Variants
The Journal of the European Society for Textual Scholarship

12-13 | 2016
Varia

CODEA: A "Primary" Corpus of Spanish Historical Documents

Ruth Miguel Franco and Pedro Sánchez-Prieto Borja

Electronic version
URL: http://journals.openedition.org/variants/364
DOI: 10.4000/variants.364
ISSN: 1879-6095

Publisher
European Society for Textual Scholarship

Printed version
Date of publication: 31 December 2016
Number of pages: 211-230
ISSN: 1573-3084

Electronic reference
Ruth Miguel Franco and Pedro Sánchez-Prieto Borja, « CODEA: A "Primary" Corpus of Spanish Historical Documents », *Variants* [Online], 12-13 | 2016, Online since 01 May 2017, connection on 30 April 2019. URL : http://journals.openedition.org/variants/364 ; DOI : 10.4000/variants.364

CODEA: A "Primary" Corpus of Spanish Historical Documents

Ruth Miguel Franco and Pedro Sánchez-Prieto Borja

Abstract: The aim of this article is to describe CODEA, *Corpus de Documentos Españoles Anteriores a 1700*, an online corpus containing 1,500 documents dating from the eleventh to the seventeenth centuries and to present a new development, CODEA+ 2015. These corpora have been created by GITHE (*Grupo de Investigación de Textos para la Historia del Español*). CODEA is a primary corpus, meaning that the GITHE is responsible both for the edition of the documents and the construction of the corpus. Also, the editing of the documents is carried out using scientific criteria which have been defined by Red CHARTA on the basis of extensive research. Finally, the usability of the corpus is increased by tools such as research engines and an advanced display of the results on graphics and maps, which turn the corpus into a powerful research tool on linguistic variation, both diachronic and geographic. *Keywords*: corpus, documentary editing, historical edition, palaeography, Spanish history, linguistic variation.

Textual corpora have undergone an extraordinary development since the 1980s (Lüdeling, Kytö and Kytö 2008; Renouf and Kehoe 2009), as one can see from projects such as PILEI, a macrocorpus directed by José Antonio Samper Padilla aimed to measure dialectal variation or PRESEEA, directed by Francisco Moreno, centred around sociolinguistic variation, or VARILEX project (Ueda and Ruiz Tinoco 2007–2008).[1] Nowadays it is virtually impossible to imagine historical linguistics without corpus linguistics; according to Enrique-Arias (2009, 11), we could say that corpus linguistics is not just a tool for diachronic studies, but a specific methodological approach. Developments in diachronic corpora have a great deal to do with improvements in tools for corpus exploitation, but also with corpus design, especially in the initial phase of selection of materials. Spanish diachronic corpora have taken advantage,

[1] This work has been supported by national research funds "Edition and study of documents from Toledo (XVI-XVII centuries)" (FF12009–10877, sub FILO) MICINN and "Corpus of Spanish documents before 1800: CODEA+ 2015" (FFI2012 –33,646) MINECO.

for instance, of the resources offered by on line archives such as PARES (Portal de Archivos Españoles en Red) <http://pares.mcu.es>, which displays descriptions of collections in Spanish archives as well as individual documents.

Within this context the aim of this contribution is to contextualize and describe CODEA, Corpus de Documentos Españoles Anteriores a 1700, (<http://demos.bitext.com/CODEA>), an on line corpus of editions of old Spanish documentary texts, as well as to present some of its future developments and perspectives for study and research. We would also try to underline the great importance accorded in this corpus not only to the edition of documents, but also to the establishment and improvement of editorial criteria. In the first section below we will briefly describe the CODEA corpus, its history and some of its key features. In the next section, we will deal with the theoretical requirements for corpora of historical editions and analyse how CODEA meets them before taking a closer look at the selection and edition of the documents included in CODEA. Finally, in the last section we will present some of CODEA's future developments, paying special attention to its use for research and study.

The CODEA corpus

CODEA is an on line corpus containing editions and images of 1500 historical Spanish documentary texts, ranging from the 11th to the 17th centuries (exactly, from 1097 to 1696) and kept in different archives all throughout Spain (León, Asturias, Cantabria, Castille, Basque Country, Navarre, Aragón, La Rioja, Castilla–La Mancha, Extremadura, Murcia and Andalusia). The CODEA corpus has been developed by GITHE, acronym of "Grupo de investigación de Textos para la Historia del Español". GITHE is based in the University of Alcalá, but it has a much wider geographical span as it works within the CHARTA Network, an international group focusing mainly on old documentary texts and sharing a methodological approach as well as research interests. It is worth mentioning that one of CHARTA's main concerns is the establishment of standards for edition of documentary texts that could be used by every researcher in the network. The aim of this effort is not only to improve our editions in general, but to be able to create over the

years a great corpus of texts edited in a similar way by many different research groups all over the word. In this corpus, Spanish documents from Latin American archives are included, which are very interesting texts that nevertheless pose particular problems.

CODEA offers a triple view of the document: a facsimile of the original document, a palaeographical transcription and a critical presentation, carried out according to the CHARTA Network's editorial criteria for historical texts, which will be commented on later. Up till now, CODEA has proved useful for linguistic research, mainly because of the reliable edition of the texts, conducted according to sound scientific standards. CODEA was born around 1998, but it was not planned as an on line, independent corpus: originally, it was only a collection of documentary texts to be included in the CORDE, Diachronic Corpus of Spanish Language (http://corpus.rae.es/CORDEnet.html) of the Real Academia Española. Later on, a team of former diachronic linguistics students of the Universidad de Alcalá prepared some editions of documents. These editions became the first volume of *Textos para la Historia del Español*, which became a series, with nine volumes published to date (Sánchez-Prieto Borja 1991–2014). As years have passed, other philologists and linguists, as well as researchers in different fields — such as experts in computer science — joined the group.

At present, the CODEA team keeps working on new documents and reviewing the old ones. We believe that some of the outcome of this effort can be shared, so that different researchers that face the challenges of editing historical materials can benefit from them.

Theoretical and practical requirements for historical corpora

The foundations of the work with historical corpora rest on the idea that studies in diachronic linguistics (as well as other historical studies, like palaeography) must always be grounded in a corpus of dated texts, which must include legal and administrative documents. Documentary texts provide a number of advantages over literary texts: we usually know the date and place when and where they were written; they are written in a non-literary style and include fragments that could be close to the common language and, last but not least, they have not been altered, neither linguistically nor in their redaction, by centuries of manuscript transmission.

But such a corpus must meet certain requirements. In the first place, it must be a primary corpus: with primary corpus we intend a corpus made up from texts directly edited by the corpus researchers, not former editions by other authors. This is, a same research group takes care of searching for the texts, selecting and editing them. We arrive thus to the second requirement for a quality corpus: the corpus texts have to be transcribed and edited according to previously set scientific criteria. This means that all the text included is homogeneous and do not show any differences due to the editor's decisions. Also, the corpus must contain a large number of texts in order to have a representative sample. Finally, the corpus must include not only texts and search engines, but also quantitative and qualitative analysis about topics such as the space-time distribution of linguistic variants.

To sum it up, we assume that only a corpus produced out of documentary texts directly selected from archives and transcribed and edited by scholars within the same corpus can fulfil the requirements of representativeness, reliability, quotability and retrievability. In this respect, we would like to underline that textual scholarship and corpus linguistics, specially when dealing with diachronic corpora, are closely linked, since the quality of the corpus is based on the quality of the editions that compose it. We have taken into account these requirements when designing CODEA, so that the corpus could be said to meet the aforementioned standards.

In the first place, in its actual state, CODEA can be said to be a representative sample, because it comprises a wide geographical, chronological and sociolinguistic range of texts. Also, it is reliable, because these texts are purpose built editions, carried out according to scientific and philological criteria. Texts in CODEA follow the editorial criteria of the International CHARTA Network and have been widely applied and amply recognized as useful; nevertheless, they are not static but develop progressively to meet the new challenges that editors face when working with different texts. Therefore, these editing standards, based on decades of research (Sánchez-Prieto Borja, 1998), have just been published in a new and improved version (Sanchez-Prieto Borja 2011a). The new criteria of CHARTA are more precise in aspects that were formerly left at the discretion of editors. For instance, now these criteria provide editors with guidelines to face material accidents in the texts, such as

deletions or corrections or to indicate different hands in the copy of the documents.[2] These norms are particularly useful to correct, in the critical presentation, minor copy mistakes, such as syllable repetition (for instance, *cartarta* for *carta* in a document from the Royal Chancery).

Furthermore, the accuracy of these editions can be easily checked, since a facsimile of the original document is also available within the same corpus. Finally, one can see the quotability and retrievability of CODEA in the fact that it provides free and universal access to all of its contents, which can also be downloaded.

Selection and edition of CODEA documents

Let us now take a closer look at some of these concepts. As for representativeness, a much larger number of texts and documents are easily available nowadays than it was in the past, thanks to the digitalization and on line availability of some archives. But this fact make us face the need of selecting materials for our corpora. There are two main methodological approaches. On one hand, we have the model of macrocorpus, like *Corpus del Español* by M. Davis, the aforementioned CORDE, CHICA, *Digital Corpus of Old Catalan* for Catalan or TMILG, *Digital Medieval Treasure of Galician Language*,for Galician. On the other, there are corpus that focus on a specific textual genre like *Camoes* for Portuguese theater, *Medieval Bible* by Enrique-Arias for translations of the Bible into Spanish or the *The Anglo-Normand On-line Hub*, which displays searchable Anglo-Norman texts. These two models, the macro-corpus and the specific corpus, are complementary, but lately, the tendency in Europe is to follow the specific model, which can reach both a better definition

[2] When some letters or characters are illegible or have been lost in a damaged area of the document (stained, broken or bent parchment, for instance), the editor will include in the palaeographical transcription one asterisk for each of the missing characters, for instance: "d** vez*nos" (*two neighbours*). If they do not know the exact number of characters missing, they may use three spaced asterisks in brackets, for instance: "dos v[* * *]". Additionally, the cause of damage will be stated in the palaeographical transcription, in italics and in brackets, like: [*roto*] 'broken', [*doblez*] 'folded', [*mancha*] 'stain'. On the critical presentation, any fragments reconstructed by the editor will be displayed in angle brackets, for instance: "d<o>s vez<i>nos".

of the focal point of research and a higher quality in displayed materials.

Corpora can be specific regarding their contents, but also in their aims. Most diachronic corpora are planned for linguistic study, even for a very specific purpose, like CHICA, which is the first step towards a historical grammar of Catalan, or CORDE, which will provide data for the new historical dictionary of Spanish language (Pascual and Domínguez, 2009). Unfortunately, corpora built by historians are rather rare, but we could mention the excellent *Codice Diplomatico della Lombardia Medievale*, by the Università di Pavia.

In this respect, we should note that large corpora are very often "secondary", this is, built up from available editions of each text, most of the times of uneven quality and based on different editorial criteria. But even specific corpora are sometimes constructed used previous editions, like the *Corpus of Middle English Prose and Verse*, which includes editions from the nineteenth century.

Most of these corpora are different as well because of the way in which texts are displayed, this is, the final aspect of the edition on the screen; the choices have an impact on their reliability and use for study. Broadly speaking we could say that there are not noticeable tendencies in this respect in English, French or Italian philology. In general, on line editions are less palaeographical and do not mark the resolving of abbreviations (like the *Anglo-Norman on-line Hub*). But sometimes on line text can be reproduced almost exactly the original script, like *Corpus of Middle English Prose and Verse,* which keeps both *þ* (thorn) and *th*. So, different corpora choose to show different bits of the information contained in the texts.

Since it is by all means impossible to enclose all the information in the original document in just one edition, the CODEA team, following the CHARTA standards, has decided to draft and display multiple editions of documents. Aware of the actual prospects of CODEA, we have made some sensible choices among the different options that the multi-edition furnishes us with: the documents are presented on line in a triple visualization. First, a palaeographical transcription which is both accurate and manageable, since it does not include any special characters, as we can see, for example, in document 259, Archivo Histórico National, Clero, Palencia, folder 1657, number 13 (dated in 1257):

> [1] Esta es la pesquisa q*ue* mando fazer do*n* ferrant gonçalez de Sojas m*er*jno mayor de Castiella. a pelay diaz de forna alcalde del Rey & a gut*ier* yuan*n*es de fresno pesquiridor del Rey por demanda [2] q*ue* demandaua el concejo de ag*ui*lar alos de valuereçoso q*ue* son solariegos del abb*a*t de ag*ui*lar

Secondly, a critical presentation edited according to the before mentioned criteria: graphical features with no phonetic relevance are standardized and the text presents the necessary accents, as well as punctuation conforming to the document's syntax. In the case of the aforementioned document 259, it would be as follows:

> [1] Esta es la pesquisa que mandó fazer don Ferrant Gonçález de Sojas, merino mayor de Castiella, a Pelay Díaz de Forna, alcalde del rey, e a Gutier Ivañes de Fresno, pesquiridor del rey, por demanda [2] que demandava el concejo de Aguilar a los de Valvereçoso que son solariegos del abbat de Aguilar.
>
> [This is the enquiry that sir Ferrant González de Sojas, main judge of Castille, ordered Pelay Díaz de Forna, royal judge, and Gutier Ivañes de Fresno, royal detective, to do, through a demand by which the village of Aguilar demanded the men of Valvereçoso that life and work in the lands of the abbot of Aguilar]

Lastly, as we have already mentioned, the CODEA web displays a facsimile of the original document. The CODEA team has recently signed an agreement with the National Office for Archives and Libraries in Spain, which allows us to use digitalized images of document in national archives and upload them into the CODEA web. Up to date, images of documents from the Archivo Municipal de Toledo and Archivo Municipal de Guadalajara are displayed in our website, and soon the uploading of facsimiles from the Archivo General de Simancas y Archivo Histórico Nacional will be completed.

This is an example of the visualization of a document in the CODEA corpus; both palaeographical transcription and critical edition are displayed in parallel columns; the image opens in a new window that can be moved around the screen (Figure 1). Each one of the displays of the document in the CODEA web provides data for different kinds of research. The facsimiles of the documents can

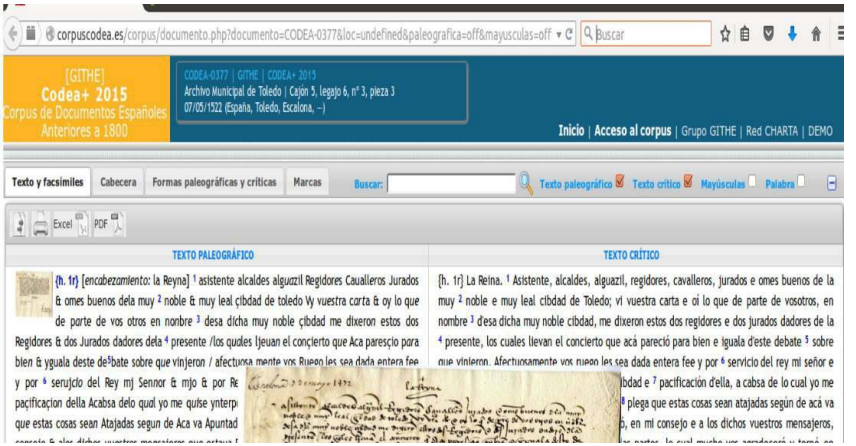

Figure 1: Screenshot CODEA: palaeographical transcription, critical edition and digital facsimile

be used for palaeographical studies. Palaeographical transcriptions of documents are useful for graphic and phonetic studies. Finally, critical presentations provide materials for syntactic and lexical studies, as well as for historical studies in general.

Future developments of CODEA

As for CODEA's future developments, we are committed to bringing the corpus to a new phase, with funding from the Ministry of Economics of Spain. First, the corpus' size will be increased, with a thousand new documents added in order to reach a total amount of 2500 by 2015. Secondly, also the chronological and geographical span of the corpus will be expanded.

The corpus will be extended chronologically, to reach from the origins of Spanish to the eighteenth century. Up to 2% of documents in Latin will be included, since it is very difficult to set boundaries between Latin and Romance in the Early Middle Ages. On the other hand, the eighteenth-century documents will be a novelty, since this century is very scarcely represented in corpora (although archives are plenty of documentary materials from this period), and also very little studied.

From a geographical point of view, CODEA will include Castilian documents from bilingual areas of the Iberian Peninsula:

Galicia, Valencia and the Basque Country. The difficulties that these documents pose have been the cause for them to be excluded from CODEA until the present date, although there are some other corpora within the CHARTA network that do work with this sort of texts; for instance, Enrique-Arias 2012 deals with problems of edition and study of Catalan-Spanish bilingual letters.

Also the sociolinguistic levels represented in the corpus will be extended. Currently, documents given by private individuals were less than 30% of the total amount and these were predominantly conveyances[3]. Previous experience shows that lexical and syntactic efficiency increases as the percentage of private documents increase. We believe that next versions of CODEA should reach at least a 43% of private citizen documentary texts, in order to get closer to real usage of language[4].

Nevertheless, the importance of legal and official documents can not be disregarded, since administrative language has played a crucial role in the shaping of modern Spanish. For instance, some textual connectors such as *en consecuencia, por tanto, por consiguiente* (having all of them meanings close to consecutive "so"), very common nowadays in spoken Spanish, sprung in administrative language and moved downwards in the sociolinguistic scale. In consequence, it is possible and productive to study the influence of the official language in diachronic change of the Spanish language, in the context of multi-causality in linguistic change.

As for the sociolinguistic side of the project, there are two very important items that will be included. On one hand, women's writing, as eighteenth century feminine epistolaries are common and available. On the other, documents written by historically socially deprived people, like travellers or Spanish *gitanos* (gypsies). We would also like to mention that the Spanish Institute of Gypsy Culture is supporting this project.

In order to identify and edit all these documents, the CODEA team will visit and explore national and city archives. Unfortunately,

[3] Percentages of documents currently in CODEA: public documents 72.77% (chancery: 20.26%; ecclesiastical: 38.77%; municipal: 6.61%; judicial: 7.11 %); private citizen documents: 27.23%.

[4] Percentages of documents in future developments of CODEA: public documents 57% (chancery: 17%; ecclesiastical: 20%; municipal: 5%; judicial: 15%); private documents: 43%.

not all the archives in which we are interested have detailed, web-compliant catalogues, so on-site work has proved necessary. It is worth mentioning the relevance of close collaboration with staff in archives, since their knowledge of the funds can point out important documents not described or poorly described in the archive inventories. This has been the case in the Archivo Municipal de Toledo, which participates in the CODEA project.

Documents from the following archives will be examined and incorporated into the corpus:

1. Archivo Histórico Nacional, specially its rich Inquisition collection
2. Archivo General de Simancas, which keeps many remarkable epistolary collections, written by kings (p. ej., Fernando II el Católico), women or even some common citizens whose cultural level was quite low
3. City archives from Andalusian capitals
4. For bilingual communities, we will examine the Archivo de la Corona de Aragón (Barcelona), as well as the great documentary stock of Valencia and Balearic Islands, Galicia and the Basque Country for Castilian and bilingual texts.

Markup or tagging of corpora is nowadays standard practice; it allows quick and detailed information retrieval (Isasi 2010; Spence, Isasi, Pierazzo and Vicente, 2012). Markup of external features of texts is most common, since it broadens the perspectives of studies conducted on them, like the *Nouveau Corpus d'Amsterdam*, for instance. Tagging of linguistic features is also usual, as one can see in the projects on European Dialect Syntax. Most corpora use TEI markup, although many textual and linguistic features are not included in the standard. Nevertheless, some centres, like the Department of Digital Humanities at King's College London, apply the TEI standard consistently to their corpora, like EPIDOC, *Epigraphic Documents in TEI XML* or *The Gascon Rolls Project*. In CODEA+ 2015, the following information will be displayed in the header of each document and encoded for future research. This header is different from the ones currently used in CODEA, which are simpler and include less external data:

GITHE *(research group)*

CODEA *(corpus)*
0274 *(number of document within the corpus)*
AMTO A.S. 602, cajón 8, legajo 1, nº 37 *(shelf mark)*
1515 julio 20 (Burgos, España) *(date and place)*
Castellano (*language)* Cancilleresco *(type of document according to issuer)* Pragmática (*type of document according to contents)* Gótica cursiva *(hand)* Autor: hombre *(gender of author)*
Pragmática de la reina doña Juana en la que prohíbe vestir telas de seda, plata y oro a excepción de las personas de la realeza (*resumée)*
Pedro de Quintana (la fize escrevir) *(author)* Papel *(material)* 265 x 160 aprox. *(measures)* Buen estado de conservación (information about conservation, particularities, etc)
Mª Jesús Torrens Álvarez *(transcriber)*
Carlos Martín Sánchez *(reviewer 1)*
Cristina Castillo Martínez *(reviewer 2)*

Table 1: Document header information in CODEA+

Development of complex search engines is essential for discovering the benefits of the corpus and the new search tools will play a very important role in CODEA. Search engines will include functionalities such as search by lemma (for instance, when searching one verb the search results will include all forms, regular and irregular, including old morphological variants), form, high or low frequencies and lexical bundles within the document. These engines will also be able to search the information included in the head: chancery, archive, date and place and so on. Users of CODEA will be able to visualize the results of these searches built into maps, tables or statistical graphics.

The functionality of these engines greatly improves when searching a completely lemmatized text. It will be the critical presentation the one to be lemmatized, since edition criteria have been devised to disambiguate homographies (*en/én; al/ál; y/ý*). Since automatic lemmatization usually does not go beyond 95% of text forms, we will go for interactive lemmatization of the whole text to reach 100% of forms, with a methodology already successfully carried out by Horcajada (Universidad Complutense de Madrid) and Ueda (University of Tokyo; Ueda and Perea 2010).

CODEA for research and study

To date, the 1500 documents of CODEA have been the basis of several studies, conferences and scientific articles, and their results are a great feedback for the corpus, since they contribute to chart future decisions and develop further applications. We will shortly comment on only two of these works; a complete bibliographical list of the research that has been carried out up till now is available in the GITHE website (e.g. Pato and Felíu Arquiona 2005 or Ueda, in press).

First, CODEA allows to carry on studies on historical geographical linguistics, meaning the mapping of graphical, graphic-phonetic, morphological, lexical and syntactic forms in different time periods. In this line, Sánchez-Prieto Borja (2011b) has studied the geographical distribution of words meaning "plot of land" in the Iberian Peninsula the from thirteenth to the fifteenth centuries. Geographic data are crucial for research in historical linguistics; CODEA+ 2015 will provide us with the graphical distribution onto a map of the search engine results for a certain linguistic feature. Such maps enable the researcher to see the different areas in the development of Spanish language, and to place different phenomena — which in traditional studies were thought to be general — within their original areas. For instance, the documents show that indefinite pronoun *algún-alguno,* generally considered to be common to all the Castilian area, seems instead to have its roots in the Leonese area, in North West Spain, and to have spread Eastward and Southward.

Secondly, a very innovative research concept is the use of CODEA for automatic dating of undated documents (they represented up

to 6% of the CODEA collection in 2012). Our research team, together with the University of Tokyo, is currently developing a computer tool for interactive dating of undated documents. Researchers Ueda and Kawasaki, from University of Tokyo have written a complex program to date these documentary pieces with a very small margin of error (Kawasaki 2014). In the first phase of the work, it was necessary to establish a large number of linguistic and non linguistic parameters, which range from hands or seals to syntax or notarial formulae. The tests conducted so far show that not only noticeable phonetic of morphosyntactic features are relevant for dating a text, but also small graphical variation, like *saber* / *ssaber*, *aver* / *auer*. Once the parameters were fixed, they statistically measured the chronological co-occurrence of certain linguistic features in the dated documents (Díaz Moreno, Martínez Sánchez, Ramírez Luengo and Sánchez-Prieto Borja, 2015; different resources for dating undated texts are commented on in Gervers 2000).

In order to check the reliability of this method, several tests have been conducted on dated documents. For instance, when applying this dating methodology to documents such as CODEA 2011 nº 468 (León, 1464), a date ranging from 1451 to 1475 is inferred, which is perfectly coherent with the document's real writing date. In CODEA 2015 this technique will be applied to all the undated pieces of the corpus, and its results will be available for study.

Conclusion

We have tried to show how scholarly edition of texts in all its phases, from the identification and reading of exemplars to the very last graphic choices, is of the outmost importance for historical corpus linguistics. Edition of documentary texts for research and study should be based on a deep knowledge of historical linguistics and carried out according to previously set editorial criteria. Creation of primary corpora, this is, corpora built from *ad hoc* editions, is crucial for historical research and study and it is clear that information retrieval heavily depends on the way in which the texts are published. Therefore, a corpus is not only a tool for different studies, but a methodological approach to edition that be taken into account in textual scholarship. CODEA, with its triple presentation, together with accurate and sound editorial criteria, aims to provide

resources for many different studies, from graphematics to general history, which is impossible with a single edition.

And, last but not least, collaboration among textual scholars is crucial to enrich and broaden any editorial project. CODEA aims to promote, within the bounds of the International CHARTA Network, scientific exchange among research groups on Hispanic documents from many different countries, as well as to stimulate methodological transfer among groups committed to editing and studying documents in different languages.

As a conclusion, we hope that CODEA will contribute to give the Spanish research on corpus linguistics a prominent position in the international scene and hopefully contribute to improve electronic edition of documentary texts also in other languages.

Addenda

The time elapsed between the writing and the publication of the following article has been enough for the initial perspectives to come true and, for once, even before the planned deadline. The results of each search are now quantified as graphics, according to four parameters. First, a time axis; the dates of emission of the documents build a linear graphic which, at the same time, displays the distribution along the centuries of the form searched for. Second, a geographical axis, where the search results are displayed according to countries, provinces and villages. Third, a typological axis shows the number of occurrences in each documentary and diplomatic typology, context of emission and, specially, women's writing. Finally, a codicological axis indicates the distribution of the results by archive and type of writing.

A list of key words (up to ten) has been added to the heading of each document. A full list of key word for the 1,500 documents currently included in CODEA is available, so it is possible for the user to select the required key word and search for it in the whole corpus. Thus, we have created a true semantic map (or referential map) of the corpus.

However, the most interesting new feature is the extensive incorporation of tools which display on a map the location of writing of each document in which the results of a given search are found. This tool allows to search for several forms at the same time

and to visualize them with different icons on the map. In this way, CODEA+ 2015 becomes a true linguistic and diachronic atlas of Spanish (*Atlas Lingüístico Diacrónico y Dinámico del Español*, ALDIDI). In addition, it is interactive, as the user can set different parameters for each search, selecting time limits, regions, documentary type, context of emission (chancellery, ecclesiastical, private), participation of women, scribe, etc.

The advanced view of results gives an immediate and detailed idea of the weight of geographical factors in linguistic variation in the Iberian Peninsula from the first Spanish texts to the nineteenth century. The linguistic variation located and quantified in the CODEA graphics and maps not only lexical, but also phonetic, morphologic and syntactic: users can search variants such as *cosa/cossa*, *otro/otri/otre/otrie*, the geographical distribution of *venta/vención/vendición/vendimiento*, or collocations like *no ... ning*/no ... alg**)

In conclusion, CODEA+ 2015 is fully working with multi factor analysis, since the diatopic data can be combined with sociolinguistic and diplomatic data. But the CODEA team will carry on working and very soon new developments will be included, for instance, a link to the *LETRAS y NÚMEROS* programmes, established by Hiroto Ueda of the University of Tokyo.

Bibliography

Biblia Medieval, <http://www.bibliamedieval.es>.

CHARTA: *Corpus Hispánico y Americano en la Red: Textos Antiguos*, <http://www.charta.es>.CHICA: *Corpus Informatitzat del Català Antic*. <http://www.chica.cat>.

CODEA+ 2015: *Corpus de Documentos Españoles Anteriores a 1800*, <http://corpuscodea.es/>.

Codice Diplomatico della Lombardia Medievale, <http://cdlm.unipv.it>.

CORDE: *Corpus Diacrónico del Español*, <http://corpus.rae.es/CORDEnet.html>.

Corpus del Español, <http://www.corpusdelespanol.org>.

Corpus of Middle English Prose and Verse, <http://quod.lib.umich.edu/c/cme>.

Díaz Moreno, Rocío et al. 2015. "Hacia una cronología evolutiva del español". In Francisco Javier de Cos Ruiz and Mariano Franco Figueroa (eds.), Vol. 1 of *Actas del IX Congreso Internacional de*

Historia de la Lengua Española. Madrid and Frankfurt: Iberoamericana Vervuert, pp. 435–49.

Enrique-Arias, Andrés, ed. 2009. *Diacronía de las lenguas iberorrománicas: Nuevas aportaciones desde la lingüística de corpus*. Madrid and Frankfurt: Iberoamericana Vervuert.

——. 2012. "Retos del estudio sociohistórico del contacto de lenguas a través de un corpus documental. El caso del castellano en contacto con el catalán en Mallorca". *Revista de Investigación Lingüística*, 15, pp. 23–46.

EPIDOC: *Epigraphic Documents in TEI XML,* <http://epidoc.sourceforge.net>.

European Dialect Syntax. <http://www.dialectsyntax.org/wiki/Projects_on_dialect_syntax>.

Gervers, Michael, ed. 2000. *Dating Undated Medieval Charters*. Suffolk: Boydell & Brewer.

GITHE: *Grupo de Investigación de Textos para la Historia del Español*. <http://www.textoshispanicos.es>..

Isasi, Carmen. 2010. "Edición digital: retos nuevos en los nuevos recursos." In Mariña Arbor Aldea and Antonio F. Guiadanes (eds.), *Estudos de edición crítica e lírica galego-portuguesa.* Special issue of *Verba: Annuario Galego de Filoloxía*, 67, pp. 353–38.

Kawasaki, Yoshifumi. 2014. "Datación crono-geográfica de documentos medievales españoles." *Scriptum digital*, 3, pp. 29–63.

Lüdeling, Anke, Anne Kytö and Merja Kytö, eds. 2008. *Corpus Linguistics: An International Handbook*. Amsterdam: Walter de Gruyter.

Nouveau Corpus d'Amsterdam, <http://www.uni-stuttgart.de/lingrom/stein/corpus/#nca>.

PARES: *Portal de Archivos Españoles*. <http://pares.mcu.es>.

Pascual, José Antonio and Carlos Domínguez. 2009. "Un corpus para un nuevo diccionario histórico del español." In Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid and Frankfurt: Iberoamericana Verbuert, pp. 79–93.

Pato, Enrique and Elena Felíu Arquiola. 2005. "Alternancia de formas, nivelación e inferencia semántica: El caso de los participios en *-udo* del español medieval." *Revue de Linguistique Romane*, 69, pp. 437–63.

PRESEEA: *Proyecto para el Estudio Sociolingüístico del Español de España y América*, <http://PRESEEA.linguas.net>.

Renouf, Antoinete and Andrew Kehoe, eds. 2009. C*orpus Linguistics: Refinements and Reassessments.* Amsterdam: Rodopi.

Sánchez-Prieto Borja, Pedro, ed. 1991–2014. *Textos para la historia del español I.* 9 vols. Alcalá de Henares: Universidad de Alcalá.

——. 1998. *Cómo editar los textos medievales. Criterios para su presentación gráfica.* Madrid: Arco/Libros.

——. 2011a**.** *La edición de textos españoles medievales y clásicos*. San Millán de La Cogolla: Cilengua.

——. 2011b. "Ensayo de geografía lingüística histórica: términos para 'parcela de terreno agrícola' en las fuentes documentales de la Edad Media." In Sara Gómez Seibane and José L. Ramírez Luengo (eds.), *Maestra en mucho. Estudios filológicos en Homenaje a Carmen Isasi Martínez.* Buenos Aires: Ediciones Voces del Sur, pp. 271–302.

Spence, Paul et al. 2012. "Cruzando la brecha: la marcación digital con criterios filológicos." In Mª Jesús Torrens and Pedro Sánchez-Prieto Borja (eds.), *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Bern: Peter Lang, pp. 465–84.

The Anglo-Normand On-line Hub, <http://www.anglo-norman.net>.

TMILG: *Tesouro Medieval Informatizado da Lingua Galega,* <http://ilg.usc.es/tmilg>.

Ueda, Hiroto. In press. "La apócope extrema medieval en la fonética castellana y en la escritura a la francesa: Observaciones en el Corpus de Documentos Españoles Anteriores a 1700 (CODEA)." In Juan Sánchez Méndez and Mariela de la Torre (eds.), *Problemas y métodos en la edición y el estudio de documentos hispánicos antiguos*. Valencia: Tirant Lo Blanch.

Ueda, Hiroto and Antonio Ruiz Tinoco. 2007–2008. "The Varilex Project: Spanish lexical variation." *Linguistica Atlantica. Journal of the Atlantic Provinces Linguistic Association (Canada)*, 27–28, pp. 117–21.

Ueda, Hiroto and María Pilar Perea. 2010. "Método general de lematización con una gramática mínima y un diccionario óptimo. Aplicación a un corpus dialectal escrito." In Isabel Moskowich-Spiegel Fandiño et al. (eds.), *Visualización del lenguaje a través de corpus*. A Coruña: Universidade da Coruña, pp. 919–32.

——. 2011. "Applying quantitative analysis techniques to *La flexió verbal en els dialectes catalans*". *Dialectologia et Geolinguistica: Journal of the International Society for Dialectology and Geolinguistics*, 18, pp. 99–114.

BOOK REVIEWS