon; the suffix *-ino/-ina* in Extremadura; the suffix *-uco/-uca* in Cantabria; the suffix *-ejo/-eja* in La Mancha, and the suffix *-iño/-iña* in Galicia (NGLE 2009: §9.1j, Bouzouita, Castillo & Pato 2008: 74). All of these variants are documented in PS and/or ODE, as shown by the examples we set out in (5):

(5) a. No te canso más. Darás más besitos a mi hija (PSCR7619)
b. Y él era mercadercete de agujas y de los tales (PSCR7066)
c. Un espejillo pequeño viejo (AL1704I0001)
d. Un mantico de paño negro pequeño (GR1552I9115)
e. me costó treinta doblones por mano de clavel para mi brujina (PSCR5417)
f. Dos potesinos pequeños con una poca de aceite (BA1601I7042)
g. Se pondrán muy perdidos, [...] por los papelucos que han recogido (PS8092)
h. Tengo más tres bacías de azófar medianas y una bacineja de lo mismo (BA1644I7028)
i. Recibí la de usted. Remito el paniño que me pide (PS6040)

In this work we shall focus on the employment of the suffixes -*ito* and -*ico* in nouns, once again paying particular attention to their geographical distribution during the classic and modern Spanish periods. Interested readers can find general studies on the diminutive in contemporary

[25] The complete transcription of this letter, along with the facsimile of the manuscript, are accessible via the following url: http://teitok.clul.ul.pt/postscriptum/index.php?action=file&cid=PS6044.xml.

Spanish in the works of Gooch (1970), Montes Giraldo (1972), Uritani & Berrueta (1985), Lázaro Mora (1999) and Callebaut (2011). Among the historical approaches, particularly worthy of mention are González Ollé (1962) for the medieval period and Náñez Fernández (2006 [1973]) for the early modern period, together with other studies included in works of greater scope, such as Alvar & Pottier (1983: 363-380), Urrutia & Álvarez (2001: 123-127) and Penny (2010: 319-323).

4.2. Data collection

As in the case of laísmo, retrieving data from a corpus on uses relating to diminutives —and these uses only— entails a number of difficulties. In principle, users interested in studying phenomena of a morphological nature, and particularly affixed elements, can resort to wildcards operators in order to obtain those words that start with (prefixes), contain (infixes, interfixes) or end in (suffixes) a given sequence of characters. Linguistic corpora, in fact, normally include search interfaces that admit this type of wildcard queries.

Note, however, that this strategy will return a higher number of examples than intended; in other words, it will include an indeterminate group of false positives in the results. In the case of diminutives, this excess group will comprise a minimum of two categories: (i) words that end in the requested sequence, but the sequence is not a diminutive (ii) words that end in the requested sequence and the sequence is a diminutive that has lost its semantic value and, as a result, no longer functions as such. Starting out from the affixes *-ito* and *-ico*, first case examples would be words such as *delito*, *hábito*, *médico* and *hocico*, with second case examples being words such as *pajarita*, in the sense of "tie that is tied in the form of a short bow", and *abanico*, in the sense of "instrument comprising a group of joined ribs". In short, the results extracted from the corpus via graphic word endings shall include both pure graphic coincidences and lexicalised suffixes.

We again find ourselves facing a linguistic phenomenon that can particularly benefit from small corpora for its filtering and analysis. The successive revision campaigns that have been carried out on the linguistic annotation of the PS and ODE corpora have led to the incorporation and refining of information regarding the employment of appreciative suffixes. This information has been included in the grammatical tags corresponding to nouns. In this way, the word *amigazo* is assigned the tag NCMS00A (Noun, Common, Masculine, Singular, Augmentative; and, similarly, the word *anillito* is assigned the tag NCMS00D. In both corpora, therefore, the final position of the noun tag is reserved for marking whether the noun in question contains an augmentative (A) or diminutive (D) suffix.

Using this strategy has enabled us to identify the total number of occurrences of the diminutives *-ito* and *-ico* in common nouns. Below we set out the syntax in CQP that is necessary to apply in each corpus to retrieve this information:

```
In PS:      [pos="NC.+D" & nform=".+i(t|c)(o|a)s?"] :: match.text_lang = "ES"
In ODE:     [pos="NC.+D" & nform=".+i(t|c)(o|a)s?"]
```

As occurs when obtaining occurrences of laísmo, in this case we have also made use of a combined search with the attributes *pos* and *nform*. The first attribute has been assigned a value that permits us to obtain all common (C) nouns (N) that include a diminutive suffix (D). The intermediate positions considered for nouns correspond to gender and number, but are irrelevant for this search, hence they have not been specified; instead, a regular expression (.+) has been applied, which represents "one or more characters" to thus obtain all possible combinations of this grammatical tag in these positions. As regards the normalised form, the syntax necessary for obtaining all words ending in the sequences *-ito* and *-ico* has been specified, likewise considering the feminine and plural forms. The result of these queries is summarised in Table 4:

	PS (Spanish)	**ODE**	**Total**
-ITO	95	283	378
-ICO	33	151	184
Total	128	434	562

Table 4. Distribution of the diminutives *-ito* and *-ico* in Post Scriptum and ODE

ODE constitutes a data source that is particularly appropriate for obtaining employments of diminutive suffixes. Their productivity in this regard clearly derives from some of the text types included in the corpus: inventories of goods. These lists of objects, some of considerable length, return a varied and extensive number of nouns, through many of which it is possible to document linguistic features of the vernacular, such as uses of the diminutive. Furthermore, the data yielded by PS, despite being more reduced, are useful for enabling control of a larger geographical space. We will therefore make use of both datasets for the analysis of *-ito* and *-ico* that we offer in the following section.

4.3. Results and analysis

As we have already mentioned, at present the most common diminutive in Spanish-speaking countries is *-ito*. This predominance also applies to peninsular Spanish, where the suffix *-ito* gained ground on other affixal solutions from at least the 18th century (Náñez Fernández 2006[1973]: 231). In regards to the diminutive *-ico*, the *Nueva Gramática de la Lengua Española* explains that its current usage spans “zonas nororientales y meridionales de la Península Ibérica” (NGLE 2009: §9.1j). This affirmation is specified and broadened in different specialised studies, in which the geographical distribution of this suffix is, on the whole, delimited to the areas of Navarre, Aragon, Murcia and Granada (Náñez Fernández 2006[1973]: 33, Bouzouita *et al*. 2017: 74). The use of *-ico* has also been documented in certain regions of Leon and Zamora (Borrego Nieto 1996: 146-151).

With these references on the spread of both suffixes in present day peninsular Spain, we then contrast it with the results returned by PS and ODE for the early modern period. To do so, we have again imported the data from these corpora to the QGIS 3.10 program in order to trace an historical-dialectical map on the phenomenon under study. In this case, we have opted to take the province as a reference indicator and group the occurrences of *-ito* and *-ico* for each. After that, the contrast established between both solutions in different peninsular provinces was projected on the map. This strategy has brought us to select only those provinces for which we have a minimum number of examples. In this way we have ruled out those provinces from our counting where the sum of *-ito* and *-ico* usage returned a result of fewer than five occurrences. As a result, the final data that have been subject to georeferencing stands at a total of 512 employments (330 cases of *-ito* and 182 cases of *-ico*), and are those that appear in Table 5. The data in this table appear distributed by province and in descending order from the percentage of *-ico* documented in each one. The resulting map is shown in Figure 3:

Province	***-ito***	***-ico***	***% -ico***
Granada	15	69	82.1
Almeria	19	54	74.0
La Rioja	3	8	72.7
Caceres	10	20	66.7
Jaen	5	5	50.0
Toledo	3	2	40.0
Malaga	15	5	25.0
Salamanca	7	1	12.5
Cuenca	8	1	11.1
Seville	35	4	10.3
Badajoz	13	1	7.1
Madrid	117	8	6.4
Huelva	15	1	6.3
Cadiz	65	3	4.4

Table 5. Distribution of -*ito* and -*ico* by province.

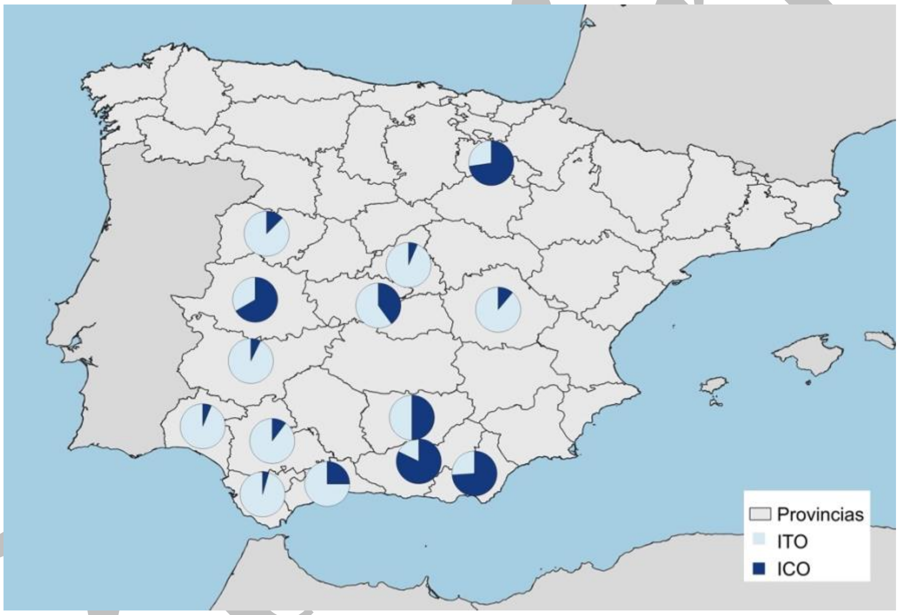

Figure 3. Geographical distribution of -*ito* and -*ico*

From the map we have obtained (Figure 3) at least two questions arise worth drawing attention to. Firstly, and leaving aside the chronological factor, it is enlightening the dialectal situation produced by our data as regards the distribution of -*ito* and -*ico* in Andalusia, and which leds to the establishment of an evident distinction between the eastern and western part of this territory. Granada and Almeria are the Spanish provinces that reflect a greater intensity of the suffix -*ico*, and, as the map shows, this intensity drops slightly as we move towards western Andalusia. Thus, Jaen displays the same number of occurrences of -*ito* and -*ico* (5/5), whereas in the case of Malaga there is an imbalance in favour of the former (15/5). Lastly, in the westernmost provinces —Cadiz, Seville and Huelva— the intensity of the -*ico* solution is exceptionally low, varying between 4% and 10%.

This contrast in the use of diminutives is observed more clearly when comparing examples that show analogous syntactic structures. ODE, in fact, contains examples where a single noun is constructed with *-ito* (6) and *-ico* (7) in similar contexts and in texts that have been produced in western or eastern Andalusia, respectively:

(6) a. Dos laminitas con marco dorado en doce reales (HU1715I0321, Huelva)
b. Ítem dos mesitas de pino también viejas (CA1767I0307, Cadiz)
c. Un espejito de manos con su marquito encarnado (SE1710I0327, Seville)
d. Una toallita de tafetán listado viejo (CA1708I2525, Cadiz)

(7) a. Cuatro laminicas con sus molduras en diez y seis reales (GR1727I2711, Granada)
b. Ítem una mesica de pino usada (AL1803I0052, Almeria)
c. Un espejico pequeño con su marco tallado en siete reales (GR1748I2555, Granada)
d. Tres toallicas de seda listadas de diferentes colores (GR1707I2710, Granada)

Secondly, it is interesting to see that *-ico* is also documented in geographical points that currently do not have an association with the use of this diminutive. The zones of Caceres and La Rioja stand out due to the number of cases observed, and this also applies to Toledo, albeit to a lesser degree, as for this latter province there is a particularly low number of occurrences (3 cases of *-ito* and 2 of *-ico*). There is a total of 30 occurrences of the *-ico* form between the three. These data invite the reflection of how usage of this diminutive may have spread in times past, and because of this it is worth paying attention to the chronological distribution of these uses in the corpora. In this regard, mention can be made of the 30 examples of *-ico* documented in the aforementioned three provinces between the 16th and 17th centuries.

In the case of the data for Caceres, there are 15 examples taken from ODE and 5 from PS. All of the examples of ODE are documented in inventories of goods produced during the first half of the 17th century. The 5 remaining examples are documented in two letters written in 1548 by Pedro de Orellana, a franciscan monk from Trujillo (Caceres). In (8) we show some of these examples taken from both corpora:

(8) a. Unos brinquiños de barro y de vidrio y una cestica y unas cucharas. (CC1615I7076)
b. Dos toallicas de cabeza viejas (CC1639I7097)
c. Ese librico de los afeites, van los cuadernos por su abecedario y cuenta (PS4139)
d. y en llegando aquí dé, so estas ventanas que taparon, con una piedrecica dos golpes (PS4138).

The 8 examples attributed to La Rioja were extracted from the PS corpus and belong to the same author, Antonio de Medrano, a priest from Fuenmayor accused of the heretical crimes of *alumbrismo* and *epicureísmo* by the Toledo inquisition in 1530. The examples are distributed into various notes written by this author from prison to his brother Bernardino Díaz de Medrano, requesting he send him different items of food. In (9) we recover some examples, all produced around 1530:

(9) a. Y envíeme siempre un pastelico a comer y cenar (PS4015)
b. Y compre una cestica para en que venga fresca la verdura (PS4029)
c. Envíenos un panecico de sal, y las aceitunas que sean mejores (PS4027)
d. Envíeme una redomica de agua rosada y una vasija de vidrio para enfriar vino (PS4016)

Finally, two cases of *-ico* documented in Toledo, both taken from PS. The oldest (10a) is used in a letter from an uncertain date, written between 1639 and 1642 by a certain María Bautista Beléndiz, from Toledo. Note the presence of the suffix *-illo* together with *-ico* in this example. The second case (10b) is documented in a letter from 1684 written by a linen and silk merchant called Pedro Molina Palacios:

(10) a. Te pido y ruego que a la mujer que vivía en el aposentillo [...] le des dos bolsicos de ámbar que están en el baúl (PSCR5422)
b. Y en cuanto a llevar las medias, si VM tiene disposición de llevarlas consigo en su cabalgadura, será preciso hacer dos frangoticos (PS5017)

This group of examples, located in geographical areas that are removed to a greater or lesser extent from the regions to which a dialectal *-ico* in contemporary Spanish is attributed, ultimately reveal that the use of this suffix must have covered a more extended area during the 16th and 17th centuries than the present day. This situation in reality is well situated in the panorama normally described for the system of evaluative suffixes during the Spanish Golden Age, where the suffix *-ico* was in an intermediate position between *-ito* and *-illo* in terms of frequency of use:

> En la morfología derivativa los diminutivos más frecuentes eran, por este orden, *-illo*, *-ico* e *-ito*. […] El sufijo *-ico* era la forma cortesana en el siglo XVI, sin las connotaciones aragonesas y murcianas de hoy, pero en el siglo XVII aumentó el prestigio de *-ito* e *-ico* ganó rusticidad y evocación dialectal (Girón Alconchel 2004: 861).

Finally, our data are also compatible with the results unearthed by prior studies on the scope of the diminutive *-ico* during the classical Spanish era. Calderón Campos (2019: 119), for example, offers accounts of *-ico* in the 17th century from the CORLEXIN corpus and obtains statistically significant results for provinces such as Leon, Murcia, Navarre, Albacete, Teruel, Zamora, Almeria, Huesca and Alicante. In light of the results shown here, it would be possible to add the province of Caceres, at the very least, to the above list. This would mean extending the majority usage of *-ico* found in Leon and Zamora towards the south, with the possible exception of Salamanca, the explanation of which would require further data to be obtained[26]. The cases of Toledo and La Rioja would also deserve somewhat larger accounts, both in terms of number of occurrences and number of informants. Notwithstanding, it is reasonable to think that during the classical Spanish era there would have been a greater intensity of the suffix *-ico* at least in La Rioja, taking into account its closeness to Navarre.

5. Seseo

5.1. Description of the phenomenon

Not only do PS and ODE permit reliable information to be obtained in regards to morphosyntactic phenomena, they are also extremely useful for analysing questions of a phonetic nature. We shall illustrate this type of approach focusing on one of the phonetic features that has received most attention on the part of researchers: seseo. It is known that in the transition from medieval Spanish into classical Spanish a phonological restructuring in the sibilant consonant system was carried out. This restructuring gave rise to two solutions that continue to prevail in contemporary Spanish: maintaining an opposition between the voiceless fricative alveolar phoneme /s/ and the voiceless interdental fricative phoneme /θ/ or neutralising said opposition in favour of a single phoneme. In the latter case, the phoneme resulting from the neutralisation may have ciceant timbre, giving rise to the phenomenon of ceceo, or siseant timbre, giving rise to the phenomenon of seseo (Moreno Fernández 2004: 976).

[26] Out of the group of provinces represented on ODE, Salamanca has the lowest number of words at the time of writing these lines. It is, furthermore, one of the least represented provinces in the CORLEXIN corpus.

Seseo thus describes a variety of Spanish pronunciation characterised by the absence of the /θ/ phoneme, all occurrences of which are replaced by the /s/ phoneme[27]. As a result, seseant speakers pronounce pairs of words such as *casa~caza*, *cien~sien* and *cocer~coser* in the same way. The seseo or confounding system is considerably more widespread, given it is used in the Spanish of Latin America, the Canary Islands and a large part of Andalusia, and in determined points of Murcia and Badajoz (NGLE 2011: §5.5k).

There is an extensive bibliography around seseo, both from the contemporary perspective, focusing generally on social distribution and/or articulatory description of the phenomenon in different regions, and the historical perspective, normally concerned with unearthing documentation that permits its appearance to be dated with relative precision. In this work we shall limit ourselves to observing the distribution of seseo in peninsular Spanish during the early modern period. A complete and updated state of the art regarding seseo may be found in Núñez-Méndez (2021). Amongst those works of a diachronic nature, mention should be made of Amado Alonso (1955), Catalán (1956), Lapesa (1957) and Alvar (1983). Other noteworthy studies include those by Cock (1969) and Frago (1992), along the more recent ones by Kania (2016) and Kauffeld (2016).

5.2. Data collection

The obtaining of historical data on seseo, and in general on any phonetic aspect from handwritten documentation, must face up to the obvious limitation that writing imposes as a sole data source. Therefore, to work with reliable data it is essential that the corpus used fulfil three fundamental characteristics, namely: (i) the application of the same editorial principles throughout the entire corpus, (ii) the creation of transcriptions in which the spelling of the original source is scrupulously respected, and (iii) the preparation of a level of edition, parallel to the original transcription, with normalised spelling according to the contemporary standard.

The two first aspects guarantee the quality of transcriptions, true to the original text and free of discrepancies. The third aspect increases the possibilities of exploiting the corpus, as it brings with it two additional advantages: on the one hand, it enables the retrieval of all spelling variations of a word, given they are all linked to the same normalised form, which may be used as a reference at the time of constructing the query; on the other hand, permits the easy detection of phonetically-based spelling variation, via the comparison of the original forms with their normalised counterparts. It is in this way possible to identify in the corpus cases of seseo (*sinco~cinco*), rotacism (*corchón~colchón*), lambdacism (*almario~armario*) or alterations of vowel sequences (*sepoltura~sepultura*), amongst other phenomena of interest.

Note that none of the three aforementioned conditions are met in the historical reference corpora for Spanish. The size of the latter can only be achieved resorting to existing editions, which will be based on inconsistent editorial criteria and, in the case of texts based on handwritten documentation, will not always faithfully reproduce the orthography of the original source (Honkapohja *et al*. 2009: 456-458, Rojo 2021: 78). As a result, the work involved in the spelling normalisation of texts is not longer beneficial for this type of corpus, given they lack a palaeographical version that permits the contrasting and exploitation thereof.

[27] The performance of the /s/ phoneme in seseant Spanish zones has been described in different ways: dorsoalveolar, perdorsoalveolar, predorsodental and dental (Kania 2016: 236, note 3). In non-seseant or distinguishing zones, the typical performance of this phoneme is apicoalveolar.

The PS and ODE corpora do boast these three particular features, making them useful digital resources for researching phonetic issues from their corresponding graphical representations. For the specific case of seseo we shall limit ourselves to the ODE corpus for two fundamental reasons. Firstly, only ODE is enriched with specific annotation for retrieving seseant forms from the corpus. PS also includes linguistic annotation on aspects with a phonetic base, but not on seseo in particular[28]. Secondly, ODE contains greater and more varied documentation originating from the south of the peninsular, and is therefore more appropriate than PS for observing the distribution of seseo, both for the number of documented cases and geographical areas covered.

The annotation of seseant forms in ODE was carried out automatically from the comparison of the spellings *s* and *ss* present in the original words with the spellings *c(e)*, *c(i)* and *z* present in the normalised counterparts, as long as these spellings appeared in the same position within the two compared forms. We then carried out a manual review to correct a number of false positives (e.g. *paresçer~parecer*). The final result stood at 4565 occurrences of seseo, which can be retrieved using the following query in CQP syntax:

```
[ltags=".*seseo.*"]
```

The *ltags* attribute is reserved in ODE for tagging linguistic information that is not contemplated in the grammatical tags based on the EAGLES standard. This generally involves information of a phonetic nature. The regular expression used in this query (.*) represents “zero or more characters” and here it is necessary for including in the result those words that have not just been tagged as seseo, but also illustrate other types of phenomena: for example, the form *sincoenta* (*cincuenta*), which as well as seseo shows a vowel alteration, and the form *carsones* (*calzones*), which along with seseo constitutes an example of rotacism.

5.3. Results and analysis

As put forward in the previous section, a total of 4565 cases of seseo have been obtained in the ODE corpus. In (11) we reveal a small sample of this broad group. As these examples show, transcribed here in their original form to identify the spellings with phonetic value, seseant forms are documented in texts produced in different regions and eras from the corpus.

(11) a. dexando el suelo parejo conforme las d(ic)has dos **sanjas** (AL1593D9128, Almería)
b. Treynta pares de **calsetas** de hilo hordinarias (BA1645I7001, Badajoz)
c. un colchon de **lienso** blanco **apresiado** en treintta y tres r(eale)s (HU1702I0315, Huelva)
d. **mosa donsella** de edad de veinte años (SE1780I7048, Seville)
e. **Dose** libras y media de **tosino** a ocho r(eale)s (CA1800I2111, Cádiz)
f. han **meresido** y **meresen** el mejor **consepto** (MA1829D9098, Malaga)

From this group we proceeded to account for the occurrences of seseo in each of the provinces represented in ODE. Given that the number of words per province varies considerably, we also proceeded to calculate the normalised frequencies, in other words, the cases of seseo per million words in each territory. The result obtained is shown in Table 6. The data appear ordered according to normalised frequency (*ipm* stands for *instances per million*):

[28] In PS, the cases of seseo are annotated with a more general tag (“consonant_system”), which also includes other types of phenomena having to do with consonantal changes, such as those illustrated with the pairs *abuja~aguja*, *arvierto~advierto*, *calunia~calumnia* or *llamao~llamado*.

Province	***words***	***seseo***	***seseo (ipm)***
Huelva	28388	943	33218
Seville	30544	906	29662
Cadiz	42411	504	11883
Badajoz	78168	888	11360
Malaga	84802	547	6450
Granada	174143	456	2618
Almeria	110708	254	2294
Caceres	36094	37	1025
Others	118952	30	305
Total	704210	4565	

Table 6. Distribution of seseo in ODE.

Lastly, and following the same strategy as for the two preceding phenomena, we loaded the results into the QGIS 3.10 geographic information system to project them on a map and facilitate their interpretation. In order to create the seseo map, the normalised frequencies have been classified by range and a symbology of graded colour has been applied, to such an extent that the provinces can be contrasted in regards to the intensity of the phenomenon. The result is that shown in Figure 4:

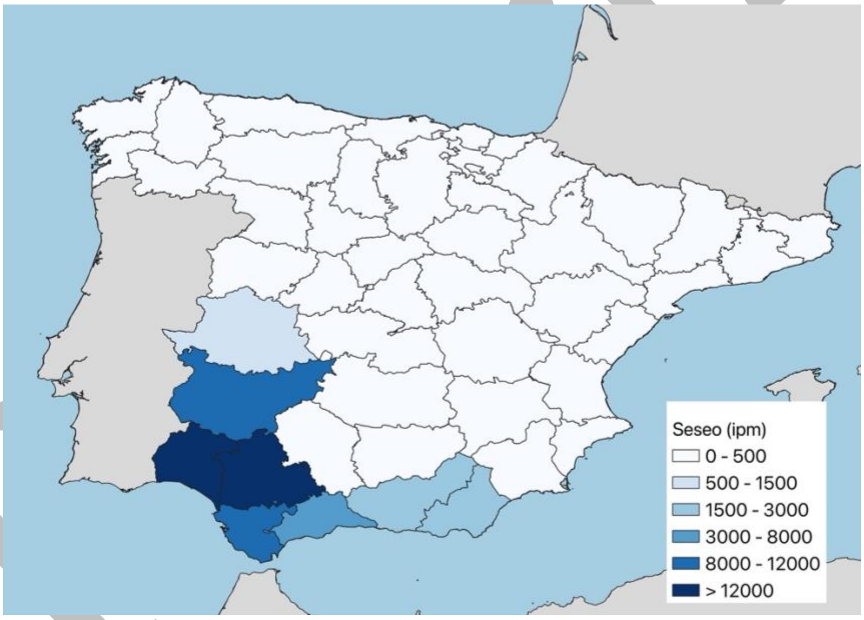

Figure 4. Geographical distribution of seseo.

Note that all of the data obtained have been projected onto the map, regardless of the era to which they correspond. We are aware that it would be interesting to establish at least a comparison by centuries to observe the geographical distribution of seseo at different stages of the early modern period. ODE is however a resource under construction and we do not currently have data produced in different centuries for a good number of geographical areas: for example, all of the documentation related to Extremadura in the corpus is from the 17th century, and all which corresponds to eastern Andalusia is from the 18th century. In this work, we will thus limit ourselves to offering and commenting briefly on this overview.

The most noteworthy aspects shown by this map can be summarised in two points. Firstly, it is observed that the provinces of Huelva and Seville comprise a focal point for the spread of seseo, which clearly extends in vertical and horizontal axes. The former affects both the province of Cadiz along the south, and the province of Badajoz, along the north, in equal measures. Although the presence of seseant forms in the documentation related to Badajoz may seem surprising, it is also widespread in the data produced by the CORLEXIN corpus, as shown by the information provided by Calderón Campos for the 17th century (2019: 116). This spread of seseo towards the north is even reflected in the documentation related to Caceres included in ODE. At the time of writing these lines, this corpus has 37 seseant forms distributed in nine inventories originating from the aforementioned province. We put forward some examples in (12) [29]:

(12) a. tres savanas de **lienso** nuebas y una de estopa (CC1615I7082)
b. Seis fanegas de **seuada** que estan sembradas (CC1639I7096)
c. la tapa del pimentero dorado y **sinselado** (CC1621I7091)

Regarding the spread of seseo in Andalusia, our data show an intensity from the west to the east, which implies two well differentiated areas: a seseant western area, comprising the provinces of Huelva, Seville, Cadiz and, to a lesser extent, Malaga, and an eastern area taking in Granada and Almeria, and with a clearly lower intensity of seseo. In regards to the provinces of Córdoba and Jaen, at the time of writing ODE lacks sufficient documentation, to which we cannot provide meaningful data. Notwithstanding this, using the data from CORLEXIN provided by Calderón Campos as a basis, both provinces would have no problem in fitting into this gradual panorama we are discussing (Calderón Campos 2019: 116-117).

Secondly, it is noteworthy that the impact of seseo in the other provinces represented in ODE is practically anecdotal: 30 seseant forms in the group comprising the texts produced in Madrid, Valladolid, La Rioja, Cantabria, Toledo, Palencia, Zamora, Burgos, Navarra, Alava, Soria, Leon, Gipuzcoa, Teruel and Salamanca; the equivalent to 0.6% of the total seseant forms documented in ODE. It should also be pointed out that over half of these occurrences —16 out of 30— appear in a single document from Madrid —with the code MD1711I0105—, wherein the spellings *c(e)*, *c(i)* and *z* are clearly predominant and even appear in the context immediately prior to or following the supposedly seseant forms. See examples in (13):

(13) a. Dos cortinas de barragan p(ar)a silla de manos guarnecidas de **flequesillo** azul
b. Veinte y dos zençerros medianos de yerro. Otro **senzerro** grande de yerro
c. clabos de media chilla de ala de mosca **viscainos**. Catorze doçenas de tragacetes de pino.

In summary, this general seseo distribution, with different degrees of intensity for the southern part compared to very sporadic —possible accidental— occurrences in the rest of the peninsular, reveals two questions that are especially relevant in regards to the exploitation of the ODE data. On the one hand, it is demonstrated that as a whole the confounding spellings (*sinco~cinco*, *lienso~lienzo*, *asul~azul*…) cannot be interpreted as casual spelling errors; rather, they reflect authentic representations of the spoken word in writing and, as a result, constitute variants indicative of seseo. The valid of the ODE data is thus confirmed for carrying out phonetic analysis. On the other hand, it is demonstrated that the geographical origin of the scribes who produced these texts had to be very close, if not identical, to that of the places where they were written, which shows the usefulness of these corpora for undertaking dialectal studies.

[29] For documentation related to Extremadura collected in the ODE corpus, see González Sopeña (2022).

6. Conclusions

The creation of diachronic corpora in electronic format has meant an unquestionable advancement for research into language history, putting at the service of the specialist, and any interested user, a large collection of historical texts with which it is possible to work on an empirical basis.

In the case of Spanish, we are fortunate to be able to rely on a number of historical reference corpora, such as CORDE, CDH and CdE and, together with them, in recent years new and varied specialised historical corpora have started to be constructed. This work has focused on two of the latest: PS, comprising private letters, and ODE, made up of witness statements, inventories of goods and medical certificates. Both compile documentation from the early modern period and both constitute, in the diachronic sphere, examples of small and tidy corpora, in other words, reduced size corpora that show careful elaboration throughout all stages of their construction: from the selection of texts to their edition in digital form, to their assigning of metadata and linguistic annotation.

The advantages these two corpora afford to the historical research of language can be summarised in six aspects, which have been highlighted throughout this work: (i) a careful selection of texts with content that approaches to a greater or lesser extent the pole of communicative immediacy; (ii) the control of extralinguistic variables, amongst which of particular importance is the geographical origin of authors and scribes, which provides a trustworthy data source for undertaking dialectal studies; (iii) access to facsimile images of documents, permitting the contrasting of transcriptions with original manuscripts to verify their content; (iv) the creation of both transcriptions faithful to the original and textual versions with normalised spelling, the comparison of which enables the identification of phonetic phenomena, such as seseo; (v) detailed grammatical tagging based on the EAGLES standard, manually revised and adapted for annotating specific phenomena, such as laísmo and appreciative suffixes; and (vi) a search engine that admits queries based on CQP syntax and permits precise and exhaustive information retrieval.

Because of these textual, technical and methodological characteristics, not only do PS and ODE complement each other, they are also complementary to the large historical corpora of Spanish. We are confident that their appropriate exploitation, of which we have in this paper merely been able to offer a small sample, enables continued advancement in knowledge of peninsular Spanish throughout the early modern period.

Reference list

Alonso, Amado (1955): *De la pronunciación medieval a la moderna en español*. Vol 1. Madrid: Gredos.

Alvar, Manuel (1983): ‘A vueltas con el seseo y ceceo’. In Francisco Marcos Martín (ed.): *Introducción plural a la gramática histórica*. Madrid: Cincel, pp. 130–144.

Alvar, Manuel; Pottier, Bernard (1983): *Morfología histórica del español*. Madrid: Gredos.

Borrego Nieto, Julio (1996): ‘Leonés’. En Manuel Alvar (dir.): *Manual de dialectología hispánica. El español de España*. Barcelona: Ariel, pp. 139-158.

Bouzouita, Miriam; Castillo, Mónica; Pato, Enrique (2018): ‘Dialectos del español. Una nueva aplicación para conocer la variación actual y el cambio en las variedades del español’. *Dialectologia: revista electrònica*, 20, pp. 61-83.

Calderón Campos, Miguel (2019): ‘La configuración de la variedad meridional en el reino de Granada’. In Eugenio Bustos Gisbert & Juan Pedro Sánchez Méndez (eds.): *La* *configuración histórica de las normas del castellano*. Valencia: Tirant Humanidades, pp. 109-134.

Calderón Campos, Miguel; Vaamonde, Gael (2020): 'Oralia Diacrónica del Español. Un nuevo corpus de la Edad Moderna'. *Scriptum Digital*, 9, pp. 167-189.

Callebaut, Sien (2011): *Entre sistematización y variación. El sufijo diminutivo en España y en Hispanoamérica*. Gent, Belgium: University of Ghent. MA thesis.

Cano Aguilar, Rafael (1996): 'Lenguaje espontáneo y retórica epistolar en cartas de emigrantes españoles a Indias'. In Thomas Kotschi, Wulf Oesterreicher & Klaus Zimmermann (eds.): *El español hablado y la cultura oral en España e Hispanoamérica*. Madrid: Iberoamericana, pp. 375-404.

Catalán, Diego (1956): 'El ceceo-zezeo al comenzar la expansión atlántica de Castilla'. *Boletín de Filología*, 16, pp. 305–34.

Claridge, Claudia (2008): 'Historical corpora' Anke Lüdeling & Merja Kytö (eds.): *Corpus Linguistics. An International Handbook. Volume 1*. Berlin: Walter de Gruyter, pp. 242-259.

Cock, Olga (1969): *El seseo en el Nuevo Reino de Granada 1550–1650*. Bogotá: Instituto Caro y Cuervo.

Dollinger, Stefan (2004): 'Philological computing vs. philological outsourcing and the compilation of historical corpora: a Late Modern English test case'. In Christiane Dalton-Puffer et al (eds.): *Vienna English Working Papers (VIEWS)* 13, pp. 3–23.

Echenique Elizondo, María Teresa (1981): 'El sistema referencial en español antiguo: leísmo, laísmo y loísmo'. *Revista de Filología Española*, 61, pp. 113-157.

Evert, Stephan; Hardie, Andrew (2011): 'Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium'. *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham: University of Birmingham.

Fernández-Ordóñez, Inés (1999): 'Leísmo, laísmo y loísmo'. In Ignacio Bosque & Violeta Demonte (coords.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 1322-1397.

Fernández-Ordóñez, Inés (2001): 'Hacia una dialectología histórica. Reflexiones sobre la historia del leísmo, el laísmo y el loísmo'. *Boletín de la Real Academia Española*, LXXXI, pp. 389-464.

Flores Cervantes, Marcela (2006): 'Leísmo, laísmo y loísmo'. In Concepción Company Company (dir.): *Sintaxis histórica de la lengua española. Primera parte: la frase verbal*. México: UNAM, Fondo de Cultura Económica, pp. 671-749.

Frago Gracias, Juan Antonio (1992): El seseo: orígenes y difusión Americana. In César Hernández Alonso (ed.): *Historia y presente del español de América*. Valladolid: Junta de Castilla y León, pp. 113–142.

Francis, W. Nelson (1992): 'Language corpora B. C.'. In Jan Svartvik (ed.): *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, pp. 17-32.

Girón Alconchel, José Luis (2004): 'Cambios gramaticales en los Siglos de Oro'. In Rafael Cano Aguilar (coord.): *Historia de la lengua española*. Barcelona: Ariel, pp. 859-894.

Gómez Seibane, Sara (2012): *Los pronombres átonos (le, la, lo) en el español*. Madrid: Arco/Libros.

Gómez Seibane, Sara (2013): *Los pronombres átonos (le, la, lo) en el español: aproximación histórica*. Madrid: Arco/Libros.

González Ollé, Fernando (1962): *Los sufijos diminutivos en castellano medieval*, Madrid: CSIC.

González Sopeña, Inmaculada (2022): 'Documentación notarial extremeña del siglo XVII en Oralia diacrónica del español (ODE): el léxico de la vida cotidiana a través de inventarios de bienes pacenses', *Romanica Olomucensia*, pp. 13-30.

Gooch, Anthony (1970): *Diminutive, Augmentative and Pejorative Suffixes in Modern Spanish: A Guide to Their Use and Meaning),* Oxford: Pergamon Press.

Gries, Stefan (2009): *Quantitative Corpus Linguistics with R: A practical introduction*. Londres: Routledge.

Honkapohja, Alpo, Samuli Kaislaniemi & Ville Marttila (2009): 'Digital Editions for Corpus Linguistics: Representing Manuscript Reality in Electronic Corpora', in Andreas H. Jucker, Daniel Schreier & Marianne Hundt (eds.): *Corpora: Pragmatics and Discourse,* Amsterdam/New York: Rodopi, pp. 451–475.

Janssen, Maarten (2016): 'TEITOK: Text-Faithful Annotated Corpora'. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2016) ELRA*. Portorož, Slovenia, pp. 4037-4043.

Janssen, Maarten; Vaamonde, Gael (2020): 'Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK'. In Rosario Álvarez & Ernesto González Seoane (eds.): *Calen barbas, falen cartas. A escrita en galego na Idade Moderna*. Santiago de Compostela: Consello da Cultura Galega (Ensaio & Investigación), pp. 271-292.

Jensen, Kim Ebensgaard (2014): Linguistics and the digital humanities: (computational) corpus linguistics. *MedieKultur: Journal of Media and Communication Research,* vol. 30, No 57, pp. 115-134.

Kania, Sonia (2016): 'Diachronic perspectives on varieties of Spanish pronunciation: seseo and yeísmo'. In Eva Núñez-Méndez (ed.): *Diachronic Applications of Hispanic Linguistics*. Newcastle, UK: Cambridge Scholars, pp. 200–238.

Kauffeld, Cynthia (2016): Andalusian Spanish: A diachronic survey of its origins and footprints in the Americas. In Eva Núñez-Méndez (ed.): *Diachronic Applications of Hispanic Linguistics*. Newcastle, UK: Cambridge Scholars, pp. 167–199.

Koch, Peter; Oesterreicher, Wulf (2007): *Lengua hablada en la Romania: Español, Francés, Italiano*. Versión española de Araceli López Serena. Madrid: Gredos.

Lapesa, Rafael (1957): 'Sobre el seseo y el ceceo andaluces'. In *Estructuralismo e Historia. Miscelánea Homenaje a A. Martinet* Vol. 1. La Laguna: Universidad de la Laguna, pp. 67–94.

Lapesa, Rafael (2000 [1968]): 'Sobre los orígenes y evolución del leísmo, laísmo y loísmo'. *Estudios de morfosintaxis histórica*, vol. I. Madrid: Gredos, pp. 279-310.

Lázaro Mora, Fernando A. (1999): 'La derivación apreciativa'. In Ignacio Bosque & Violeta Demonte (coords.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 4645-4682.

Leech, Geoffrey (1992): 'Corpora and theories of linguistic performance'. In Jan Svartvik (ed.): *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, pp. 105-122.

Leech, Geoffrey; Wilson, Andrew (1996): *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG-TCWG-MAC/R.

Mair, Christian (2006): 'Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora'. In Antoinette Renouf & Andrew Kehoe (eds.): *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, pp. 355–376.

Marquilhas, Rita (2000): *A Faculdade das Letras. Leitura e escrita em Portugal no séc. XVII*. Lisboa: Imprensa Nacional-Casa da Moeda.

Mayer, Charles F. (2008): 'Pre-electronic corpora'. In Anke Lüdeling & Merja Kytö (eds.): *Corpus Linguistics. An International Handbook. Volume 1*. Berlin: Walter de Gruyter, pp. 1-13.

McEnery, Tony; Xiao, Richard; Tono, Yukio (2006): *Corpus-based language studies: An advanced resource book*. London: Routledge.

Mendoza García, Eva María (2011): ‘Alianzas familiares y transmisión de oficios públicos: los escribanos de Málaga en el siglo XVII’. In Jaime Contreras Contreras & Raquel Sánchez Ibáñez (coords.): *Familias, poderes, instituciones y conflictos*. Murcia: Universidad de Murcia, pp. 141-154.

Montes Giraldo, José Joaquín (1972): ‘Funciones del diminutivo en español: ensayo de clasificación’. *Thesaurus* XXVII, pp. 71-88.

Moreno Fernández, Francisco (2004): ‘Cambios vivos en el plano fónico del español: variación dialectal’. In Rafael Cano Aguilar (coord.): *Historia de la lengua española*. Barcelona: Ariel, pp. 973-1009.

Náñez Fernández, Emilio (2006 [1973]): *El diminutivo. Historia y funciones en el español clásico y moderno*. Madrid: UAM Ediciones.

NGLE = RAE / ASALE (2009): *Nueva gramática de la lengua española. Morfología. Sintaxis*, 2 vols.. Madrid: Espasa.

NGLE = RAE / ASALE (2011): *Nueva gramática de la lengua española. Fonética y fonología*. Madrid: Espasa.

Núñez-Méndez, Eva (2021): ‘An overview of the sibilant merger and its development in Spanish’. In Eva Núñez-Méndez (ed.): *Sociolinguistic Approaches to Sibilant Variation in Spanish*. London/New York: Routledge, pp. 9-72.

Pérez Teijón, Josefina (1985): *Contribución al estudio lingüístico del siglo XVIII. Los sainetes de Juan Ignacio González del Castillo*. Salamanca: Universidad de Salamanca.

Octavio de Toledo y Huerta, Álvaro S. (2016): ‘Sin CORDE pero con red: algotras fuentes de datos’. *Revista Internacional de Lingüística Iberoamericana (RILI)* 28, pp. 19-47.

Octavio de Toledo y Huerta, Álvaro S. (2019): ‘Sintaxis de la prosa del instante: La lengua de una tradición efímera en los albores del siglo XIX’. *Anuari de Filologia. Estudis de Lingüística* 9, pp. 91-144.

Penny, Ralph (2010). *Gramática histórica del español*. Barcelona: Ariel.

Rodríguez Puente, Paula (2018): ‘En busca de lo hablado en lo escrito en los corpus diacrónicos del español: una comparativa con los corpus anglosajones’. *E-Scripta Romanica*, 5, pp. 89-127.

Rojo, Guillermo (2021): *Introducción a la lingüística de corpus en español*. London: Routledge.

Sáez Rivera, Daniel M. (2008): ‘Leísmo, laísmo y loísmo en el siglo XVIII en España: gramática y norma’. In Concepción Company Company & José Guadalupe Moreno de Alba (eds.): *Actas del VII Congreso Internacional de Historia de la Lengua Español*, I. Madrid: Arco/Libros, pp. 1087-1104.

Sánchez-Prieto Borja, Pedro; Vázquez Balonga, Delfina (2018): ‘Toledo frente a Madrid en la conformación del español moderno: el sistema pronominal átono’. *Revista De Filología Española*, 98(1), pp. 185–215.

Uritani, Nozomu; Berrueta, Aurora (1985): ‘Los diminutivos en los atlas lingüísticos españoles’. *Lingüística Española Actual* 7, pp. 203–236.

Urrutia Cárdenas, Hernán; Álvarez Álvarez, Manuela (2001): *Esquema de morfosintaxis histórica del español*. Bilbao: Universidad de Deusto.

Vaamonde, Gael (2015): ‘Distribución de leísmo, laísmo y loísmo en un corpus diacrónico epistolar’. *Res Diachronicae*, 13, pp. 58-79.

Vaamonde, Gael (2018): ‘La multidisciplinariedad en la creación de corpus históricos: El caso de Post Scriptum’. *Artnodes*, 22, pp. 118-127.