Rita Marquilhas*

Non-anachronism in the historical sociolinguistic study of Portuguese

DOI 10.1515/jhsl-2015-0013

Abstract: This paper discusses the methods that historical sociolinguists can use in order to avoid anachronism. It is argued that there are four practical ways of triggering a sense of *scale*, both for external variables that correlate with past language use and for the linguistic data that we inherited from past societies: by learning from social and cultural historians, by visiting judicial archives, by making scholarly digital editions, and by using corpus linguistic statistical procedures. The case study focused on here is Portuguese in the Early Modern period, from the sixteenth century to the early nineteenth century. The size of an ideal sample of informants is discussed, based on the demographic history of Portugal. Furthermore, social categories are established relevant in the context of Portuguese cultural history, taking into account the social world that made sense to Early Modern people. Next, I introduce a corpus of Early Modern letters containing a close-to-conversational register, and discuss two case studies. An analysis of spelling variation in the corpus shows the diachronic dialectal spread of a merger of sibilants. The statistical analysis of keywords shows the pervasiveness of register markers as well as some typical uses found in epistolary communication by social actors from different social strata.

Keywords: Early Modern Portuguese, social status, digital editions, corpus linguistics

1 Introduction

Non-anachronism is the avoidance of a chrono-centred, culture-centred bias in the interpretation of historical data. The risk of this bias is very present, for instance, among the preoccupations of cultural historians, who like to repeat what Jorge Luis Borges once wrote in his essay *Magias parciales del Quixote*: “Cervantes has created for us the poetry of seventeenth-century Spain, but neither that century nor that Spain were poetic to him” (Borges, 1964 [1952]: 193). Because historical

***Corresponding author: Rita Marquilhas,** Universidade de Lisboa, Lisbon, Portugal, E-mail: rmarquilhas@letras.ulisboa.pt

sociolinguistics is also a way of practicing history, awareness of the danger of anachronism must also characterize this discipline.

In historical sociolinguistics, the aim is to establish a possible correlation between linguistic data from the past and the social reality of the same period. However, concerning data on past language use, only fragments survive over time. Furthermore, past social categories are always difficult to grasp, given the fact that the observer does not belong to the community of speakers that she or he is hoping to analyse. The gaps between the fragments must be filled, and the precise contours of the social landscape must be reconstructed without tolerance for crude generalizations coming from the observable present (Bergs, 2012).

As stressed by other authors who have recently concerned themselves with the topic, there are two important attitudes to assume in order to prevent the danger of anachronism in historical sociolinguistics. First, we need to know when to abandon inquiries that only make sense in the light of the abundant data available to contemporary sociolinguistics, and second, we should invest far more in procuring background information about the historical social, demographic, cultural and economic matters than we would have to do if we were studying the varieties of contemporary languages (Raumolin-Brunberg, 1996; Nevalainen, 2011; Bergs, 2012). Only then can legitimate generalizations be made, in conformity with the uniformitarian principle applied to the historical study of language, which assumes that "the forces operating to produce linguistic change today are of the same kind and order of magnitude as those which operated five or ten thousand years ago" (Labov, 1972: 275).

Another way of dealing with the epistemological problem of anachronism is to adopt the *scale* metaphor, which is useful in sociolinguistics, as convincingly argued by Blommaert (2007: 3): "[T]he layered and polycentric nature of sociolinguistic phenomena should be seen as tied to differences between 'scales'[...]. [I]ntroducing this notion of scales strengthens the social-theoretical foundations of sociolinguistic analysis". In his study, Blommaert's specific concern is with social interactions as power relations, and he examines inequality problems connected to language in the globalized world. He analyses the classic case of identical words that change meaning depending on the participants involved, the case of the institutional identities that, once verbalized, prevail upon others, and the case of low-scale and up-scale language varieties, i.e. sub-standards and standards. But his larger preoccupation holds for historical sociolinguistics too: "to the extent that we intend to perform sociolinguistic studies that have a degree of sociological realism, the model of society we use should be as close as possible to the real thing" (Blommaert, 2007: 15).

Similarly as in sociolinguistics, research in historical sociolinguistics must design different ways of producing a sense of scale or a sense of perspective regarding its objects. Because such objects are detached from our immediately verifiable present, both the variables and the linguistic data that theoretically inter-correlate must be analysed and explained only when such sense of scale is triggered for them.

In this paper, I will use a concrete case, the case of Portuguese in its Early Modern period (sixteenth to early nineteenth centuries), and I will try to trigger a sense of scale both for the extra-linguistic variables that correlated with language at the time and for the linguistic data that we inherited from society in that period. The first objective is attained by collecting relevant information in the works of cultural and social historians who specialize in the study of Portuguese's Early Modern reality. The second objective demands a twofold strategy: (i) working in the archives in the search of contextualized original texts of a quasi-spontaneous nature and (ii) using digital editions with corpus linguistics methods, namely with the corpus routine of exploring keyness statistics and making comparisons with reference corpora.

2 Demographic data and social structure for the Portuguese Early Modern period

2.1 The total population and its sampling

Although it is impossible to "re-engineer pre-existing datasets" in historical research, as Hernández-Campoy and Schilling put it, the sociolinguistic approach to historical texts has, nevertheless, a number of ways to ensure that the evidence used is minimally representative. It is a question of being careful when assessing "the types of speaker, population, segments and forms of language [the] samples encompass" (Hernández-Campoy and Schilling, 2012: 64).

Insofar as the variable *population* is concerned, the sampling of historical Portuguese benefits from the fact that the history of Portugal's population has been repeatedly revisited and refined since the 1980s. Therefore, historians can tell us today, with reasonable confidence, what the changing size of the kingdom' population during the Early Modern period was. Table 1 is an excerpt from a larger chart published by the population historian Teresa Rodrigues (2009) that gives us a rough idea of how the population nearly tripled in Portugal, from just over 1 million people to a population of 3 million, over 300 years:

Table 1: Portugal: size of the population between the sixteenth and the early nineteenth centuries, based on Rodrigues (2009: 519).

Years	Mainland Portugal	Azores and Madeira Islands	Total
1530	1,120,000	no data	1,120,000
1580	1,200,000	no data	1,200,000
1620	1,200,000	no data	1,200,000
1640	1,900,000	no data	1,900,000
1700	2,050,000	125,897	2,175,897
1732	2,143,368	159,921	2,303,289
1768	2,409,698		2,409,698
1801	2,931,930	283,400	3,215,330
1821	3,026,450		3,026,450
1835	3,061,684		3,061,684

Data like these provide a basis for a historical sociolinguist to begin building samples of language informants from the past. Indeed, if we want to grasp the importance of independent historical variables (socioeconomic structures, networks of interpersonal relations, differentiated behaviours related to gender etc.) for the process of variation and change in a language in a given period, the first move has to be that of finding reliable information on *how many speakers there were* in the first place.

The fact that the study of the variation in such a population's language presupposes a correlation to social factors makes the proportionate representation of the successive chronologies mandatory. A greater number of individuals, as centuries passed, meant more speakers interacting; hence, more opportunities for language variation processes to occur and surface in written sources. A representative collection of sources is the way to grasp such possible effects, leaving the sampling refinements *by quotas* for different variables to be completed at a later stage (Milroy and Gordon, 2003: 30–32).

Labov suggested that a sample of 0.025% to 0.003% would be sufficient in a sociolinguistic survey of a speech community (Labov, 2006 [1996]: 447). Although the small size of the sample Labov assembled in his sociolinguistic survey of New York City was criticized for its lack of representativeness (Milroy and Gordon, 2003: 29), here, a percentage of 0.025% applied to different cells in the fourth column of Table 1 results in a sum of approximately 3,000 speakers (see Table 2). This is not much, given the extent of the timespan, but it means that more than 150 different speakers can be studied for each successive generation, whereas Sankoff (1980: 76) has suggested that such a limit is an acceptable one for present-day surveys: "The literature, as well as our own

Table 2: Desirable size of a speakers' sample representing proportionally the Portuguese population in the Early Modern period.

Years	Total	0.025%	Selection of speakers
1530	1,120,000	300	
1580	1,200,000	300	300
1620	1,200,000	300	300
1640	1,900,000	475	475
1700	2,175,897	544	544
1732	2,303,289	576	
1768	2,409,698	602	602
1801	3,215,330	804	
1821	3,026,450	757	757
1835	3,061,684	765	
		5,423	**2,978**

experience, would suggest that even for quite complex communities samples of more than about 150 individuals tend to be redundant".

In the fourth column of Table 2, we can see that 0.025% of Portugal's population, as determined for the years 1580, 1620, 1640, 1700, 1768 and 1820, leads to a total of 2,978 subjects. If the subjects represent different layers of society and a testimony of their speech is collected, successively, from the late sixteenth century (300), the early seventeenth (300), the late seventeenth (475), the early eighteenth (544), the late eighteenth (602), and the early nineteenth (757), then an acceptable first sample is obtained for a sociolinguistic approach to Portuguese in the Early Modern period.

The numbers in Table 2 mean that it would be necessary to track down almost 3,000 reliable Portuguese subjects of the Early Modern period for whom informal records have been kept. *Informal* here means non-literary and mostly spontaneous in nature. The need for the collection of spontaneous speech never ceases to be stressed in the historical sociolinguistics literature (cf. Weinreich *et al.*, 1968: Section 2.1). *Reliable* means randomly found, documented in the original handwriting and traceable in spatial, historical and social terms.

2.2 The social hierarchy

For the above sample to be refined by quotas into a second one, reflecting the divisions between social groups in Early Modern Portugal, it is necessary to determine what social statuses and ranks made sense in their speakers' communities.

Dividing a society into cohesive groups is a matter that involves the well-known problem of objectivity in the social sciences. In this field, there is never a sufficient zooming in on the observed subjects because the interacting social actors are themselves both the holders and the "dupes" of a sociological knowledge that the observer means to devise. Giddens (1984: 335) has put the problem in the following way:

> It is surely plain that the 'revelatory model' of natural science cannot be directly transferred to the social sciences. [...] The natural sciences can in principle demonstrate that some of the things that the lay member of society believes about the object world are false, while others are valid. It is more complicated, for better or for worse, in the social sciences.

To solve the said "complication", Giddens recalled Peter Winch's concept of *mutual knowledge*, a "knowledge which sociological observers and lay members of society hold in common" (Giddens, 1984: 336). To address such mutual knowledge is to suspend a militant scepticism regarding beliefs held by social actors. They must first be treated as descriptions, sometimes even authentic ones, of social life instead of as beliefs. The beliefs of the social actors are the necessary condition for the beginning of a social enquiry: "we cannot describe social activity at all without knowing what its constituent actors know, tacitly as well as discursively" (Giddens, 1984: 336).

When the social science in question is historical, the sharing of mutual knowledge becomes even more challenging because it involves three sides instead of two: the historian's, the empowered voices of the society under scrutiny (because their discourses were more coherent and enduring), and the disempowered voices, who left only unevenly preserved pieces of discourse that were mostly mediated by others.

Having all of these caveats in mind, we can turn to cultural historians and carefully incorporate their representations of Early Modern Portuguese social groups into our language informants' ideal sample. What was the image of the social world shared in those groups? We only have the testimony of how that image was textualized. Throughout the Middle Ages and the Early Modern times, up until the mid-eighteenth century, amid a Christian environment that was solidly set, the Western world (things and people) received its central meaning from the explanation contained in the Book of Genesis: the creation narrative (Hespanha, 2003: 72–85, 129). The order of things resulting from God's creation was used as a blueprint for the order of things in the medieval and Early Modern models of the world, where little place was reserved for the notions of individuality or free will. Inequality was blatant, but it was accepted as belonging to a balance that reflected nature's structure, which itself reflected God's perfection. The idea was not only shared by sophisticated thinkers but seems also to have been pervasive in society:

> Para além das conceções refletidas dos filósofos e dos juristas, a ideia de uma ordem objetiva e indisponível das coisas dominava o sentido da vida, as representações do mundo e da sociedade e as ações dos homens. Antes de provir de uma norma de direito formal, a ordem era um facto espontâneo da vida.
>
> [Beyond the reflexions made by philosophers and juridical thinkers, the idea of an objective, unarguable order of things dominated the sense of life, the representations of the world and society, the actions of individuals. Before being dictated by formal law, the order was a spontaneous fact of life.] (Hespanha, 2003: 76)

Such order could include the sense of a well-known division, inherited from medieval imagery, between three estates, corresponding to the three main functions men could play in society, which were the functions of fighting, praying or labouring: the nobility, the clergy and the people. Multiple sources repeat the same textualization of this sense of social division, be they preserved from eleventh-century France (i), from thirteenth-century Castille and León and fifteenth-century Portugal (ii), or from seventeenth-century France (iii):

(i) Triple then is the house of God which is thought to be one: on Earth, some pray [*orant*], others fight [*pugnant*], still others work [*laborant*]; which three are joined together and may not be torn asunder; so that on the function [*officium*] of each the works [*opera*] of the others rest, each in turn assisting all. (Year [1030], Adalberon, bishop of Laon, *Carmen ad Rotbertum regem*; English translation of the original Latin in Duby, 1980 [1978]: 5)

(ii) Defenders constitute one of the three means through which God desired the world to be sustained. For, just as those who pray to God for the people are called preachers; and those who cultivate the earth and perform the work in it, by means of which men must live and be supported, are called labourers; those on the other hand, whose duty is to protect all are called defenders. (Years [1252–1284], *Las Siete Partidas* compiled by Alfonso X, *partida* 2, title 21; English translation of the original Spanish in Burns 2001: 417; same text in 1446, *Ordenações Afonsinas* book 1, title 63)[1]

(iii) For we cannot live together in a condition of equality, but of necessity it must be that some command and others obey. [...] Some are particularly dedicated to the service of God, others to conserve the state by the arms, others to nourish and maintain it by the exercises of peace. These are our three orders or estates-general of France: the clergy, the nobility and the

1 Original Portuguese: "[D]efensores som huùs dos três estados, que Deos quis, per que se mantevesse o Mundo, ca bem assy como os que rogam polo povoo chama oradores, e aos que lavram a terra, per que os homeës ham de viver, e se manteem, som ditos manteedores, e os que ham de defender som chamados defensores", cf. Coelho (1998: 121–122).

> third estate. (Year 1610, Charles Loyseau, *Traité des ordres et simples dignités*; English translation of the original French in Lloyd, 1994: 5–7).

Such obvious intertextuality referring to a tripartite society stands more for a *topos* in Western literate culture than it stands for the description of a real-world, invariable separation between three stable groups, which could not possibly persist across such a wide time-span. Thus, a historical sociolinguistic categorization of an Early Modern society such as Portugal's should not blindly slice a sample of informants into three classes: the clergy, the nobility and the people. However, neither should it passively import the categorizations proposed for other speech communities of the same Early Modern time, such as the ones considered for Tudor and Stuart England (Nevalainen and Raumolin-Brunberg, 2003). In a world much less globalized than today's, a parallel for the English visible social climbers, for instance, barely existed in Portugal. Portuguese magistrates, military officers and public financiers could, from the sixteenth century onward, enter intermediate degrees of nobility, either by receiving knightly orders or by earning the concession of a coat of arms (Monteiro, 1997: 368), but that happened only in rare cases. According to Hespanha (2006: 121), social mobility in Early Modern Portugal was almost never seen, most unexpected and hardly ever hoped for.

More telling than the three estates imagery is the fact that political power was more pluralistic than state-centred in Early Modern Europe. Before the mid-eighteenth century in Portugal, the power of the king was that of only one among others: the church above all, as well as the local communities, the lords, the institutions – such as the university or the professionals' corporations –, the families (Hespanha, 1994). Political pluralism meant that there were several centres of punishment but also of benefice granting. Some institutions were central: the monarchy, the church, the Inquisition and the university and all their members had the privilege to be judged by a jury of their peers.

In light of the above, a first provisional social classification for the language informants, especially men, of Early Modern Portugal would include:

1. nobility (men and women)
2. clergy (priests, monks and nuns)
3. Inquisition (men)
4. military officers (men)
5. knightly orders (men)
6. university (men)
7. magistrature (men)
8. public finance (men)
9. farmers, traders and crafters (non-privileged men, judged by the ordinary justice)

A complementary set covers the individuals to whom the juridical protection of the king was granted due to the humbleness of their condition:

10. the poor (men and women)
11. the non-privileged women (i.e. related to farmers, traders and crafters, especially their widows)
12. the orphans (boys and girls)
13. the peasants (men and women)
14. the Africans and Amerindian indigenous (men and women)

A religiously influenced reasoning allowed for the humble to be both the victims of social discrimination and the beneficiaries of spiritual benevolence, including juridical favour in terms of guilt, proof, presumed innocence and good faith (Hespanha, 2003: 80).

Having said this about Early Modern Portugal's social divisions, we can now add that a refined sample of language informants for that time should reflect a division between 14 different groups. A balanced number of linguistic performances for each group would hopefully be present, granted that all of the kingdom's geographical areas with differing dialects are also represented.

A final issue closely connected to the differences between social groups is literacy. Illiteracy rates prevent the ultimate balance from being attained, obviously, although the Early Modern Portuguese (and Iberian) population was likely more literate than generally believed by social and economic historians who repeat Cressy's claim that Mediterranean Europe in the Early Modern period, with neither a protestant literacy campaign nor a complex and dynamic economy, has remained illiterate (Cressy, 1980: 179–181). Several studies contribute to the idea that there was no striking difference between Early Modern levels of literacy in Portugal, Spain or England (Rodríguez and Bennassar, 1978; Silva, 1986; Marquilhas, 2000). Nevertheless, Portuguese women were likely literate only when they were nuns, noble women or urban inhabitants. Rural women who did not belong to a social *élite* were usually illiterate (Marquilhas, 2000).

3 The data

3.1 The letters data

As argued above, for the historical sociolinguistics approach to Portuguese in the Early Modern period, a reliable sample should include almost 3,000 subjects belonging to 14 different social groups. Do any sources informing on their

contextualized spontaneous linguistic performances survive in such a high number and in a random aggregation? They do. As was investigated between 2008 and 2010 within the *CARDS, Unknown Letters* project (Marquilhas, 2012), and now within the *P.S. Post Scriptum* project that has succeeded it (Vaamonde *et al*., 2014), there are unpublished but contextualized Portuguese documents preserved in sufficient numbers to enable the composition of a historical sociolinguistic sample that comes near to the model sketched above.

The *P.S. Post Scriptum* team (as the *CARDS* team did before) locates, publishes and studies documents with linguistic interaction, viz. letters in Portuguese and Spanish. The present paper focuses on Portuguese. These documents are extraordinarily rich sources for the history of the language as spoken on an everyday basis from the sixteenth to the early nineteenth centuries. Most letters were addressed to relatives, friends, sweethearts or mere acquaintances, and were archived by the civil and ecclesiastical courts, together with documentation that contextualizes them. This documentation also sheds light on the sociological profiles of the writers and recipients, as what interested the court was the incriminatory value of the letters. The discourse used in them comes from very diverse social strata: they were written by and to men, women and even children, the powerful and the disempowered, from all parts of Portugal and its overseas territories. In some cases, we know the life stories of these people, despite their modest origins, through the testimony of defendants or witnesses.

The main ecclesiastic source for the *P.S. Post Scriptum* documentation is the archive of the Portuguese Inquisition (1532–1821). The Portuguese Inquisition was a special canonical court that received a Papal commission for prosecuting heresies, and the Portuguese Inquisitors judged, over three centuries, a total sum of nearly 40,000 proceedings (*processos*),[2] mostly concentrated on accusations of Judaism, but also covering other practices seen by the Catholic Church as being both sins and crimes (Bethencourt, 1994: 42–43). Inquisition suspects were frequently arrested with all of their assets, and whenever those assets included suspicious private correspondence, the Inquisitors regarded it as criminal proof (*Regimento* 1640, 2, 5 – see Franco and Assunção, 2004: 301). On the other hand, witnesses could also offer pieces of private correspondence, especially when they wanted to attest for the existence of a living first spouse, in bigamy cases, or when they wanted to prove superstitious or heretic propositions made by a third party.

2 In Marquilhas (2012), I wrote that the total of Inquisition *processos* was nearly 18,000, instead of 40,000. The first figure refers to Lisbon Inquisition only, while the second one of 40,000 refers to the archives of all three Inquisitions: Lisbon, Coimbra and Évora (Portugal, 2008–2015).

As for the ideal Portuguese common justice source for the documentation in question, it is the archive of the Royal Appeal Court (*Casa da Suplicação*), whose time-span was exactly that of the Portuguese Early Modern period (sixteenth century-1833). The Royal Appeal Court was an institution whose judges had the last word on crimes against social order, as established in civil law compilations (*Ordenações*), in Royal laws, and in previous court decisions. For punishing suspects, the judges had to decide in terms of equity (Hespanha, 2003: 132–134). Thorough inquiries (*inquirições*) and proof gathering (*instrumentos de prova*) were mandatory, which gave rise to a large amount of paper work lengthily described in the law compilations. It also gave rise to a precious collection of letters written by a large majority of ordinary people: people who were either writing on a criminal impulse, as is the case of extortions, menaces and insults, or were just elaborating on everyday matters, oblivious to the existing chance that their letters could become a piece of evidence to be exhibited in court. In many cases, only bad luck determined if the evidence would arrive in court: the addressee did not destroy the paper contrarily to indications received, a neighbour had success in revenge, the porter was intercepted, etc. This means that some randomness was indeed involved in specific letters finding their ways into the courts archives while others did not.

We would call these letters today, straightforwardly, *private* letters, but the concept of private is more modern than Early Modern. The next section will cover this problem.

3.2 The back stage register

The search for data of a spontaneous and conversational nature, such as sociolinguistic approaches to language prefer, puts forward the need to clarify the conception of the public versus private dichotomy in pre-modern times. As mentioned before, the centrality that the mental categories of "individual" and "will" conquered in our current modern societies cannot be uncritically transposed to the Early Modern universe of reference, and thus the term *back stage epistolary register* seems more appropriate than *private letters* to refer to the written conversational register of the *P.S. Post Scriptum* sources. An alternative term could be *graphic immediacy* (Koch, 1997) but, as will be argued below, the metaphor of the theatre appears to be the perfect tool when the researcher needs a reliable criterion to choose document sources that are likely more conversational than others, even before a detailed analysis of their contents. The metaphor of the immediacy-distance continuum is more useful for other purposes, especially for the comparison between different levels of speech that are already collected and ready for analysis (Hennig, 2011: 30).

In pre-modern times, prior to the formation of a *bourgeois public sphere* as theorized by Habermas (1989 [1962]), even if the concept of *public* did not yet exist, there were still obviously other oppositions between different spaces of interaction in society. At the very least, there was the whole process of staging that is set up everywhere at any time for communication to function. According to Goffman, this operates in two different metaphorical spaces – front and back stage – each implying the adoption of different roles by social actors (Goffman, 1990 [1959]). The front stage is a type of team performance for the benefit of a public that expects codified conventions, while back stage performance can be more relaxed and held with social actors that are trusted, in the sense that they belong to the same team. Back stage speakers can be less wordy; they can ridicule the audience; they can behave more spontaneously, more emotively, and in a less civilized manner; and they can conspire for the benefit of future performances. The back stage may also be infiltrated by members of the audience, who can move here provided they take care with their disguise.

As hinted above, these theatrical metaphors are very useful for linguists trying to uncover the conversational register of a former speech community through documents from the past. Family letters of all types are of interest of course, as the domestic space is the setting par excellence of back stage acting. Moreover, we should not forget that the model of the family in the Early Modern time was much broader than the nuclear model that took over in Western modernity, and included many more individuals than those who shared affinities of blood such as the servants of the household (Ariès and Duby, 1986).

Back stage is also the dimension for the sending of irate letters, love letters, obscene letters, anonymous letters (whether written for extortion or revenge), indiscreet letters, gossipy letters – provided that it can be established with some confidence that the author and addressee trusted each other enough to perform quite differently together on stage if necessary. In this sense, dealing with well-contextualized documentation from a situational point of view, as is the case with judicial documentation, is very useful.

The most varied inner states are documented in the *Post Scriptum* letters. Equally varied are the argumentative strategies designed to trigger, in the reader, bodily and emotional states and processes. Back stage letters were normally written to convey news and requests to the addressees, but some were sent so that addressees would feel frightened, embarrassed or distressed; they could otherwise be sent to instill hope, joy and even sexual excitement. To cover all pragmatics, the *P.S. Post Scriptum* team looks in the Early Modern courts archives for letters that correspond to the widest possible range of speech acts (Searle, 1979).

3.3 The digital edition and corpus annotation

In terms of dealing with the letters as sources, the *P.S. Post Scriptum* team makes use of all possible innovations in philology, the digital humanities in general and corpus linguistics in particular to grasp their cultural, social and linguistic significances. Consequently, the treatment of these original manuscripts has to begin with conservative transcription (semi-diplomatic), which gives rise to a genetic critical edition that reconstructs the writing process. It uses the XML-TEI universal format in terms of digital publishing (TEI Consortium 2015) and feeds a biographical database of authors and addressees whose information may be compared with the textual contents of the letters that they wrote or received and contextualized within the communicative situation in which they participated.

Using automatic tools, the word strings of the manuscripts' transcriptions are tokenized, i.e. broken up into graphemic units – tokens – aligned one-by-one with modern orthographic words, their lemmas and more abstract, morphosyntactic categories. Thus, the scholarly digital edition becomes a historical corpus, to be offered online in a fully searchable format along with images of the original manuscripts (http://ps.clul.ul.pt).

The first tool to be used by the *P.S. Post Scriptum* team to align the original spellings with standardized ones and to annotate the corpus for parts of speech was the *eDictor* tool (Paixão de Sousa *et al.*, 2013), developed at the University of Campinas, Brazil, by a team lead by Charlotte Galves. Later, because of the convenience in preserving the TEI format in a textual mark-up language and the needs for working online in the cross-search of linguistic and extra-linguistic variables, a new tool was developed: *TEITOK*, created by the computational linguist Maarten Janssen while working closely with the *Post Scriptum* team. This is an online environment that makes it possible for an annotated corpus and a TEI scholarly edition to have one and the same XML support. As described by its developer, TEITOK is:

> a web-based system for viewing, creating, and editing corpora with both rich textual mark-up and linguistic annotation. For visitors, the system provides a graphical user interface in which the annotated document can be visualized in a number of different ways, depending on what the user is interested in. And for administrators of the corpus, TEITOK uses the same interface to easily and efficiently edit the underlying XML document (Janssen, 2014).

This has been, in short, the programme of the *P.S. Post Scriptum* project, which involves linguists, historians and computational linguists and which will be completed in 2017. The sample of almost 3,000 different informants will perhaps only be finished much later, but a historical corpus of 1,000,000 Portuguese words can be ready by 2017. From the perspective of historical sociolinguistics,

the project is important because of what it might yield about the internal and external factors of variation and change in the Early Modern centuries of Portuguese language history.

4 Applying corpus linguistics methods to the data

4.1 *P.S. Post Scriptum* sample

For the moment, only partial analyses can be made on the already treated letters, i.e. letters already found, contextualized, transcribed into an XML-TEI format, indexed for extra-linguistic variables, standardized for spelling and annotated for morphosyntactic categories. The data from the early nineteenth century are much easier to detect, not only because the size of the population was greater but also because the court proceedings are best preserved. Additionally, literacy rates were likely higher: in the second half of the eighteenth century, Portugal had known a reformation of its educational system led by the Marquis of Pombal after 1759 and aimed at centralizing, standardizing and secularizing education (Maxwell, 1995: 96).

The sample comprises the text of 470 original letters (120,000 tokens) written by 470 different native speakers of Portuguese (cf. Tables 3 and 4). The subjects lived mostly in the first three decades of the nineteenth century (260); the others lived in the eighteenth (91), the seventeenth century (94) and only a few in the late sixteenth (25). More than three quarters were men (377) and less than one quarter women (93). Three quarters did not belong to the higher ranks of society (343): they were either under-privileged (167), or anonymous people (176). One quarter of the subjects were privileged (127). These latter 127 include

Table 3: Social representativeness of the sample.

Social-Economic categories	Higher ranks subjects	Lower ranks subjects	Unknown subjects	Total of subjects
Late 16th cent.	5	10	10	25
17th cent.	29	29	36	94
18th cent.	39	22	30	91
Early 19th cent.	54	106	100	260
Total	**127**	**167**	**176**	**470**

Table 4: Gender representativeness of the sample.

Gender	Men	Women	Total
Late 16th cent.	19	6	25
17th cent.	69	25	94
18th cent.	65	26	91
Early 19th cent.	224	36	260
Total	**377**	**93**	**470**

the ones with the right to receive their peers' justice and the ones that could, theoretically, move to an intermediate rank of nobility. That means, as seen in Section 2.2, nobility (men and women), clergy (men and women), inquisitors, military officers, knights, university men, magistrates and public financiers. The 167 under-privileged belonged either to the intermediate, non-nobilitating categories of farmers, traders and crafters, or to the list of humble people who were, as also seen in 2.2, the poor, the women (especially the widows), the orphans, the peasants, and the African and Amerindian indigenous. The 176 anonymous were, in most cases, seen by their contemporaries as too insignificant to be identified in terms of rank or occupation. They are nevertheless considered separately because of the hypothesis that some reference to their identities, however illustrious, was lost.

Although the sample has a low representativeness in terms of diachrony, it is nonetheless wide-ranging in terms of social diversity. Furthermore, it contains letter registers written in all of the language varieties of what would later be perceived as the four most linguistically contrastive areas: (i) the upper north and north-eastern, (ii) the lower-north, (iii) the coastal central, and (iv) the inner central and southern (Cintra, 1983 [1971]; Álvarez Pérez, 2014).

In the next section, we will see that the sample indeed seems valid in terms of dialectal evidence: its subjects show a high awareness of an important historical and dialectal feature – the merger of sibilants.

4.2 Spelling variation and sibilants mergers

The sample just described allows for the observer to detect what the main source of variation in the subjects' spellings is: the mismatch between the options of the Latin alphabet inventory, the phonology of Portuguese and the different phonetic realizations in the dialects.

Most of the sample's informants had received little or no classical education, due to their mostly non-privileged social provenance, and Portuguese

orthography before its twentieth century reform, though varied, was mostly etymological (Ministério do Interior, 1911: 3846). This means that the changes in the phonology of Portuguese that distanced it from the phonology of Latin had no stable written correspondence. No easy alignment was present between the graphemes and phonetic realizations in what concerned the new Romance palatal consonants, the affricates, the fricatives, the nasal vowels and the non-stressed vowels. The ones knowing Latin wrote a Latinized construction of the words they wanted to use. Those with limited literacy tended to turn to a phonetic usage of the Latin alphabet.

Two of the main Portuguese dialectal phonology features surfacing in the data are the *sesseio* of the lower north varieties and the *ceceio* of the central and southern ones (Cintra, 1983 [1971]). Both words, *sesseio* and *ceceio*, mean to mirror in Portuguese the popular Spanish terms *seseo* and *ceceo*. What is at stake is a merger of consonants that occurred in several mediaeval Romance areas, including parts of Portugal (Cardeira, 2003), Andalucía (Penny, 2000: 118–120) and northern France (Posner, 1996: 251–252).

In their *sesseio* or *ceceio*, some Portuguese varieties did not preserve minimal pairs of this type:

<paço> 'palace' < PALATIU- *versus* <passo> 'step' < PASSU-
<cozer> 'to cook' < COCERE (vulgar Lat.) *versus* <coser> 'to sew' < CONSUERE

The original system, in Old Portuguese, had affricates as in *paa[ts]o* 'palace' and *co[ʤ]er* 'to cook', and apical sibilants as in *pa[s̻]o* 'step', *co[z̻]er* 'to sew'. Nevertheless, this productive opposition was only kept (with simplification of the affricates) in the Portuguese upper-north and north-eastern dialects, as well as further north in Galician. In the other Portuguese varieties, the consonants underwent different mergers, either losing the laminal series [s, z] that originated in the mediaeval affricates [ts, ʤ], or losing the apical series [s̻, z̻], a process that began in the thirteenth century both from a southern focus of *ceceio* and from a northern focus of *sesseio*, according to Cardeira (2003) (Table 5).

In the *Post Scriptum* sample, letter writers show a growing tendency between the sixteenth and the nineteenth centuries to employ only two consonant letters in the intervocalic context: one for the voiced sibilant with merger, and one for the voiceless. Otherwise, they also experiment with all sorts of hypercorrections involving *s*, *ss*, *c*, *ç*, and *z*. Example (1) presents a representative case. The hand varies widely in the spelling of sibilants, with many hypercorrections, but a stable system for <c> = [s] and <z> = [z] seems also to be followed, particularly in intervocalic contexts. Even a voiced [z] resulting from a sandhi process by which final -[s] becomes [z] before a vowel initial word is spelled <z> (cf.

Table 5: The merger of sibilants in the dialects of today's European Portuguese.

	Upper-north and North-east (no merger)	**Lower north (merger in *sesseio*)**	**Coastal centre, inner centre and south (merger in *ceceio*)**
paço <PALATIU- (lat.)	laminal voiceless sibilant pa[s]o	apical voiceless sibilant pa[ṣ]o	laminal voiceless sibilant pa[s]o
passo <PASSU- (lat.)	apical voiceless sibilant pa[ṣ]o	apical voiceless sibilant pa[ṣ]o	laminal voiceless sibilant pa[s]o
cozer <COCERE (lat.)	laminal voiced sibilant co[z]er	apical voiced sibilant co[ẓ]er	laminal voiced sibilant co[z]er
coser <CONSUERE (lat.)	apical voiced sibilant co[ẓ]er	apical voiced sibilant co[ẓ]er	laminal voiced sibilant co[z]er

dellez e for *deles e* 'of-them and'). Other examples, highlighted in bold, are *coceco* for etymological *sucesso* with two intervocalic [s], and *couza milagroza* for etymological *cousa milagrosa*, both with intervocalic [z]:

(1) [...] E serto que he lastima di dizer o **coceco** / de que modo socodeo q de uma escoadra de navios não / escapou nehü nem ce calvou nehuã gemte mais que dois / omes de que eu fui um **dellez e** caimos em umas taboas e a nado / **couza milagroza** e caimos numa terra donde Chamão / coichim [...] Code PSCR1408, Letter from a braid manufacturer to the woman he wanted to marry, telling her the fake news of the death of her first husband, a sailor. Supposedly from Goa, India, but really from Lisbon, year 1634.

'Certainly, it is sad to tell you the facts, how they happened, because from a squad of ships not one was saved, nor did anyone survive apart from two men, me being one of them, and we got out on some wooden boards, and swam, miraculous thing, and got out at a place named Cochim.'

The phonetic transcription of the [+voice] and [–voiced] features, as well as the hypercorrections give a wide range of spellings in the whole collection of the *P.S Post Scriptum* letters. Random examples are *prezo* for *preso* ('prisoner'), *asim*, *acim*, *açim* for *assim* ('thus'), *faser* for *fazer* ('to do'), *çaber* for *saber* ('to know'); *peso*, *pesso*, *peco* for *peço* ('I beg'), *cenão* for *senão* ('otherwise'), and so forth. The merger that had begun in the thirteenth century and only showed timidly in the notary documents of the former centuries (Cardeira, 2003) was gaining larger and larger territory.

The diagnosis of a growing importance given by speakers of the Early Modern period to these mergers is allowed by the tool DICER: Discovery and Investigation of Character Edit Rules, developed by Alistair Baron for the statistical analysis of graphemic variants compared with extra-linguistic factors (Baron, 2011). As it is not possible to analyse the variants here in much detail, a list with the name "Portuguese Post Scriptum by eDictor" is publicly available at the DICER site (http://corpora.lancs.ac.uk/dicer/). The tool lets the observer compute and hierarchize graphemic variants, given that the input data have an xml format of the type shown below, where both the original spelling and its normalized twin are registered for each word in the manuscripts:

```
<normalised orig="prezo" auto="false">preso</normalised>
<normalised orig="asim" auto="false">assim</normalised>
<normalised orig="acim" auto="false">assim</normalised>
<normalised orig="açim" auto="false">assim</normalised>
<normalised orig="faser" auto="false">fazer</normalised>
<normalised orig="çaber" auto="false">saber</normalised>
<normalised orig="peso" auto="false">peço</normalised>
<normalised orig="pesso" auto="false">peço</normalised>
<normalised orig="peco" auto="false">peço</normalised>
<normalised orig="cenão" auto="false">senão</normalised>
```

The data from the 470 letter writers of the *P.S. Post Scriptum* sample, as analysed by the DICER tool, show that the graphemic variation that most largely betrays the merger of sibilants corresponds to the voiced sibilants in an intervocalic context: the use of <-z-> for <-s-> (e.g., *prezo* for the etymological *preso* 'prisoner'). Putting aside the sixteenth century, which is comparatively badly represented, <z> for the etymological <s> is the fifth most frequent spelling variant in the early seventeenth century, the third in the late seventeenth, the second in the early eighteenth, and the first in subsequent time-spans: the late eighteenth and the early nineteenth centuries.

Even if more data are needed to confirm this pace of the merger of sibilants in the history of Portuguese dialects, the ordinals of the DICER test can be regarded as an indication of two things. Firstly, they show how increasingly clear the merger was for speakers living in the late seventeenth century. Secondly, they reveal how badly the etymological writing for sibilants adjusts with the phonotactics of Portuguese. In this language, the medial context is indeed the major position for different consonants to occur, including the binary opposition /s/≠/z/, while the syllable final and word final positions impose several restrictions on sibilants (Mateus, 2003: 995).

4.3 Lexical analysis, genre and register markers, universes of reference

As stated above, the sample used here presents some lack of balance, given the overrepresentation of the later period (1800–1833). Nonetheless, there is a feature that all 470 subjects in the sample shared: they all practiced the same textual genre, the letter genre, with its inevitably associated register. As argued by text linguists from various theoretical backgrounds (e.g. Adam, 2003 [1992]; Fairclough, 2003; Koch and Oesterreicher, 2007 [1990]; Biber and Conrad, 2009), text type is a major variable whose values correlate with the linguistic behaviour of social actors. One such behaviour involves the use of the lexicon, particularly the use of so-called keywords.

Keyness, as Scott (2010: 43) puts it, "is a quality possessed by words, word-clusters, phrases etc., a quality which is not language-dependent but text-dependent. That is [...], words are not generally or simply key in a given language, but they may be key in a given text, or in a given set of texts, or in a given culture".

Technically, the measuring of keyness is a process that demands that the input data, however variable the spellings of its source texts, is completely standardized, so the vestiges of any diplomatic transcription cannot be kept: they have to be temporarily erased because they create noise within lexical comparison operations. This type of manipulation of the data is a strong reason for researchers in historical linguistics to have their editions in a richly annotated digital format, with rigorous alignment between diplomatic, philological and linguistic labels.

The software for the lexical analysis approach must allow for the computation of concordances, wordlists, clusters and keywords, which is the case with WordSmith Tools (Scott, 2011), which enables any consistent corpus to be powerfully searched for keyness. As for the method for measuring the keyness in a text, it demands first that the analyst obtains two lists of words with information both on the words' absolute frequencies, and on their relative frequencies. The first list of words belongs to the text or the set of texts that are the major target of the analysis; for instance, a set of letters from a given time-span. The second list of words comes from a reference corpus; for instance, a corpus of texts from the same time-span belonging to all possibly different textual genres, including letters. If a statistical test is run simultaneously on the two lists – normally the log likelihood test or the chi-square test with the *p value* threshold set to ≤ 0.05 – then a third list is obtained with the indication of the most salient words in the first list (the keywords) against the reference corpus.

Several examples of keyness evaluation can already be obtained with the *P.S. Post Scriptum* corpus, using the above-described sample of letters and its 120,000 tokens. As for the reference corpus, it contains 2.2 million words from literary works of the same Early Modern period collected by the Tycho Brahe corpus (Galves and Faria, 2010). The texts used here, coming from the Tycho Brahe collection, are the prose, including published letters, of 52 highly literate men and 3 highly literate women who lived between the sixteenth and the nineteenth century. As a matter of fact, the Tycho Brahe original corpus also includes other sub-sets – fifteenth-century literature and miscellaneous, non-literary, collections – but they both had to be excluded from the reference sample here for anachronism (and loss of control of the subjects) to be avoided.

The top key lexicon of the *Post Scriptum* sample compared to the Tycho Brahe's corpus is shown in Table 6.

Table 6: Keyness of the *P.S. Post Scriptum* lexicon compared to the Tycho Brahe reference corpus; statistical test: log likelihood; p value ≤ 0.05.

	Keyword	Freq.	Keyness		Keyword	Freq.	Keyness
1	VM 'Your Honour' / 'you' sing.	1050	6351,12	11	NÃO 'no' / 'not'	1942	393,70
2	VSa 'Your Lordship' / 'Your Ladyship'	231	1241,83	12	PRESO 'arrested' / 'prisoner'	113	368,41
3	ME 'me' accusative and dative	1342	1092,98	13	VMs 'Your Honours' / 'you' pl.	58	311,86
4	EU 'I'	1096	870,30	14	PORTADOR '(letter) carrier'	78	295,91
5	SAÚDE 'health'	219	609,10	15	DINHEIRO 'money'	156	284,20
6	LHE 'him' / 'her' dative	927	461,59	16	MANUEL (named entity: person)	148	268,41
7	MINHA 'my' fem. sing.	480	458,84	17	ADEUS 'goodbye'	94	267,43
8	JOSÉ (named entity: person and saint)	217	445,66	18	CASA 'home'	277	255,60
9	POIS 'well' / 'then'	420	419,39	19	MARIA (named entity: person and saint)	178	254,86
10	ANTÓNIO (named entity: person)	189	393,77	20	ESTIMAREI 'I-hope'	63	251,70

The scale effect that we obtain from this first lexical analysis tells us that the *Post Scriptum* documents are, above all, objects of observation belonging to a special category of texts: the texts bearing the letter genre markers and the conversational register markers. The words for first and second-person reference, both nominal and pronominal (see bold highlights in example 1), have the

higher keyness, as expected in the conversational register (Biber and Conrad, 2009).

(2) [...] **eu** ja estou milho/or ja **me** alevanto a 3 dias com toce mas es/tou mto fraca cinto **me** mto disvaida da ca/beca e trago muito fastio não poso comer / nada apetece parece **me** q aqui acabarei / a vida mas iso he o mesmo como **Vmce** ten/ha caude ainda q **eu** mora não inporta [...] Code CARDS0008, Letter from the wife of a grocer to her lover, the husband's cashier; both author and addressee were in prison, accused of having attempted to murder the woman's husband. Lisbon, year 1822.

'I am better now, I managed to get out of bed these last three days, with a cough, but I am very weak, I feel my head empty and I have no appetite, I cannot eat, nothing tempts me, it seems I will end my life here, but that doesn't matter, as long as you are healthy, even if I die, it doesn't matter.'

In second place, in terms of keyness, are the genre markers belonging to the lexicon of politeness clichés, much used in letter openers and closers: references to health, God, Saint Joseph and Saint Mary. An example:

(3) Estas faço pa saber da **saude** de vmce a ql estimarei seja / vmce perfeita em compa de minha snra Ma Thereza de-/Jezus, e de minhas snras mossas todas, e de toda a mais obri-/zação de caza dezejando, q lhe asista aquella felis **saude** e / em compa do snr **Menino Ds** da Prociunculla. A minha q / me asiste, como he agrado de **Ds**, he bóa e offereço ao dispor / do servisso de vmce; e de minha Snra Esta faço só por dar a vmce / gosto de ver as minhas Letras pa vmce se rir hü bucadinho / e mais toda a caza. [...] Code CARDS3048, Letter from a freed slave, a woman, to her former owner, a farmer. Here she was writing, exceptionally, in her own hand, while normally she dictated her letters. Brazil, Rio de Janeiro, year 1752.

'I write these to enquire about your health, which I hope is perfect, in the company of my lady Maria Teresa de Jesus and my lady girls, all of them, and the remaining household, hoping you will have that happy health in the company of the Baby Jesus of Porciúncula. My health, pleasing God, is good and I offer me to your service and that of my lady. I only write this to give you the pleasure of seeing my letters, so that you can laugh a bit and all your household too.'

The third keyness factor corresponds to contingencies from the social-historical context, the universe of reference explicitly lexicalized in the letters text. The *Post Scriptum* data were often concerned with crimes, a fact that is manifested by

the high keyness of the words *preso* ('prisoner / arrested') and *cadeia* ('jail'). For example:

(4) Snr R Conigo / Dizem os **prezos** que / estão na **cadeia** que se / acha na dita hum **pre/zo** que Dis que lheve / a Ds o diabo i os santos i a m/aria santisima [...] Prezos. CARDS2036, Anonymous accusation letter, [Coimbra], [1790–1799]

'Mr Reverend Deacon, We the prisoners who are in jail say that we have here in the same jail a prisoner who says: – To the Devil with God, the Saints and the Holy Mary [...] [Signed] Prisoners'

As for the high keyness of the form *que*, this derives from the fact that the token is mostly the pronoun in relative clauses in the reference corpus, whereas it is mostly a complementizer, a causal conjunction, or an operator in cleft sentences in the *P.S. Post Scriptum* letters.

(5) depois **q** marxei dessa sidade inda os / meos olhos se não enxugarão de me ver tão desgr/açada e lonje da ma familia **q** me parese **q** ja / não a ei de ver. CARDS0043, Letter from a woman to her beloved, a soldier who had been arrested; she herself had been banned from her city, Lisbon, because of a theft accusation. Lavre, 1832

'Since the moment that I left that town my eyes never got dry because I see myself in such a disgrace and far from my family, that it seems that I will never see them again.'

As stressed in text linguistics studies, while relative clauses can have the pervasiveness of a register marker in highly elaborate texts, especially of the more informational genre, the same does not occur in informal conversation, which conversely has the *that*-complement clauses as a marker (see Biber and Conrad, 2009: 256–259 for typologically different languages).

What these types of tests signify is that there is a possibility of obtaining descriptions that are very similar to these if lexical statistics focus on a historical corpus of other languages also consisting of letters set in a back stage of human interaction. What is specific to Portuguese is its particular forms of address, the most common Christian names, the names of the supernatural entities most venerated in the period (Catholic saints), and the functional word that is most recurrent in strategies of syntactic focalization (i.e. 'que').

The main lesson to learn from this lexical analysis of the *P. S. Post Scriptum* data has an epistemological flavour. Further research on the linguistic features of the corpus and the hypothetical connections to external factors has to adopt

only intra-genre and intra-register comparisons. The discourses of the under-privileged letter writers have to be compared either with the *élite* letter writers' discourses of the same time or with data coming from the close-to-conversational register of the under-privileged from other periods.

Given this provisional conclusion, a second line of enquiry has to be directed to the key lexicon in two different sub-sets of the *P.S. Post Scriptum* letters: the set of letters written by upper social rank individuals and the set written by the non-privileged, the humble and the anonymous (see categories in Section 2.2.). Table 7 contains the results of a keyness test in which letters from the upper ranks are the analysis target and letters from the lower ranks the reference measure.

Table 7: Keyness of the *P.S. Post Scriptum* lexicon of the upper ranks' letter writers compared to the lower ranks' letter writers; statistical test: log likelihood; p value ≤ 0.05.

	Keyword	Freq.	Keyness	N	Keyword	Freq.	Keyness
1	PATERNIDADE 'paternity'	35	43.67	10	DOMINGOS (named entity: person)	25	19.42
2	VOCÊ 'you' sing.	20	30.44	11	RAMOS (named entity: person and religious)	7	18.31
3	ILUSTRÍSSIMAS 'very illustrious'	12	24.96	12	FREITES (named entity: place)	7	18.31
4	CERTAMENTE 'certainly'	9	23.54	13	RAMALHOSA (named entity: place)	7	18.31
5	CATIVO 'captive'	16	21.95	14	CIÊNCIA 'science'	7	18.31
6	FREI 'brother' (title)	16	21.95	15	FÉNIX 'phoenix'	7	18.31
7	DE 'of'	1534	21.93	16	VARELA (named entity: person)	7	18.31
8	FURTADO (named entity: person)	8	20.92	17	MISSA Mass	12	18.26
9	PAIS (named entity: person)	14	20.07	18	SOLDADO 'soldier'	13	17.97
10	DOMINGOS (named entity: person)	25	19.42	19	FALSOS 'fake' pl.	10	16.60

Table 8 contains the result of a swapping operation by which the letters from the lower ranks and the anonymous show their keyness compared to the letters of the upper ranks.

What happens as a result of this second approach involving a smaller scale is that the level of keyness attained is much lower than the level observed in Table 6. Because of that, some forms, the ones with lower frequency, do not allow for generalizations. Such is the case of the innovation *você* ('you'), the form that

Table 8: Keyness of the *P.S. Post Scriptum* lexicon of the lower ranks' letter writers compared to the upper ranks' letter writers; statistical test: log likelihood; p value ≤ 0.05.

	Keyword	Freq.	Keyness		Keyword	Freq.	Keyness
1	VM 'Your Honour' / 'you' sing.	1019	49,98	11	MÃE 'mother'	112	14,80
2	POIS 'well' / 'then'	413	48,70	12	SENHORA 'lady'	158	14,46
3	BENTA 'holy' fem. sg.	33	20,94	13	LOGO 'later' / 'immediately'	181	13,64
4	SENHOR 'lord' / 'sir'	559	19,98	14	ESTAMOS '(we) are'	21	13,33
5	ME 'me' accusative and dative	1389	19,91	15	RECADOS 'recommendations' / 'regards'	20	12,69
6	NÓS 'we'	93	19,43	16	TERRA 'homeland'	121	12,49
7	NOS 'us' accusative and dative	153	17,77	17	FARÁ '(you sing.) will do''	46	11,73
8	NÃO 'no' / 'not'	1755	16,84	18	ADEUS 'goodbye'	51	11,69
9	LHE 'him' / 'her' dative	840	16,68	19	OFERECE '(he/she) gives'	18	11,42
10	NOVAS 'news' fem. pl.	75	15,80	20	ÁGUA 'water'	27	11,11

would become a 2nd person pronoun in twentieth-century Brazilian Portuguese, after much variation in the two former centuries (Lopes, 2006). *Você* started in the grammaticalization of *Vossa Mercê*, abbreviated *VM*, *VMce*, *VMe* in the letters. Although it appears as a keyword in the *Post Scriptum* upper ranks letters, it actually occurs only 20 times, always in the same irate letter, a context that could trigger atypical politeness strategies:

(6) [...] Vamos ao seu cazo, sim vou fallar com / **voce**: De que razão he munido para dizer a / Estevão: Ainda **voce** não tem vergonha de /aqui aparecer? [...] Code CARDS0164, Letter from a Navy officer to the enemy of his brother, a judge. Lisbon, year 1818.

'Let's take your case. Yes, I am talking to you: What reason do you have to say to Estêvão: "Have you no shame in showing up here?"

Nevertheless, other general observations can be advanced because the genre and register markers, while becoming much less contrastive, do not disappear completely. Only this time they do seem to correlate to the different uses of letters by socially contrasting subjects. It seems, for instance, that a new discourse marker linked to the conversational register, *pois* ('well', 'right'), was appearing in the lower ranks letters, accompanying the older values of an adverb ('then') or a causal conjunction ('because') (Costa, 2014):

(7) **[...] pois** se tu me / não vales peLo amor deos eu is/tou perdido eu pesote pelo amor / deos q te Lenbres q eu sou teu mar/ido **pois** torna a pedir a Senhora / q pedia por mim que me valha / pelo amor deos **pois** não tenho /quem me valha. Code CARDS0033, Letter from a tailor to his wife, a former housemaid, asking her to find the help of powerful people who would succeed in freeing him from prison. Lisbon, year [1791].

'Then, if you don't help me, for God's sake, I will be lost. I ask you, for God's sake, remember I am your husband. Well, ask the lady again, the one who interceded for me, for God's sake, because I have no one who will help me.'

In contemporary times, the form *pois* evolved into a phatic marker ('right') which is very productive in today's spoken European Portuguese (Lima, 2002). This also means that *pois* eventually lost all links to the values of social variables. Other register markers, the indexicals for the first person reference (*eu* 'I', *nós* 'we'), are also keywords of the lower ranks' and anonymous letters. These pronouns could be even more numerous if Portuguese were not a pro-drop language, with a rich verbal morphology that allows for null subjects.

(8) [...] deos queira que estes ditos nam / pasem a mais cre meu joze que se **eu** souver / alguma couza de maior sepozisam que ainda / que a xuva munta seja **eu** ei de ser a mesma / carta [...]. Code CARDS2167, Letter from a woman to her lover, whom she feared would be prosecuted by the Inquisition. Viana, Ponte da Barca, year [1774].

'God willing, these rumors are only that, [rumors]. Believe me, my José, if I learn something more certain, even if the rain is heavy, I myself will be the letter to you.'

The upper ranks' letters have high keyness in the lexicon for named entities: persons and places:

(9) [...]A compra q VS fes da Naveta foi mui boa e asertada, em **Madrid** folgarão mto de saber q ella estava na **Ilha Terceira**, tanto q **franco de Luca** mo escreveo cõ grande alvoroço. e mais a festejarião depois de chegada a **Lxa**, a / nova da perda de **Ormus** [...]. Code PSCR1259, Letter from an aristocrat, sent from abroad to his brother. Italy, Naples, year 1623.

'You buying the boat was a good and sensible choice. In Madrid, everyone was happy to learn the boat was in Terceira Island, so much so that Francisco de Lucena wrote to me in much agitation. And more they would celebrate after the arrival to Lisbon. As for losing Ormuz, [...]"

The above examples, (8) and (9), mean to illustrate the different uses that the *élite* and the non-privileged made of letters in Portugal's Early Modern times. The *élite* individuals were accustomed to such a practice (letter writing) and made it an extension of daily interaction, including using it for communicating gossip and general comments on others. The keyness of named entities comes in their texts from there. As for the under-privileged, they only used letters in exceptional occasions, either because they needed help, or because they had extraordinary news or wanted to share secrets that were too heavy to be borne. Regardless, the reference for the proper contextualization of the sender or senders ('I' / 'we') was imperative.

5 Conclusion

To avoid anachronism is one of the major challenges of historical sociolinguistics (cf. Raumolin-Brunberg, 1996; Nevalainen, 2011; Bergs, 2012). In this paper, different methods were used to trigger a sense of *scale* (Blommaert, 2007), both for extra-linguistic variables and for linguistic data from Portugal's Early Modern period. I attempted to show, in practice, how anachronism can be avoided by taking into account the results obtained within social and cultural history, by visiting judicial archives, by making scholarly digital editions, and by using corpus linguistics methods to obtain statistical results from bodies of textual data.

As to the representativeness of the historical data, an important question is what an appropriate sample of informants would be that can provide us with a reliable idea of the language as spoken on an everyday basis from the sixteenth to early nineteenth centuries, in the Portuguese society at large, and taking into account its social structure? I have argued that such a sample should mirror the fact that Portugal's population went from around 1 million to 3 million people between 1530 and 1835. The sum of its parts – 6 successive generations of speakers – should approximate 3,000 people (0.025% of the population). Furthermore, the sample should be randomly collected, representing the linguistic performances of all social strata. We can retrieve the utterances produced by such a sample of informants by scrutinizing the archives of the Portuguese Early Modern judicial institutions, especially the Inquisition archives and the Royal Appeal Court archives, which hold the testimony of those performances. These archives contain thousands of letters that are representative of Portuguese language variation in time, space and society during the Early Modern period. The letters were used as exhibits in court, and so many of them were archived

with contextualizing documentation. There is a guarantee that some degree of randomness characterizes this documentation because fate also played a part in some Early Modern subjects being summoned by the Inquisition or the civil courts instead of others.

In order to ensure that researchers can analyse their contents, replicate analyses and pursue their own hypotheses, the sources should be digitized. If the protocols for both scholarly digital editions and linguistically annotated corpora are followed, the researcher can use a number of statistical tools that will turn the data into sound historical linguistic evidence.

In the second part of the paper, I have shown how the empirical evidence, once available and searchable, can enhance our understanding of the historical sociolinguistics of Portuguese. In the case studies reported here, the socially balanced evidence revealed new data on i) the diachrony of sound changes as evidenced by spelling variation, particularly regarding the merger of sibilants in certain dialects, ii) the overwhelming effect of register and genre markers in historical data and iii) the connections that can be found between the lexicon of historical sources, on the one hand, and the social practices of different social layers, on the other.

Funding: This work has been funded by the European Research Council, ERC Advanced Grant 2011, Grant Agreement 295562.

References

Adam, Jean-Michel. 2003 [1992]. *Les textes: Types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris: Nathan.

Álvarez Pérez, Xosé Afonso. 2014. European Portuguese dialectal features: A comparison with Cintra's proposal. *Journal of Portuguese Linguistics* 13(1): 29–62.

Ariès, Philippe & Georges Duby (eds.). 1986. *Histoire de la vie privée*, 3: De la Renaissance aux Lumières. Paris: Seuil.

Baron, Alistair. 2011. *Dealing with spelling variation in early modern English texts*. Lancaster UK, Lancaster University PhD thesis.

Bergs, Alexander. 2012. The uniformitarian principle and the risk of anachronisms in language and social history. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The handbook of historical sociolinguistics*, 80–98. Chichester, West Sussex, UK & Malden, MA: Wiley-Blackwell.

Bethencourt, Francisco. 1994. *História das Inquisições. Portugal, Espanha, Itália*. Lisboa: Círculo de Leitores.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge, UK & New York: Cambridge University Press.

Blommaert, Jan. 2007. Sociolinguistics scales. *Intercultural Pragmatics* 4(1). 1–19.

Borges, Jorge Luis 1964 [1952]. Partial magic in the Quixote. In *Labyrinths: Selected stories & other writings*, 193–196. Translated by James E. Irby. New York: New Directions Publishing.

Burns, S. J., Robert I. (eds.). 2001. *Las Siete Partidas*, Volume 2: *Medieval government: The world of kings and warriors (Partida II)*. Translated by Samuel Parsons Scott. Philadelphia: University of Pennsylvania Press.

Cardeira, Esperança. 2003. Alguns dados sobre o sistema de sibilantes do português. In Ivo Castro & Inês Duarte (eds.), *Razões e emoção: miscelânea de estudos em homenagem a Maria Helena Mira Mateus* 1. 129–145. Lisboa: Imprensa Nacional–Casa da Moeda.

Cintra, Luís Filipe Lindley. 1983 [1971]. Nova proposta de classificação dos dialectos galego-portugueses. In *Estudos de dialectologia portuguesa*, 117–163. Lisboa: Sá da Costa Editora.

Coelho, Maria Helena da Cruz. 1998. Clivagens e equilíbrio da sociedade portuguesa quatrocentista. *Tempo* 3(5). 121–45.

Costa, Ana Luísa. 2014. Um Pois comentador. In João Veloso (ed.), *Textos selecionados. XXIX Encontro Nacional da Associação Portuguesa de Linguística*, 199–211. Porto: APL.

Cressy, David. 1980. *Literacy and the social order. Reading and writing in Tudor and Stuart England*. Cambridge: Cambridge University Press.

Duby, Georges. 1980. *The three orders: Feudal society imagined*. Chicago: University of Chicago Press.

Fairclough, Norman. 2003. *Analysing discourse: Textual analysis for social research*. London: Routledge.

Franco, José Eduardo & Paulo de Assunção. 2004. *As metamorfoses de um polvo: religião e política nos Regimentos da Inquisição Portuguesa (séc. XVI-XIX). Estudo introdutório e edição integral dos Regimentos da Inquisição*. Lisboa: Prefácio.

Galves, Charlotte & Pablo P. F. Faria. 2010. *Tycho Brahe parsed corpus of historical Portuguese*. http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html (accessed 1 May 2015).

Giddens, Anthony. 1984. *The constitution of society: Outline of the theory of structuration*. Berkeley: University of California Press.

Goffman, Erving. 1990. *The presentation of self in everyday life*. London: Penguin Books.

Habermas, Jürgen. 1989. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Cambridge, MA: MIT Press.

Hennig, Mathilde. 2011. The notion of 'immediacy' and 'distance'. In Mario Franco & Bernd Sieberg (eds.), *Proximidade e distância. Estudos sobre a língua e a cultura*, 15–31. Lisboa: Universidade Católica Editora.

Hernández-Campoy, Juan Manuel & Natalie Schilling. 2012. The application of the quantitative paradigm to historical sociolinguistics: Problems with the generalizability principle. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The handbook of historical sociolinguistics*, 63–79. Chichester, West Sussex, UK & Malden, MA: Wiley-Blackwell.

Hespanha, António Manuel. 1994. *As vésperas do Leviathan: instituições e poder político*. Portugal – séc. XVII. Coimbra: Almedina.

Hespanha, António Manuel. 2003. *Cultura jurídica europeia: síntese de um milénio*. 3rd edn. Mem Martins: Publicações Europa-América.

Hespanha, António Manuel. 2006. A mobilidade social na sociedade de Antigo Regime. *Tempo* 11(21). 121–143.

Janssen, Maarten. 2014. TEITOK – a Tokenized TEI environment. http://alfclul.clul.ul.pt/teitok/site/index.php (accessed 1 May 2015).

Koch, Peter. 1997. Orality in literate cultures. In Clotilde Pontecorvo (ed.), *Writing development. An interdisciplinary view*, 149–171. Amsterdam & Philadelphia: John Benjamins.

Koch, Peter & Wulf Oesterreicher. 2007 [1990]. *Lengua hablada en la Romania: español, francés, italiano*. 2nd edn revised. Madrid: Editorial Gredos.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William. 2006 [1966]. *The social stratification of English in New York City*. 2nd edn. Washington: Center for Applied Linguistics.
Lima, J. P. 2002. Grammaticalization, subjectification and the origin of phatic markers. In Ilse Wischer & Gabriele Diewald (eds.), *New Reflections on grammaticalization*. Amsterdam: John Benjamins.
Lloyd, Howell A. 1994. *A treatise of orders and plain dignities*. Cambridge: Cambridge University Press.
Lopes, Célia. 2006. Correlações histórico-sociais e lingüístico-discursivas das formas de tratamento em textos escritos no Brasil – séculos XVIII e XIX. In Guiomar Ciapuscio, Konstanze Jungbluth, Dorothee Kaiser, & Célia Lopes (eds.), *Sincronía y diacronía de tradiciones discursivas en Latinoamérica*, 187–214. Madrid & Frankfurt: Iberoamericana/Vervuert.
Marquilhas, Rita. 2000. *A faculdade das letras. Leitura e escrita em Portugal no século XVII*. Lisboa: Imprensa Nacional-Casa da Moeda.
Marquilhas, Rita. 2012. A historical digital archive of Portuguese letters. In Marina Dossena & Gabriella Del Lungo Camiciotti (eds.), *Letter writing in late modern Europe*, 31–43. Amsterdam: John Benjamins.
Mateus, Maria Helena Mira. 2003. Fonologia. In Maria Helena Mira Mateus, Ana Maria Brito, Inês Duarte & Isabel Hub Faria (eds.), *Gramática da língua portuguesa*, 988–1033. Lisboa: Caminho.
Maxwell, Kenneth. 1995. *Pombal, paradox of the Enlightenment*. Cambridge: Cambridge University Press.
Milroy, Lesley & Matthew J. Gordon. 2003. *Sociolinguistics: Method and interpretation*. Malden, MA: Blackwell.
Ministério do Interior. 1911. Relatório da Comissão nomeada, por Portaria de 15 de Fevereiro de 1911, para fixar as bases da ortografia que deve ser adoptada nas Escolas e nos Documentos oficiais e outras publicações feitas por conta do Estado. *Diário do Governo* 213. 3845–3851.
Monteiro, Nuno Gonçalo. 1997. Elites locais e mobilidade social em Portugal nos finais do Antigo Regime. *Análise Social* 32(141). 335–368.
Nevalainen, Terttu. 2011. Historical Sociolinguistics. In Ruth Wodak, Barbara Johnstone & Paul Kerswill (eds.), *The SAGE handbook of sociolinguistics*, 279–295. Los Angeles: SAGE.
Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England*. London & New York: Longman.
Paixão de Sousa, Maria Clara, Fabio Kepler & Pablo Picasso Feliciano de Faria. 2013. ***E-DICTOR***. Version 1.0 beta 10, 2013. http://edictor.net/download (accessed 1 May 2015).
Penny, Ralph J. 2000. *Variation and change in Spanish*. Cambridge & New York: Cambridge University Press.
Portugal. 2008–2015. *Arquivo Nacional Torre do Tombo*. http://digitarq.arquivos.pt/ (accessed 1 May 2015).
Posner, Rebecca. 1996. *The Romance languages*. Cambridge & New York: Cambridge University Press.
Raumolin-Brunberg, Helena. 1996. Historical Sociolinguistics. In Terttu Nevalainen & Helena Raumolin-Brunberg (eds.), *Sociolinguistics and language history: Studies based on the Corpus of Early English Correspondence*, 11–37. Amsterdam: Rodopi.

Rodrigues, Teresa Ferreira (ed.). 2009. *História da população portuguesa: das longas permanências à conquista da modernidade*. Porto: CEPESE & Edições Afrontamento.

Rodríguez, Marie-Christine & Bennassar, Bartolomé. 1978. Signatures et niveau culturels des témoins et accusés dans les procès d'inquisition du ressort du Tribunal de Tolède (1525–1817) et du ressort du Tribunal de Cordoue (1595–1632). *Cahiers du Monde Hispanique et Lusobrésilien* 31. 19–46.

Sankoff, Gillian. 1980. A quantitative paradigm for the study of communicative competence. In Gillian Sankoff (ed.), *The social life of language*, 47–79. Philadelphia: University of Pennsylvania Press.

Scott, Mike. 2010. Problems in investigating keyness, or clearing the undergrowth and marking out trails….In Marina Bondi & Mike Scott (eds.), *Keyness in texts*, 43–57. Amsterdam & Philadelphia: John Benjamins.

Scott, Mike. 2011. *WordSmith Tools Version 6*. Liverpool: Lexical Analysis Software.

Searle, John R. 1979. *Expression and meaning. Studies in the theory of speech acts*. Cambridge: Cambridge University Press.

Silva, Francisco Ribeiro da. 1986. A alfabetização no Antigo Regime. O caso do Porto e da sua região (1580–1650). *Revista da Faculdade de Letras – História, Série II* 3. 101–63.

TEI Consortium (eds.). 2015. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.8.0. http://www.tei-c.org/Guidelines/P5/ (accessed 1 May 2015).

Vaamonde, Gael. 2015. *Userguide for digital edition of texts in P.S. Post Scriptum*. http://ps.clul.ul.pt/index.php?action = home (accessed 1 May 2015).

Vaamonde, Gael, Rita Marquilhas, Ana Luísa Costa & Clara Pinto. 2014. Post Scriptum: archivo digital de escritura cotidiana. In Sagrario López Poza & Nieves Pena Sueiro (eds.), *Humanidades Digitales: desafíos, logros y perspectivas de futuro*. [Special issue]. Janus [online] Anexo 1. 473–482.

Weinreich, Uriel, William Labov & Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. Austin: University of Texas Press.