Post Scriptum: Digital Archive of Everyday Writing[1]

Gael Vaamonde, Ana Luísa Costa, Rita Marquilhas, Clara Pinto, Fernanda Pratas

CLUL, University of Lisbon

(gaelvmnd@gmail.com)

Abstract:

This paper addresses the problem of how to create compatibility between digital scholarly editions designed to support historical studies (including language history) and robust annotated corpora providing evidence for the study of language change within diachronic linguistics. This project (Post Scriptum or Project P.S.) aims to collect and publish Portuguese and Spanish personal letters written by people from all social backgrounds from the 16th to the 19th century. The documents, which are largely unpublished, were mainly kept as material evidence in civil and religious court proceedings. The paper will discuss the methodological options that resulted in the online digital edition of these Early Modern texts, as well as issues related to their automatic modernization, and linguistic annotation.

Keywords: corpus annotation, scholarly digital edition, history of the Spanish language, history of the Portuguese language, letters

1. *Introduction*

The aim of the research project *P. S. Post Scriptum: Digital Archive of Everyday Writing in Portugal and Spain in the Early Modern Period* (henceforth Post Scriptum) is to collect and publish private letters written in Portuguese and Spanish

[1] The contents of this paper were also presented in Spanish, in a much more abbreviated version, at the HDH 2013 Conference, Humanidades Digitales: desafíos, logros y perspectivas de futuro, A Coruña (Spain), 9-12 July 2013.

during the Early Modern period (from the 16th century to the first third of the 19th century) by people from different social backgrounds.[2] It started in 2012, after two previous experiments with Portuguese sources only, from the Early Modern and Modern times, namely projects CARDS, Cartas Desconhecidas (Unknown Letters) for 1500-1900, and FLY, Forgotten Letters Years 1900-1974.

The letters studied by Post Scriptum, and previously by the CARDS project,most of which are unpublished, have mainly survived because their authors came into contact with the Inquisition or the various civil and ecclesiastical courts, which used private correspondence as evidence in judicial proceedings. This documental evidence was usually supplemented by *sociological interrogations*, carried out by inquisitors and judges, and this does enable researchers to contextualize the interpersonal relationships involved. These written sources often use a form of rhetoric that is almost oral, and deal with everyday subjects that, till now, have largely remained unstudied except in isolated cases.

Thus, an electronic database is being created consisting of 7000 of these letters (3500 for each language), which can serve as a working tool for different humanistic studies, particularly in disciplines such as modern history, cultural history, textual criticism, historical linguistics and corpus linguistics. The online digital edition offers a heterogeneous collection of letters written in different social contexts for various communicative purposes. In the domain of historical linguistics, the dialogic nature of these private documents can compensate to some extent for the lack of oral sources. Firstly, the spontaneity of the interactions can offer a window onto everyday discourse (Nevalainen and Tanskanen 2007). Secondly, a repertoire of informal private letters produced by people with little formal education and written as if they were spoken, constitutes an extraordinary resource for the phonological,

[2] The Post Scriptum project is funded by the European Research Council (7FP/ERC Advanced Grant – GA 295562) and is currently under way at the University of Lisbon Centre for Linguistics (CLUL). In addition to the authors of this text, the team includes the following researchers: Guadalupe Adámez, Sandra Antunes, Catarina Carvalheiro, Tiago Castro, Elisa García, Raïssa Gillier, Mariana Gomes, Ana Leitão, Laura Martínez, Víctor Pampliega, Liliana Romão, Carmen Serrano and Leonor Tavares.

morphological, syntactic and textual study of a particular historical period. Finally, biographical data about anonymous individuals, their lifestyles and social interrelations, is an important asset not only from the historical and cultural perspective but also for modern historiography.

This paper describes the method used in Post Scriptum to digitally edit these documents and make them available online, and discusses questions related to the modernization of the texts and their subsequent processing at the levels of POS annotation, parsing, discourse annotation and keyword indexing. It focuses firstly on the process of locating and selecting the letters, and then explains the procedure used for the digital editing of the documents, involving transcription into XML format. It then goes on to describe the process of text modernization and linguistic tagging, followed by a discussion of the relationship between different editing tasks. The paper closes with some general conclusions about the project.

2. *The search for and selection of letters*

In a project that aims to construct and publish a database, the first important task is locating and selecting the data. In the task plan, the first two years were reserved almost exclusively for the search for texts and their transcription into digital format (see next section), which gives an idea of the difficulties involved.[3]

As regards data location, private correspondence dating from the 16th to 19th centuries is recovered through the *in situ* consultation of historical archives, focusing on those with an extensive repertoire of judicial and/or Inquisition sources, as mentioned above. The Spanish letters are being found by consulting document sources in the General Historical Archive (Madrid), the Simancas General Archive (Valladolid) and the Archive of the Kingdom of Galicia (La Coruña), as well as

[3] With regard to the search in Spanish archives, one letter is found for every 10 judicial proceedings scanned; in Portuguese archives, the average is one letter per 11 proceedings.

other archives in other parts of Spain (Asturias, Barcelona, Cuenca, Guadalajara, Murcia, Orense, Pontevedra and Toledo).[4] For the Portuguese letters, most of the judicial and Inquisition documentation is centralized in the National Archive of Torre do Tombo (Lisbon), though other archives have been taken into consideration, such as those in Évora, Braga and Oporto, and also in Goa (India) and Cape Verde (Western Africa).

This use of different archives from across the whole of the Iberian Peninsula – not to mention extra-European former colonies – is designed not only to increase the chances of finding letters but also to ensure geographic and linguistic representativity. A selection strategy was also adopted to control spatial and temporal variability. To ensure chronological balance, the 3500 letters selected for each language are distributed in the following manner: 500 for the 16th century (judicial documentation from this period is difficult to find in the archives); 1250 for the 17th century; 1250 for the 18th century, and 500 for the 19th century. The letters finish in 1834, representative of the administrative reforms that occurred at the end of the *Ancien Régime*.

Finally, another selection control is applied in function of the type of offence connected to the letter. Prior experience in consulting document sources has shown that judicial proceedings associated to certain offences are more likely to include missives as exhibits (for example, charges of bigamy or solicitation).[5] To avoid imbalance in this respect, archives containing many documents were approached selectively, with the consultation of a minimum number of cases for each offence.

Ultimately, the tasks of searching for and selecting the letters are oriented towards the construction of a balanced corpus that is representative of the research interests of Post Scriptum.

---

[4] In the near future, it may be possible to consult collections in archives from Granada, Seville and Zaragoza.

[5] The crime of solicitation includes any type of provocation, insinuation or seduction of the penitent by the confessor in Catholic confession.

3. *Palaeography and digital editions*

When a letter has been located, the next step is to transcribe the manuscript, converting it into a machine-readable format. This involves a series of technical and methodological decisions.

The data, textual and extratextual, is encoded using the language XML. The XML files can be read by all word processors without any loss of information, which facilitates their conversion into other formats and avoids electronic processing problems.

In order to ensure practicality and integration with other similar electronic corpora, we used the annotation standards proposed by the TEI (*Text Encoding Initiative*) consortium, an international convention that has already been consolidated for the virtual processing of primary sources. Of all the projects that follow the TEI directives, the *Digital Archive of Letters in Flanders* (DALF) stands out, as it is specifically concerned with epistolographic editions. Consequently, the schema used in Post Scriptum is the one provided by the DALF project (DALF.dtd), which still follows, for now, the TEI-P4 Guidelines, instead of TEI-P5.

A conservative attitude has been adopted for the transcription of the manuscript, giving rise to a semipalaeographic edition of the original text. That is to say, only word segmentation and the written variants of *i*, *j*, *u* and *v* are standardized.

The adoption of conservative solutions in the sources' edition is a solid principle within philology. Indeed, the textual critic, eager to dismiss all possible mismatch between his or her interpretation and the author's intention regarding the final text, follows a well-known golden rule in the philological edition of literary works; as Michele Barbi formulated it, the textual critic must establish, grounded in the oldest manuscripts, an orthographic system that can faithfully stand for the language of [the author] and his times while using today's graphical signs or, as he said in Italian in his 1907 edition of Dante's La Vita Nuova, … fissare col sussidio dei più antichi Mss. un sistema ortografico che riesca, quanto è possibile, a

rapresentarci fedelmente la lingua di Dante e dei suoi tempi coi segni grafici oggi in uso (Barbi 1907: xvi quoted in Roncaglia 1975: 89).

The principle was designed by philologists in order to ensure that literary inheritances were maintained intact and true along their cultural tradition, and it was certainly disputed, and dismissed, by the Nouvelle Critique, in terms of literary criticism (Barthes 1966). Nevertheless it still stands today as a useful principle for such manuscripts as the Post Scriptum letters, which are most of them original ones, written by common people, or by élite people in common moments of their lives; they all had to write because, by an infinite number of reasons, dialogue was impossible.

The importance of the above mentioned writings does not come from their literary quality, nor from the inferable conscience of their authors. As it happens, these are valid sources both for language historians and for cultural historians because the writings can be seen as fragments of a practice, given the meaning of the practice term in Social Theory (Bourdieu 1977, De Certeau 1984, Postill 2010). Here are the pen-and-paper artefacts manually produced by thousands of individuals who lived in some point of the Early Modern Ages in Portugal, Spain or their Empires. They gave bodily responses to everyday problems, responses that involved a wide range of uses, from the language use to the politics use, and led to all sorts of tactics, in calligraphy, spelling, epistolarity, verbal politeness, enforcement, obedience, or subtlety. Seen as this sort of artefact, it is understandable that the smallest details of the manuscripts ought to be kept by us as editors, much in the same way as archaeologists keep their material findings as intact as possible.

Accordingly, in the Post Scriptum semipalaeographic edition, aspects such as line breaks, spelling, abbreviations, deletions, corrections, tears or creases in the paper and the direction of the writing are maintained, using the XML tags defined by the TEI-P4 and DALF projects. This enables the production of an electronic edition

that is philologically rigourous, as can be seen in the following example (Figure 1), in which the tags *</lb>*, *<del>*, *<add>* and *<abbr>* mark line changes, deletions, authorial additions outside the line and abbreviations, respectively:

@@ Insert Figure 1 here.
Fragment of a transcribed letter with TEI-XML tags (PS6122, letter to a friend written by a Spanish prisoner, 1824)

The information about each manuscript is not limited to the semipalaeographic transcription of the text itself, but also includes data of an extratextual nature, such as the characteristics of the physical support (graphic layout of the text, size of paper, state of conservation) and a historical and contextual description of the letter. Whenever possible, it also provides biographical data about the authors and addressees (name, place of birth, occupation, religion, marital status, etc), which are compiled in an independent database.

4. *Linguistic research*

Research in historical linguistics is enriched within Post Scriptum by taking advantage of the progresses that have been made in the domains of corpus and computational linguistics. Thus, the annotation of linguistic type occupies an important place in the project. The intention is to discursively and morphologically annotate the whole corpus (through parts-of-speech tagging) and syntactically annotate at least a significant portion of the data.

Nowadays, there are various automatic tagging programmes available to facilitate the tasks of POS annotation and parsing, but the spelling of the input data needs to be standardized for maximum profitability and reliability. Thus, in Post

Scriptum, linguistic annotation is preceded by spelling standardization. This section will explain the methodology used in the standardization and annotation processes for both Portuguese and Spanish.

4.1 *Spelling standardization*

The original manuscripts display great variations in spelling, and a single word (eg. Spanish *vergüenza* [shame]) may be written in several different ways (eg. *berguensa*, *verguensa*, *berguenza*, *vergüenza*, *berguença*, *verguença*, etc). This diversity, which is in itself of philological and linguistic interest, is scrupulously respected in the semipalaeographic digital edition and the readers of the on-line digital edition can witness all the spelling variance phenomena. They can also confirm their calligraphic shape because a facsimile of each manuscript is almost always presented, along with the transcription. Nevertheless, such variance becomes counterproductive for the purpose of autonomatic annotation. Hence, a modernized version is also required.

In the modernized edition, the spelling and accentuation are standardized, punctuation is revised and the text is divided into paragraphs. All the abbreviations have also been expanded, with the exception of forms like *etc*, *PD*, *AD*, *XPTO*.[6] However, no words of the original are changed or removed, [7] and lexical archaisms and regionalisms are conserved, though they have been highlighted with the mark *sic* to enable their recovery.

The documents where the text standardization occurs are separate from the ones created for the primary data, since the *stand-off* encoding routine is followed here (Ide 2004). Standardization makes use of the tool *eDictor* (Faria, Kepler & Sousa 2010), developed by the research group working on the Tycho Brahe project. This is a user-friendly programme that enables the original word to be selected and

[6] In the semipalaeographic edition, the extension of the abbreviations is marked with the tag <*expan*> (see Figure 1).
[7] The text in the letter's outside address, if there is one, falls outside the standardized edition.

edited in accordance with current standard written forms. An important advantage of this tool is its compatibility with the extensions TXT, XML and HTML, which enables output files to be created in any of these formats. After the text has been standardized, eDictor is also able to organize the content into data tables, which allows the changes made at each editing level (i.e. standardized forms, expanded abbreviations and lexical variants) to be easily visualized.

The drawback of this process of orthographic editing is that it has to be done manually (i.e. on a word-by-word basis). Hence, the Post Scriptum team is currently working on a method of semi-automatic processing to facilitate this task. At present, the VARD, VARiant Detector tools (Rayson et al. 2005) are being tested for Portuguese although its application to the project is at an experimental phase. The current version of VARD is VARD 2, and its results when dealing with Portuguese data are being improved with the help of a statistical tool, namely DICER, Discovery and Investigation of Character Edit Rules (Baron 2011: ch. 4). DICER creates a list of edit rules on the basis of a corpus labelled with spelling variants and their modern counterparts; based on the DICER calculations, VARD 2 can now automatically normalize the Portuguese Post Scriptum letters' variants with an F-score of 83.6 per cent and a precision of 97 per cent (Marquilhas & Hendrickx [2014]). The next step is to apply the same procedure to the Spanish language and thus have a semi-automatic way of also normalizing the spelling variation in the Spanish Post Scriptum letters.

4.2 *Linguistic annotation*

The electronic texts making up a corpus may be presented in two forms: non-annotated (i.e. in their original form) and annotated (enriched with various types of linguistic information). As already mentioned, one of the objectives of Post Scriptum is to create an annotated corpus of everyday writing that will facilitate the recovery and analysis of all the linguistic information contained in the texts. At present, this

task includes three levels of metalinguistic reflection: Morphology (Parts Of Speech), Syntax, and Discourse. To complement them, the team also labels, by means of a set of keywords, the words that attest for those phenomena of language change (mainly sound change) that are not retrievable otherwise.

In all cases, it is the modernized edition of the texts that is annotated, although the linking with the palaeographic edition is to be recoverable, word by word, due to the XML format of the modernization process explained in the previous section.

The annotation tools can differ in accordance with the language of input and for POS annotation and parsing the process is always semi-automatic, i.e. the programme automatically produces a tagged text, which is then manually checked for possible errors by a team of linguists. As for discursive annotation and keyword indexing, the two processes are still manual, although the eDictor interface makes this task a considerably light one.

4.2.1 *POS Annotation*

At the first level of morphosyntactic annotation, the Portuguese part of the corpus makes use of the *eDictor* programme, which includes a morphological tagger, while the Spanish texts use the tool *FreeLing 3.0* (Padró & Stalinovsky 2012). Below is an example of each type of morphological annotation:

(1) *Isto/DEM sei-o/VB-P+CL eu/PRO hoje/ADV ,/, com/P toda/Q-F a/D-F certeza/N,/,* (Portuguese)

"This, I know it today, for sure."

Fragment of a text in Portuguese annotated with eDictor

CARDS0020, mother's letter to a son, 1827

(2) *En_en/SPS00 nuestro_nuestro/DP1MSP país_país/NCMS000 ha_haber/VAIP3S0 habido_haber/VMP00SM muchas_mucho/DI0FP0 tempestades_tempestad/NCFP000* (Spanish)

"In our country we have had many tempests."

Fragment of a text in Spanish annotated with FreeLing 3.0

PS6001, mother's letter to a son, 1736

As can be seen, eDictor gives each word in the corpus a tag representing the word class to which it belongs: eg. Portuguese *hoje* ['today'] ADV(adverb). The FreeLing analyser also permits analysis of the grammatical category: eg. Spanish *tempestades* ['storms'], lemma *tempestad*, N(noun) C(common) F(feminine) P(plural). In the case of eDictor, the tagging code is based on the manual annotation system used by the *Penn Corpora of Historical English* (Kroch, Santorini & Diertani 2010), slightly revised to make it suitable to the characteristics of Portuguese grammar.[8] As for the FreeLing analyser, tagging here is based on the EAGLES group model for the morphosyntactic tagging of lexicons and corpora for all European languages.[9]

This separate tagging of the Spanish and Portuguese parts of the corpus responds to some practical questions. It is the only way to make effective use of the technology available to support historical linguistics in these languages, while at the same time enabling cooperation with parallel projects under way elsewhere: in the universities of Lisbon and Campinas, for Portuguese (the WOChWEL and Tycho Brahe projects mentioned above) and in various Catalan universities in the case of Spanish (Sánchez-Marco et al.). Although this makes it difficult to compare data between the two languages, the tagging system used by the FreeLing POS tagger is

[8] See http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/tags.html for the complete list of tags used by the morphological analyser for eDictor [accessed 13/09/2013].

[9] See http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html#verbos for the complete list of EAGLES tags used by the morphological analyser of FreeLing [accessed 13/09/2013].

more complete and detailed than that used by eDictor; hence, it is always possible to convert the tags from the first format into the second.

4.2.2 *Syntactic Annotation*

The syntactic annotation for the Post Scriptum Portuguese corpus follows the Penn-Helsinki annotation system.

The files containing POS tagging are used as input for Dan Bikel's parser and the result is then manually corrected, using ~~the tool~~ CorpusDraw from the CorpusSearch set of tools.[10] The annotation of the Portuguese corpus is being implemented in close cooperation with the projects Tycho Brahe (Galves & Faria 2010) and WOChWEL,[11] both of them historical corpora, and follows the annotation guidelines initially created by Tycho Brahe.

The Post Scriptum corpus has raised, however, several specific problems, since we are dealing with private letters mainly written by undereducated people and, thus, with a great number of syntactic structures very close to oral speech, not commonly produced in written texts. The previous annotation guidelines did not account for most of these oral phenomena, since the Tycho Brahe corpus is mainly composed by literary work. Therefore, two different steps have been taken bearing in mind the need to keep up an alignment with Tycho Brahe and WOChWEL corpora, while still accounting for the novelties of our own corpus.

The first step has been to improve the parser's results in order to reduce the number of corrections done manually by the annotators. Using the tool Corpus Search, a series of basic coding queries have been created and then run on the output of the parser. Most of these coding queries allow us to automatically insert, remove or change elements in a syntactic tree. For instance, a verb such as Portuguese *haver*

[10] The tools are downloadable at http://corpussearch.sourceforge.net/ [accessed 26/12/2013].
[11] The WOChWEL webpage can be found here: http://alfclul.clul.ul.pt/wochwel/index.html [accessed 26/12/2013]

(*to have*) has an expletive subject which can easily be inserted automatically by means of adding a query with the following syntax:

(3) (IP* iDominates {1}HV*) AND (IP* iDominates !NP-SBJ*) add_leaf_before{1}: (NP-SBJ *exp*).

The second step has consisted in, for some cases, borrowing some of the tags used in Cordial-Sin (Carrilho 2010), such as the tag CP-D, which encodes sentences with a complementizer *que* that is considered discursive, and, for other cases, adopting a new annotation proposal. This is the case of comparative structures with *que* and *do que*, which are now being annotated in a slightly different way, with an indication of measure being marked as NP-MSR in comparatives introduced by *do que*, illustrated in (4) and as ADVP-MSR in comparatives introduced by *que* or *como*.

(4) *ter crimes mais do que NP-MSR o de triga* (Portuguese)
have-INF crimes more of_the that NP-MSR the of intrigue
"To have more crimes than the one of intrigue."
(CARDS0016, mother's letter to a son, 1827)

Another example is the annotation of structures with the exceptive *senão*, which can now be annotated in specifier position inside an NP-Acc, as in (5), or as a parenthetical phrase inside the NP-Acc as in (6):

(5) *Não me emprestaram [senão essa que i vai]* $_{NP\text{-}Acc}$ (Portuguese)
NEG me-DAT lend if-not DEM.F.SG that there goes
"They didn't lend me but this [coin] I'm sending you."
(CARDS0014, answer to a blackmail letter, 1824)

(6) *E assim não faças [nada [senão o que te digo]*$_{NP\text{-}PRN}$*]*$_{NP\text{-}Acc}$ (Portuguese)

And so not do-SBJV.PRS.2.SG nothing if-not the what you-DAT tell-IND.PRS.1.SG
"So, don't do anything but what I tell you to do."
(CARDS0018, mother's letter to a son, 1827)

Alongside with the need to correctly annotate oral phenomena, there have always been two other major priorities: fulfilling the main goal of any syntactic annotation, which is to make information easy to retrieve, and enriching annotation in a way that other corpora following the Tycho Brahe annotation guidelines could benefit from.

As for the Spanish corpus, the parsing has not started yet, although we are studying a strategy that allows for comparative searches in the two corpora, Portuguese and Spanish. The initial step is to parse the POS Spanish corpus with the FreeLing 3.0 shallow parser in order to identify syntactic contituents (NP, VP, SP, …). The parsed result will then be the basis for a manual syntactic annotation by the Penn-Helsinki system also using the CorpusSearch set of tools.

4.2.3 *Linguistic keywords*

A limited set of linguistic keywords has been specifically planned in order to complement the POS annotation and parsing of the Post Scriptum Portuguese and Spanish corpora. This keyword indexing will facilitate the retrieval of linguistic information that cannot be located within any of the other annotation levels. The current set of keywords has been reached after some experiments and a lot of debate among the team members, and along this process two main dimensions have been considered: (i) we should avoid redundancies with respect to the POS annotation and the parsing; (ii) we should not use keywords to classify linguistic phenomena that still survive in the standard variety of the language. Therefore, we have focused on: (a) old phenomena that became archaic today; (b) old phenomena that, although

absent from the standard, still survive in today's non-standard varieties. For both, the exceptions are the cases made evident by the other levels of annotation.

At this point of the project, we have around 30 keywords in the subareas of phonology, syntax and lexical semantics, both for the Portuguese and the Spanish corpora. Their selection has also obeyed to the specific purpose of not committing to any theoretical analysis, for our goal is strictly to provide information about where to find this or that interesting phenomenon. The further study and analysis of these phenomena is up to the scholars and students that use these corpora (and, thus, this information) to gather data for their own research projects.

Take, for instance, the following sentence, written in early 19th century non-standard Portuguese:

(7) *e se não te devertistes com esses quejos q(ue) lá ficarão venhão também* (Portuguese)
and if NEG 2SG had-fun.2SG with those cheeses that there stayed come too
"And if you haven't had fun with those cheeses that were there, bring them on, too."
(CARDS0011, letter by a young man to his friend, 1826)

The keyword *diphthong* has been applied here to the Portuguese word *quejos* (English *cheese*-PL), instead of an alternative that would point directly to the process of glide-loss between former (and standard) *queijos*, on the one hand, and the dialectal innovation *quejos*.

Another example is provided, taken from the Spanish data:

(8) *Da mil espresiones a todos los que tu sabes que eran de mi agardo* (Spanish)
Give thousand greetings to all the-ART.M.PL who you know that were of my pleasure
"Give a thousand greetings to all the ones you know I used to like."
(PS6031, letter by a highwayman to a friend, 1815)

Here, the word *agardo* was signalled with the keyword *complex syllable* instead of *metathesis*, a more intuitive label, perhaps. Indeed, we signal all syllables that happen to have a branching onset and a non-standard spelling with this broad keyword, which covers either 'metathesis' cases, or dissimilation, rhotacism, lateralization, loss, apart from just an impressionistic spelling, for lack of phonological conscience on the side of the letter writers.

As for the method of inserting the keywords, we have created their specific labels in the eDictor tool, which makes this level of annotation very friendly both to the person who applies it, and to the researcher who needs to retrieve this type of information from the corpora. For the latter, each keyword – or a combination of more than one, when this is the case – will be available on our website, attached to the word it applies to.

The whole set of current Post Scriptum keywords goes as below:

General: *second language acquisition*.

Phonology: *consonant system, complex syllable, dissimilation, diphtong, elision, insertion, hiatus, hiposegmentation, hypersegmentation, metathesis, sandhi, unstressed vowels system, vowel harmony*.

Syntax: *ethical dative, inflected gerund, negative concord, nominal concord, partitive, QUE complementizer, subject-verb agreement*.

Lexical Semantics: *existential TER, infinitive as imperative, negative words, possessive HAVER, QUE vs. QUEM, SER vs. ESTAR*.

4.2.4 *Discourse annotation*

The Post Scriptum linguistic annotation does not stop at the levels of morphology or syntax. It also includes the development of a discourse annotation level, which seems highly relevant in a corpus based on epistolary texts. At the discourse level, we can take letters as macro-units and further use them in the study of texts along history,

meaning *texts* as abstract, conceptual units, subject to theoretical reflection. At the same time, we can look for systematic phenomena in the thousands of linguistic options made by speakers in their concrete letter *énoncés*, and at the simultaneous social interactions practiced along their letter discourse.

In the current state of the project, the annotation of units above the sentence is being done manually, together with the spelling modernization process, by means of the eDictor tool (cf. section 4.1). The discourse labels classify the letters' internal structure: the first three are conventions coming from the DALF schema, later included in TEI-P5, *i.e.*, *opener*, *closer*, and *ps*; to these, we added *harangue* and *peroration*, two letter parts that accompanied the transmission of the epistolary model from classical rhetoric through the medieval tradition of *ars dictaminis* (Gomes 2004). As it happens, we can still recognize them in the writing practices of the Early Modern period.

In practical terms, the annotation of these textual parts has a positive effect on the other levels of corpus labeling: once isolated, they can be set as *to-be-ignored* by the automatic taggers that label the corpus for POS and syntactic categories. Indeed, the letters' formulaic parts (*opener*, *harangue*, *peroration* and *closer*) are structured on formulae and archaic expressions, so their treatment as non-anachronistic data to historical studies would give misleading results, at least in terms of diachronic syntax.

Additionally, the annotation of different text sequences can open the way to diachronic sociolinguistics studies, as well as to the pragmatics of letter writing. For instance, when we finish classifying letter writers and addressees by their social group, we will proceed to a contrastive analysis of formulae and topics present in the different letter parts.

The development of other levels of discourse annotation, as will be the case of the labeling of connective or of referential textual relations, will be the next step for the discourse annotation process.

5. *Relation between levels*

The previous sections have focused particularly on the process of editing and annotating the letters. The methodology adopted involves various tasks or work levels which follow a sequential order: the search for letters in historical archives; the transcription of the manuscript into XML; orthographic standardization, keyword indexing, and linguistic annotation. However, this is not a simple unidirectional process leading from semipalaeographic transcription to syntactic annotation, but rather a dynamic relationship between the different levels. Indeed, decisions taken at one level sometimes lead to the reformulation of others taken earlier. This is what happens, for example, with the processing of abbreviations, non-standardized contractions or punctuation.

Abbreviations, which are frequently used in everyday letter writing, represent a problem for automatic annotators, not only because they are unable to recognise the corresponding form, but also because one particular abbreviation may refer to more than one word (for example, Spanish *no.* may refer to *nuestro* [‘our’] or *número* [‘number’]; Spanish *pa.* may mean *para* [‘to’/‘for’] or *padre* [‘priest’/ ‘Father’]). In order to overcome this problem, all abbreviations are given in their expanded form in the XML transcription accompanied by the tag <*expan*>. Thus, in the semipalaeographic version, the abbreviation is kept and the content that is omitted in the original appears in brackets (*n(úmer)o*, *pa(dre)*), while in the modernized version it is expanded without any additional markings (*número*, *padre*), precisely in order to avoid errors with the automatic annotator.

As regards non-standard contractions (Spanish: *la venta destas caxas*, *en lamable compania*, *ya tescrito*; Portuguese: *deu pedir*, *pelamor de Deus*), which also appear quite frequently in the letters, these are written in full in the standardized edition of the text (Spanish: *la venta d’estas cajas*, *en l’amable compañía*, *ya t’he escrito*; Portuguese: *d’eu pedir*, *pel’amor de Deus*). However, this means that the

same number of tokens cannot be maintained in the two digital (semipalaeographic and modernized) editions of the text, something that is necessary for the word-by-word collation of the two versions on line. For this reason, the decision was taken to split the non-standardized contractions in the semipalaeographic edition (Spanish: *la venta d estas caxas*, *en l amable compania*, *ya t e escrito*; Portuguese: *d eu pedir*, *pel amor de Deus*) so that the tokens are aligned just as they are in the modernized edition. An XML tag (<note n="contraction">) leaves a record of the contraction in the semipalaeographic version (*la venta <note n="contraction">d</note> estas caxas*).

A third process that is affected by subsequent tasks is the punctuation of texts. As we have seen, the punctuation marks are corrected in the process of orthographic standardization and it is this standardized version that is used as the basis for automatic tagging. In this sense, the punctuation of the input text is important as it can facilitate the syntactic analysis of the data, which is the most complex part and which requires most revision by the researcher. For this reason, a punctuation system is used which, without disrespecting spelling norms, increases the percentage of correct parsing (for example, a tendency for short sentences).[12]

6. *Final remarks*

Post Scriptum began in the year 2012 and is funded by the European Research Council until 2017. The work carried out to date has mostly focused on the search for letters and their XML transcription, although part of the Portuguese corpus already has revised morphological annotation. The letters are currently being published in electronic format with the following information[13]:

- Visualization of the manuscript facsimile

[12] From the facsimiles of the letters, it is always possible to consult the original punctuation marks.
[13] The electronic address of Post Scriptum is http://ps.clul.ul.pt/index.php.

- XML transcription of the text (semipalaeographic edition)
- Standardized edition of the text (in the original language and in English)
- Alignment of the conservative and the standardized editions
- Morphological, syntactic, and discursive annotation
- Biographic files of participants (authors and addressees)
- Keywords (for linguistics and history)
- Contextualization

At present, any user has free access to the XML files, style sheet, DTD and POS corpus, and in the future, it should also be possible to consult the parsed corpora and extracts obtained from them. For this reason, we believe that the work done on Post Scriptum will prove useful for a broad public, ranging from non-specialists curious about the historical periods covered by the letters to historians interested in the type of sources found or linguists that wish to perform systematic searches of a corpus of almost spontaneous, annotated and unfragmented historical texts. Ultimately, what it offers is a body of data that can facilitate online research for multiple fields of study, with different electronic tools.

7. *Bibliography*

Barbi, Michele. 1907. "Introduzione". *La Vita Nuova* by Dante Aligheri. Milan: Ulrico Hoepli. XI-CCLXXXVI.

Baron, Alistair. 2011. *Dealing with spelling variation in Early Modern English texts*. University of Lancaster (PhD Dissertation).

Bourdieu, Pierre. 1977. *An Outline of the Theory of Practice*. Cambridge: Cambridge University Press (First published as *Esquisse d'une théorie de la pratique*. Geneva: Librairie Droz, 1972).

Carrilho, Ernestina. 2010. "Tools for dialect syntax: the case of CORDIAL-SIN (an annotated corpus of Portuguese dialects)". *Tools for linguistic variation* ed. by Gotzon Aurrekoetxea & Jose Luis Ormaetxea, 57-70. Bilbao: Universidad del País Vasco.

Certeau, Michel de. 1984. *The Practice of Everyday Life*. Berkeley: University of California Press (First published as *L'invention du quotidien: arts de faire*. Paris: Union Générale d'Éditions, 1980).

Faria, Pablo, Fabio Kepler & Maria Clara de Sousa. 2010. "An Integrated Tool for Annotating Historical Corpora". *Proceedings of the Fourth Linguistic Annotation Workshop*. 217-221.

Galves, Charlotte & Pablo Faria. 2010. *Tycho Brahe Parsed Corpus of Historical Portuguese*. <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html> [29/09/2013].

Gomes, Rita Costa. 2004. "Letters and Letter-writing in Fifteenth Century Portugal". *Reading, Interpreting and Historicizing: Letters as Historical Sources* ed. by Regina Schulte & Xenia Von Tippelskirch, 11–37. Florence: European University Institute.

Ide, Nancy. 2004. "Preparation and Analysis of Linguistic Corpora". *A Companion to Digital Humanities* ed. by Susan Schreibman, Ray Siemens & John Unsworth, chap. 21. Oxford: Blackwell.

Kroch, Anthony, Beatrice Santorini & Ariel Diertani. 2010. *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition. URL: <http://www.ling.upenn.edu/hist-corpora/> [29/09/2013].

Marquilhas, Rita & Iris Hendrckx. [2014]. "Manuscripts and machines: the automatic replacement of spelling variants in a Portuguese historical corpus". *International Journal of Humanities and Arts Computing* 8. [March 2014].

Nevalainen, Terttu & Sanna-Kaisa Tanskanen, eds. 2007. *Letter Writing*. Amsterdam: John Benjamins.

Padró, Lluís & Evgeny Stanilovsky. 2012. "FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2473-2479.

Postill, John. 2010. "Introduction: Theorising Media and Practice". *Theorising Media and Practice* ed. by Birgit Bräuchler & J. Postill, 1-32. Oxford & New York: Berghahn.

Rayson, Paul, Dan Archer & Nicholas Smith. 2005. *VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora*. Paper presented at Corpus Linguistics 2005, Birmingham, UK. URL: < http://eprints.lancs.ac.uk/12686/1/cl2005_vardword.pdf> [26/12/2013].

Roncaglia, Aurelio. 1975. *Prinzipi e Applicazione di Critica Testuale*. Roma: Bulzoni Editore.

Sánchez-Marco, Cristina, Gemma Boleda, Josep Maria Fontana & Judith Domingo. 2010. "Annotation and Representation of a Diachronic Corpus of Spanish". *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2713-2718.