Inmaculada González Sopeña

Language Corpora and Lexical Arabisms in the Digital Age

1 Introduction. The Era of Digital Humanities

The so-called Digital Era has meant a paradigm shift in scientific research. The industrial era has undergone a transition into a digital era since the advent of computers and similar devices, as well as since access to Internet became widespread for most of the world population since the late 20th c. and early 21st c. (Jódar Marín 2010: 3). The Internet has become a new social forum where a wealth of data of all kinds are released. A range of technological tools has thus become quickly available for research in all fields to further human knowledge and has made it accessible online. In a word, the digital revolution means a paradigm shift in research practice.[1]

Within the Humanities, the use of such resources, methods and tools has sparked an intense debate, to the extent that its role and the very view of the discipline have been revised regarding how to use and transfer this progress to society (Romero Frías 2014: 19). The concept of *Digital Humanities*[2] and its very notion as a specific field arise at this point. Roberto Busa's project on concordance design for specific search of Thomas Aquinas's works, assisted by IBM on the computational side of the project,[3] has been cited as the starting point of this process (Allés Torrent 2019, Romero Frías 2014, Spence 2014b). The project brought to light the need for analog-to-digital text conversion to enable computer-assisted data search for concepts, language structures, etc. Corpus linguistics and Philology thus became pioneers in the use of these new methods. At present, the Digital Humanities aim at "the design and use of applications and models for new teaching

1 Cf., thus, the use of the term *cyberscience* or *e-Science* (Romero Frías 2014:22).

2 This label dates back to the label *Humanities Computing* of the 1950s (Spence 2014b). The publication of *A Companion to Digital Humanities* in 2004 renamed the discipline for a wider scope and full coverage of all the changes brought by the digital revolution to such a multifaceted field as the Humanities (Romero Frías 2014, Allés Torrent 2019).

3 Namely, *Index Thomisticum*.

Note: This contribution has been supported by Grant PID2022-136256NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by the European Union's ERDF/EU. Also, it has been carried out within the framework of Grant C-HUM-038-UGR23 funded by *Consejería de Universidad, Investigación e Innovación* and by ERDF Andalusia Program 2021–2027.

https://doi.org/10.1515/9783110781465-003

and research modes" (Allés Torrent 2019: 6). To this end, they include various processes regarding data processing (creation, management, storage, reuse). In the digital era, data, which may be of various kinds, are viewed as "an items', attribute's or empirical value's smallest digital expression and symbol representation that is formalized digitally as a binary notation of 0 or 1" (Allés Torrent 2019: 12). Five major data types can be identified, according to their digital encoding: Text, numbers, images, video, and audio.

Accordingly, data encoding has given rise to "virtually unlimited research avenues, ranging from *Big Data* to *Small Data*" (Romero Frías 2014: 20). This is thanks to the new technological resources and tools available, both in general and specifically for research in the Humanities too. The *big data* boom, reportedly starting in 2011, has been defined variously, even if "most experts and academics agree that it is some kind of combination of algorithms, technologies and strategies capable of collecting and analysing large amounts of data" as fast as possible (Sánchez González 2014: 136). The aim is "to bring together, store and process as much information as possible" (Allés Torrent 2019: 19), with a view to subsequent quantitative analysis for patterns and dynamics, and for the construction of predictive models. In the Humanities, this amounts to the information that a researcher cannot read in a given period of time and whose interpretation requires computational tools (Graham, Milligan & Weingart 2015). Schöch (2013) put forward a closely related term, *smart data*: Unlike *big data*, smart data are semi-structured, explicit, enriched, clean, interconnected, and need human processing, e.g. XML-marked-up texts for digital editions.

As for Philology, the Digital Humanities have contributed a large number of new tools and resources for language research. Corpus linguistics and computational linguistics were the first disciplines to develop specific tools for retrieval, analysis, and interpretation of text data (Martínez-Gamboa 2016). The Digital Humanities therefore stand out for the design and compilation of language corpora as experimental bases for research on relevant questions in Spanish and in other languages.

This chapter is intended to briefly review the contribution of corpus linguistics to research on Spanish up to present-day. This includes a short account of corpus linguistics and its objectives (with a focus on text encoding), and a discussion of its use for research on Spanish. The chapter therefore reviews the properties of the main types of Spanish language corpora: Reference corpora (CORDE, CREA, CORPES, etc.) and specialized corpora (*CorLexIn*, ODE, CORDIAM), among others. It also reviews research on Arabisms in Spanish, and assesses how this specific field has been influenced by the digital revolution. Therefore, a bibliographical review is also included of the main research avenues on Arabisms and of the difficulties for their identification and retrieval in corpora.

In this regard, the chapter spans from the classics and from most traditional studies to the most recent corpus-based references. The resulting review shows the advantage in researching Arabisms, their variants, their text sources, and their chronological attestation based on the new methods and the new technological tools of corpora. Indeed, the new technologies have helped improve the lexico-semantic classification of Arabisms, and also revise specific statements about the periods when Arabisms were added most productively and when they were more frequently replaced by other words. These and other improvements are some of the contributions to the study of Arabisms in Spanish by TEI-compliant, XML-marked-up transcripts stored within the TEITOK platform. TEITOK (Janseen 2016) was originally designed as a platform for corpus storage and exploitation, as is the case of the most relevant corpus for this chapter, namely *Oralia diacrónica del español* (ODE). The chapter closes with short conclusions drawn from the preceding sections.

2 Corpus Linguistics and Spanish Language Corpora

Corpus linguistics, whether a discipline or a method,[4] is rather new, as it can hardly be conceived without computer technology. The first language corpus designed for computerized use, the *Brown University Standard Corpus of Present-Day American English*, dates back to 1964 (Rojo 2016: 286). A *language corpus* is:

> [. . .] un conjunto de (fragmentos de) textos, orales o escritos, producidos en condiciones naturales, conjuntamente representativos de una lengua o variedad de lengua, que se almacenan en formato electrónico y se codifican con la intención de que puedan ser analizados científicamente.[5] (Rojo 2021:1)

The need for corpora to be built and stored "in electronic format" makes computer technology a turning point for the capacity to manage text in various formats, to store data by the thousand million, and to efficiently retrieve language

4 The debate on whether corpus linguistics is a new discipline, a new approach or a new method (Leech 1992, Gries 2006, Parodi 2010) has reached the generally acknowledged conclusion that it is an approach, such that the language data are researched empirically using new tools for quantitative and qualitative analysis (Rojo 2021: 49–50).

5 "[. . .] a collection of (passages of) texts, spoken or written, produced in natural conditions, representative of a language or language variety, stored in electronic format and encoded for scientific research". (Rojo 2021:1).

data with new applications. Corpora are classified according to their text samples (originally, from communication not intended for research), representativeness, size, and also to the balance between chronological periods and text types[6] (Rojo 2021: 63–69). Regardless, electronic format remains a basic, essential requirement for digital encoding, i.e. the language data of a text require conversion into machine-readable language (Rojo 2020: 94).

Corpus encoding relies on corpus design, which, in turn, is according to the corpus purpose. Encoding has developed rapidly as a result of computational progress.[7] *Encoding* is a broad concept that encompasses information of various levels. Text transcription protocols, including decisions like what textual and paratextual information is annotated (e.g. paragraphs, side notes, comments), are examples of first level encoding. Addition of metadata of each corpus sample text (e.g. date, country, author) make a different level. Linguistic annotation, e.g. morphosyntactic tagging or lemmatization, are an additional level too.

In general, encoding starts with conversion of text samples into machine-readable data, after due arrangement and classification (Allés Torrent 2019: 10). Document conversion into machine-readable mark-up languages like XML, HTML, XHTML, or LaTex, is more and more relevant (Rojo 2021: 74). This process is in accordance with international annotation standards specifically designed for language research, e.g. the *TEI* initiative (*Text Encoding Initiative*) (2016).

Another major point in corpus design is which metadata are to be added to each sample document, i.e. the information on the author, the text, the year of production, the country of origin, the text type, and the register, among others.[8] The functionalities of a corpus become wider, if morphosyntactic tagging and lemmatization are added (Rojo 2016: 286, 2021: 2–3), as this makes available all kinds of grammatical and lexical information of each word and, thus, any grammatical question can be researched in depth. Computational linguistics has sup-

6 The use of the *Web* as a corpus has become a relevant issue over the past years. Such scientific use of the Web still suffers from shortcomings, like the dependence on appropriate search engines, difficulties for the use of regular expressions for retrieval of specific data, or the very fact that the Internet is, by nature, an ever-changing body of data, so any results may change virtually by the day (Rojo 2021: 71).

7 Note that the earlier techniques for analog-to-digital text conversion were based on OCR scanning and the subsequent creation of a .pdf file capable of sustaining unlimited searches (Rojo 2021: 88). Use of plain text for specific queries or frequency data followed afterwards. At present, mark-up language capable of telling documents from their metadata, like SGML, TEI, or CES, is used (Rojo 2021: 92).

8 In addition to data insofar as digital representations, metadata are essential for data classification (Allés Torrent 2019: 13). The possibilities are immense here, and may vary sharply according to text type.

plied a number of taggers and lemmatizers, like FreeLing, Peen Treebank, or the general guidelines of the EAGLES Consortium. Even so, all have limitations and rely on strictly linguistic decisions for the description as a tag of a given word in a given language: Which morphological elements are to be tagged in nouns or verbs? How can they be converted into machine-readable codes? How should Spanish agglutinative forms like *haberlas* be processed? Other issues must be considered too, like syntactically and semantically different homographs, e.g. *vino* ('come.PAST' vs. 'wine') or *la* ('the' vs. 'her.ACC'). Similar questions arise during lemmatization of each orthographic form, especially when it comes to subsuming a whole pronominal paradigm under a lemma, or regarding the lemmatization of locutions in Spanish. The answers to these questions are according to the purpose of each corpus.

2.1 Corpus Types in Spanish

The first Spanish corpora were built in the 1990s, and several remarks are in order prior to their description. First, corpora can be general (reference), specialized, learner corpora, and technical language corpora.[9] These may have been designed as data sources of the evolution of a language over time (diachronic), of the most recent state of language (synchronic), or of diatopic, diastratic, or diaphasic variation.

Back to the abovementioned question of size, there are Spanish diachronic and synchronic corpora of millions of words. The main difference between these and smaller corpora is that the former contain thousands of text samples, unprovided with an editing policy because they consist in previously edited text, and unprovided with a computational structure (in many cases, without encoding and lemmatization). By contrast, small, specialized corpora have been edited much more exhaustively and rely on a text conversion and language annotation policy. Smaller corpora of the diachronic kind also allow several text presentation modes, e.g. palaeographic, critical, and even facsimile editions (Rojo 2021: 181).

The Spanish Royal Academy (hereafter, RAE) has fostered two major large corpus projects of diachronic and contemporary Spanish since the 1990s: CREA (*Corpus de Referencia del Español Actual*) and CORDE (*Corpus Diacrónico del Español*). CORDE is a 250-million word corpus encompassing from the origins of Spanish up to 1974. CREA is a 160-million word corpus produced between 1975 and 2004. At the turn of the century, these corpora were supplemented with addi-

9 For a review, cf. Rojo (2021: 72–75).

tional projects: The multi-layered, 300-million-word CDH (*Corpus del Nuevo Diccionario Histórico de la Lengua Española*), and CORPES XXI (*Corpus del Español del Siglo XXI*), whose latest version amounts to over 381 million words of spoken and written sources produced in the Spanish-speaking world.

By contrast, Mark Davies' *Corpus del español* hosts several subcorpora that are not entirely well-balanced: The historical subcorpus is considerably smaller (ca. 100 million words) than the dialects subcorpus (2,000 million words collected from websites[10]). While, in general, all are large corpora and offer overviews of specific topics, they rely on technological resources that do not allow accurate data retrieval.

The diachronic corpora cited above rely mainly on literary and formal register texts. A need for other text types has been arguably considered necessary in the past few years, and a wealth of corpora of specific periods, geographical areas, and text types have mushroomed as a result. Thus, both large and specialized medium-sized and small diachronic corpora of various Spanish text types are available at present. Following the typology by Torruella & Kabatek (2018), there are corpora of European Spanish (CODEA, *Corpus Mallorca*, CODEMA, *CorLexIn*, ODE), of American Spanish (CORDIAM, COREECOM), and of both varieties, like CHARTA's network's.[11]

All in all, the above offer a range of language samples of various periods, Spanish-speaking geographical areas, and texts from outside the literary realm (correspondence, last wills, goods inventories, ordinances, diaries, legal agreements, chronicles, newspapers). All these corpora contain significant information for the accurate revision and for an improved account of the history of Spanish at all levels.

2.2 Technological Issues in the Spanish Diachronic Corpora, and New Methodological Proposals

Following the above review of Spanish language corpora, this chapter focuses on diachronic corpora, where digital progress has been slower and, often, unsatisfactory. Despite the large number of language corpus projects, many initiatives do not apply the potential of the new digital tools and applications to the design and

10 Since 2018, this corpus has included the subcorpus *NOW*, with over 7,000 million words produced between 2012 and 2018 (www.corpuesdelespañol.org).

11 Besides these references, it is worth mentioning all the language corpora built in the 1990s cited in specific volumes like Sánchez Prieto (1995), Fontanella (1993), or Rojas (2008). A comprehensive review of Spanish corpora is available in Calderón Campos (2015), and in Rojo (2016).

exploitation of language resources (Díaz Bravo 2018: 565). This is as a result of the fairly frequent design flaws and deficient selective criteria used for diachronic corpora of old documents of all stages of the history of Spanish.

CORDE, CDH, and the historical subcorpus of *Corpus del Español* (CdE) have similar limitations as regards exploitation. Thus, CORDE is not lemmatized, is not tagged, and does not allow graphical representation of queries.[12] The CDH corpus, based on newer technology, is partially lemmatized and tagged, and sustains queries by genre and subject topic. Still the classification used is so detailed that accurate data for research on diatopic and diaphasic variation are rather difficult to retrieve. Queries by absolute and relative frequency also have limitations as regards graphical representation. Additionally, these corpora contain many documents which are according to their own edition guidelines, so a range of editorial policies coexist. The historical subcorpus of CdE uses an extremely deficient classification and description, and is also flawed by frequent errors in the links supplied for retrieval of the source samples.[13]

Some shortcomings of Spanish diachronic corpora are summarized below (Díaz Bravo 2018: 580):

- A lack of balance across centuries and genres, such that medieval texts are comparatively underrepresented in the CORDE corpus and in the CDH corpus.
- A lack of unified editorial criteria across corpus projects.
- Frequently inaccurate document metadata, often lacking in uniformity across corpus projects.
- Despite the many lemmatizers and taggers available, a bias towards contemporary Spanish that demands much-needed improvement in pre-20th c. data.

Specific methods have been made available to overcome some of the abovementioned difficulties based on digital resources and tools, and which allow consistent and accurate analysis of a number of topics by use of diachronic corpora. The historical corpora *Post Scriptum* and *Oralia Diacrónica del Español* (hereafter, ODE) set several prime examples, as their research procedures and technology rely on

12 Albeit rather poorly, regular expressions are possible, with items like * or ?, as well as with logical operators like AND, OR, or NOT.

13 *Sketch Engine* is a completely different case, in that it is a digital tool for language research comprehensive of a range of subcorpora of various languages amounting to millions of words and relying on newer technology capable of sustaining multiple research options (concordances, collocations, frequency lists, etc.). Based on its more complex tagging and lemmatization, it allows queries by morphological tag, lemma, and text type, among others, as well as by other encoding procedures (https://www.sketchengine.eu/#blue).

the widely acknowledged TEI[14] standard for XML (*eXtensible Markup Language*) mark-up text encoding and organization. Additionally, TEITOK, specifically designed by Janssen (2016) as a corpus storage platform and based on TEI-compliant, XML text conversion, is gaining ground as a widespread reference of use.[15]

XML exceeds .doc and other files in many respects, e.g. by not needing specific hardware or software and thus being interoperational, by being reusable as various formats, by separating form from contents, and by being extensible as a result of not using a closed tagset. It also allows text data organization and modelling by use of accurately defined, machine-readable marks, whether the data is of the structure of the document, editorial, of a semantic nature, of abbreviations, of images, etc. The marks are formalized as tags to describe a text segment. In XML, marks are arranged as angles in customizable modules according to various purposes.

XML text formatting and the TEI *Guidelines* allow text encoding comprehensive of all the textual and metatextual properties desired in a given corpus. The basic structure in TEI-compliant XML text conversion consists of: i) a header (<teiHeader>); and ii) a body (<body>). The header includes any metadata considered relevant. In old texts, they cite the title, and data on its creation, like the source archive identification and the reference signature of the bundle and page, the text type, and, if available, information of where, when, and by whom it was produced. The following example (Figure 1) is for a last will signed in Badajoz in the 17th c., of which some metadata have been recorded in the ODE corpus:

The above data selection must have been made early in the stage of corpus design, also because it is according to the type of corpus and to the samples used.[16] Metadata encoding allows efficient data retrieval for various research projects.

Otherwise, the body of an XML, TEI-compliant document is the text transcript. Text transcription entails the use of a range of tags according to the TEI *Guidelines*, for identification of textual and paratextual information: Beginning of new page, beginning of new line, font change, side notes, crossed out material, overwriting, signatures, as well as strictly linguistic data (e.g. instances of /θ/ spelt as *s*, unstable use of liquid consonants, borrowings). Figure 2 shows the most basic structure of a page taken from a late 17th c. dowry letter:

14 The *Text Encoding Initiative* is an international standard available since 1987 specifically designed for the Humanities as regards digital editing of documents and electronic encoding (Díaz Bravo 2018). Still, its use for old documents is rather limited (Calderón Campos 2019).

15 A list of corpus projects thus built is available at <http://www.teitok.org/index.php?action=projects>.

16 These parameters are substantially different in spoken and written samples, and also if other parameters are added too, e.g. 'informal register', the speaker, their age, their sex, etc.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlnsoff="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Testamento de Franciso Hernández Sevillano</title>
                <editor id="CORTENEX">CORTENEX: Corpus de textos notariales extremeños (siglos XVI y XVII)</editor>
                <funder/>
                <respStmt>
                    <resp id="transcription">Inmaculada González Sopeña</resp>
                </respStmt>
                <respStmt>
                    <resp id="standardization">Inmaculada González Sopeña</resp>
                </respStmt>
                <respStmt>
                    <resp id="annotation">Inmaculada González Sopeña</resp>
                </respStmt>
                <respStmt>
                    <resp id="revision"/>
                </respStmt>
            </titleStmt>
            <publicationStmt>
                <publisher>UGR, Universidad de Granada</publisher>
                <pubPlace>Granada</pubPlace>
                <distributor>HUM-278. Grupo de Investigaciones Histórico-Lingüísticas y Dialectales. UGR-Junta de Andalucía</distributor>
            </publicationStmt>
            <sourceDesc>
                <msDesc>
                    <msIdentifier>
                        <country>España</country>
                        <settlement>Badajoz</settlement>
                        <institution>Archivo Histórico Provincial de Badajoz</institution>
                        <repository>Protocolos Notariales</repository>
                        <idno>AHPB PN/289</idno>
                    </msIdentifier>
                    <msContents>
                        <summary>NA</summary>
                        <msItem class="original">
                            <p><locus>59r-60v</locus></p>
                        </msItem>
```

Figure 1: A cropped screenshot of the XML header of a last will (ODE).

```
<pb n="402v" facs="BA_IMG_2791.jpg"/>
<p>
    <lb/> <gap reason="illegible"/> sarçillos con çinco pendientes de perlas
    <lb/> y dos surtijas, la una grande de piedras verdes,
    <lb/> y la otra con diez con esmaltes vlancos y verdes
    <lb/> y una cruz pequeña con siete piedras verdes,
    <lb/> todo ello de oro que peso treçe rs||reales de a ocho y m<add place="above">
    <lb/> y dos rs||reales de plata, tasado en duzientos y qua<lb/>renta y siete rs||
    <lb/> Una artesa con su banca de madera, tasada
    <lb/> en quatro ducados.
    <lb/> Un medio arcaz de madera tasado en q<add place="above">tro</add>||quatro rs
    <lb/> Un escritorio llano de madera, tasado en
    <lb/> dozienttos rreales.
    <lb/> Quatro silla de vaqueta negras, tasadas
    <lb/> a quatro ducados cada una.
    <lb/> Mas un baul de baqueta negro usado tasado
    <lb/> en sesonta rs||reales.
    <lb/> Vn arca pequeña tasada en veynte rs||reales.
    <lb/> Un catre de madera forrado en treinta y tres r<add place="above">s</add>||r
    <lb/> Vna mesa de madera tasada en treinta y
    <lb/> tres rs||reales.
    <lb/> Un bufete pequeño con su cajon tasado en onçe rs||reales.
```

Figure 2: A cropped screenshot of the XML body of a document (ODE).

Thus, specific tags are used for beginning of a new page (<pb/>), for beginning of new paragraphs (<p>), or for beginning of new lines (<lb/>). Figure 2 also shows a mark for the position of a letter above or below a line (<add>). Many tags are also enlarged with attributes and values to enable highly specific queries and meticulous detail about the information contained in a sample, e.g. <gap/> signals omission, which can be enlarged in turn with the attribute 'reason' for greater detail. In Figure 2, the latter is "illegible" (the original may have been torn or damaged). By contrast, other tags, e.g. "editorial", are used for deliberate deletion of an irrelevant excerpt.

Further, various editions (palaeographic, critical) of one and the same sample may be available, and a fully tagged and lemmatized corpus may be stored within the platform TEITOK. Thus, each text is tokenized, i.e. each word is marked up with a token or numbered angled mark detailing the data of its normalized, expanded form, its morphosyntactic tag and its lemma. In old texts, each tokenized sample is processed for orthographic normalization prior to further editing. TEITOK uses a semiautomatic tool to supply a normalized form for each word[17]:

(1) <tok id="w-76" nform="había">auia</tok>

Example (1) illustrates the tokenization and normalization of *auia*: Tokenization first converts every form into a token, and orthographic normalization then adds a new attribute (*nform*) to the XML file with a specific value, in this case *había.* The application *Neotag* (Janssen 2012), a tagger that follows the model proposed by EAGLES (*Expert Advisory Group on Language Engineering Standards*), finally tags and lemmatizes each of the resulting forms within the platform.[18] Morphosyntactic tagging translates each morpheme of the grammatical categories of Spanish into an alphanumerical code:

(2) <tok id="w-76" nform="había" lemma="haber" pos="VAII3S0">auia</tok>

17 Manual post-editing follows to revise any forms that the tool may not have normalized correctly. This is because the process starts from a training corpus large enough as to allocate orthographically standard forms during the first stages of use within the TEITOK platform. The most frequent wrong normalization mistakes in Spanish occur for homographs of the type *se* vs. *sé*, or *mi* vs. *mí.*

18 Manual post-editing of *Neotag*'s output follows for revision according to what may and may not occur in each word-class. *Neotag* relies on a probabilistic algorithm that often tags wrongly orthographically identical forms and multifunctional words, e.g. *que* (as a conjunction vs. as a relative pronoun).

Example (2) shows *Neotag*'s output of (1), whereby each token is enlarged with two new attributes for specific values: i) the attribute *lemma* (in the example, the infinitive form, because it is a verb); and ii) the attribute *pos*, where the morpho-syntactic tag VAII3S0 stands for "verb", "auxiliary", "indicative", "imperfect", "3rd person", "singular". The segment 0, which stands for the lack of any other morphological contents regardless of the lemma's word-class, must be included on account of the inflection for "gender" in other verbal forms, e.g. in participles.[19]

In the end, queries for any linguistic issue can be designed for accurate data retrieval: By transcribed forms, by standardized forms, by lemma, by tag, etc. TEI-TOK also supports CQL (*Corpus Query Language*) for queries combining many of the parameters recorded. The last part of this chapter reviews these technological possibilities in the specific field of research on Arabisms in Spanish.

3 The Research on Arabisms in Spanish

The research on Arabisms in Spanish has attracted the interest of a large number of philologists, partly for the nearly eight centuries of close contact between the varieties of Arabic and of Romance languages spoken in the Iberian peninsula since the Muslim invasion in 711 (Corriente 1977). The number and the range of studies on Arabisms is thus extremely wide and, as a result, difficult to summarize succinctly.

The lexical component of a language is a particularly challenging research field, among other reasons, because it is an ever-changing level that readily echoes diachronically external changes, whether social, cultural, political, etc. Aside from a language's morphological processes for word-formation (derivation, composition), the lexical component relies on borrowings, i.e. on elements that a language takes from another language (Gómez Capuz 2004).[20]

Borrowings presume contact between languages or language varieties, a key component in the history of Spanish. Language contact is measured in terms of *language influence* and the *intensity of the contact*, and both are evident in the concurrence of linguistic and extralinguistic factors of Arabisms: On the one

19 The sequence of letters and numbers may vary according to the word-class, e.g. the arrangement of grammatical inflection is not the same in nouns as in pronouns.

20 Borrowings have been approached from a range of theoretical positions (structuralist, formalist, funcionalist), each contributing specialized knowledge for their research, e.g. *cultural borrowings*, *integral borrowing*, *lexical* and *semantic calques*, *foreign words*, or *adapted borrowings*, among others.

hand, a large number of Arabisms were borrowed into Spanish, even if there was hardly any morphosyntactic influence for the structural difference between the languages involved; on the other, the contact went through various stages at which a range of differences and sociocultural roles between the Christian and the Muslim populations can be noticed. The latter brings to the fore the relevance of *prestige* for borrowing as a process (Giménez-Eguíbar 2016). Based on the above, the following types of studies on Arabisms can be considered:

– Studies on the periods during which most Arabisms are borrowed, and into which lexico-semantic fields they were borrowed;
– Studies on the decreased borrowing of Arabisms and on their competition with other words, or their replacement by other words (lexical replacement); and
– Studies on the phonetic processes involved in the borrowing of Arabisms, and on the great spelling diversity of Arabisms since the Middle Ages.

The three types cover tens of papers, monographs and handbooks on the periods of the history of Spanish. A good part of the research on Arabisms focuses on the Middle Ages, as this is the period of closest contact and, therefore, when the largest number of Arabisms were borrowed (Oliver Pérez 2004), considering the social constraints imposed by the Christian reconquest. The subject topics thus range from Arabisms in King Alfonso X's works and in mediaeval treaties and chronicles (García González 1998, Maíllo Salgado 1998, Neuvonen 1941, Pocklington 1984) to the role of the Mozarabic community in the integration of Arabisms into Spanish (García González 2007). The research on Arabisms reaches as late as the Modern era, mainly with a focus on loss or obsolescence, for the cultural disregard of the Muslim world that was concomitant with Humanism and the Greek and Latin revival (Walsh 1967), and for the competition or replacement by words of non-Arabic stock (Giménez-Eguíbar 2015, 2016). The research on Arabisms during the Spanish Golden Age examines words with low attestation records, or words that occur or remain in highly specific lexico-semantic fields (Calderón Campos 2010, González Sopeña 2017, Morala Rodríguez 2012a). A large amount of research on Arabisms is also available in specific Spanish-speaking geographical areas or on dialects, both synchronic (Garulo Muñoz 1983) and diachronic (Torres Montes 1996).

The 18th c. bears witness to a decrease in research on Arabisms, for the rise of French borrowings during the Age of Enlightment. Still, lexicography has produced dictionaries and glossaries of Arabisms and word stock of Eastern origin since the 19th c., e.g. by Dozy & Engelmann (1869), or by Eguílaz & Yanguas (1886). The knowledge on Arabisms is finally broadened by the lexicographic information available in general and in etymological dictionaries of Spanish (e.g. the RAE

dictionaries, or DCECH, respectively), and in dictionaries of Arabisms (Corriente 1999).[21] General descriptions of Arabisms available in handbooks of the history of Spanish can also be found, both classic (Lapesa 1981 [1942]) and newer publications (Giménez-Eguíbar 2023).[22]

Indeed, the diachronic relevance of Arabisms lies behind the substantial number of titles on the topic. Even so, a significant loss of specific information on Arabisms is lost in previous databases. This is either for the technical limitations of the corpora reviewed above which supplied the experimental evidence for many of the references cited, or for being based on non-digital sources. The paradigm shift described earlier in this chapter no longer sustains the latter approach, as the scientific community demands research data to be available online and as open access.

4 Research on Arabisms and Language Corpora: The Case of *Oralia diacrónica del español* (ODE)

In the above, this chapter has described the complexities of lexical research in general. Many of them arise because the lexical level is the most superficial, the most variable, and the most difficult to systematize in a language. Lexical research must rely on sources where all the linguistic properties of a term can be attested so they can be added to exhaustive historical dictionaries[23]: Lexical senses, semantic widening and reduction, lexical replacement, earliest attestations, etymological information, actual examples of a term, etc. Therefore, corpus-based diachronic lexical research is one of the most successful methods available at present. The abovementioned sources enlarge and improve the word stock, and record forms that are beyond what can be considered the usually acknowledged most standard, academic vocabulary (Morala 2012b: 200).

21 Federico Corriente stands out in the research on Arabisms for his dictionaries, but also for the vast number of papers and monographs he produced on the subject. Arnold Steiger (1932) is another major name, in this case for his research on the phonetic adaptation of Arabisms.

22 Some handbooks on the history of Spanish lexis also describe the words of Arabic origin in Spanish (Dworkin 2012, Colón Doménech 2002).

23 In the case of Spanish, this task has been frequently interrupted for various reasons. At present, two historical dictionaries are pending completion. CDH, which lays the foundations for the NDHDLE / *Nuevo diccionario histórico de la lengua española*, currently underway, has resumed the task of a historical dictionary of Spanish.

Specific issues arise in the case of Arabic lexis, like the great variation in spelling according to their written records, and the subsequent difficult systematization of all the possible variants of one Arabism. The description of such words in other documents was not even possible until recently, as many of the existing corpora rely on highly formal sources.

The following presents some of the advantages of new methods in the Digital Humanities. They have improved research on Arabisms in Spanish by use of the technology that is behind the corpus *Oralia diacrónica del español* (hereafter, ODE).

The methodological principles underlying the ODE corpus are part and parcel of technical resources designed specifically for corpora. The ODE corpus, a medium-sized specialized corpus, consists of three text types that go beyond the typical formal register sources used in corpora: i) goods inventories; ii) witness testimonies; and iii) incident reports (Calderón Campos 2019). Geographically, most of the text samples are from Andalusia, even if control subcorpora of other provinces in the country are also available for the period between 1492 and 1833. It is worth remarking at this stage that the ODE corpus builds on the CORDEREGRA (*Corpus diacrónico del español del reino de Granada 1492– 1833*), even if marked differences arise from the two, e.g. the geographical area covered, the use of XML-marked-up, TEI-compliant samples, or access within the TEITOK platform (Janssen 2016).

The latter two features help overcome obstacles such as the ones reviewed in Section 2 above, e.g. use of queries for transcribed text alone (i.e. excluding metadata), visualization as various types of edition (semipalaeographic, standardized), retrieval of every possible spelling variant, and availability of a corpus tagged and lemmatized according to consistent philological criteria, and of the linguistic information contained therein by use of accurate queries (Calderón Campos & Vaamonde 2020: 177).

As a result of the implementation of all levels of textual and metatextual enconding reviewed above, each XML-marked-up, TEI-compliant sample of the ODE corpus may display a header with information such as the year of production, the author, the text type, the archive, the title, the geographical area it comes from, etc. Each transcript contains tags with paratextual and with linguistic information. As all samples are morphosyntactically tagged and lemmatized, CQL queries allow exhaustive, accurate data retrieval. Figure 3 shows various query options based on the encoding described.

Data retrieval may be as transcribed, expanded, or normalized text, and by lemma or by tag. A range of parameter combinations can be used too as CQL queries (e.g. determiner + noun). Lemmatization allows retrieval of all possible variants of a term, even the least likely ones. The text type is a key point here, as the least formal texts, like those included in the ODE corpus, often contain terms

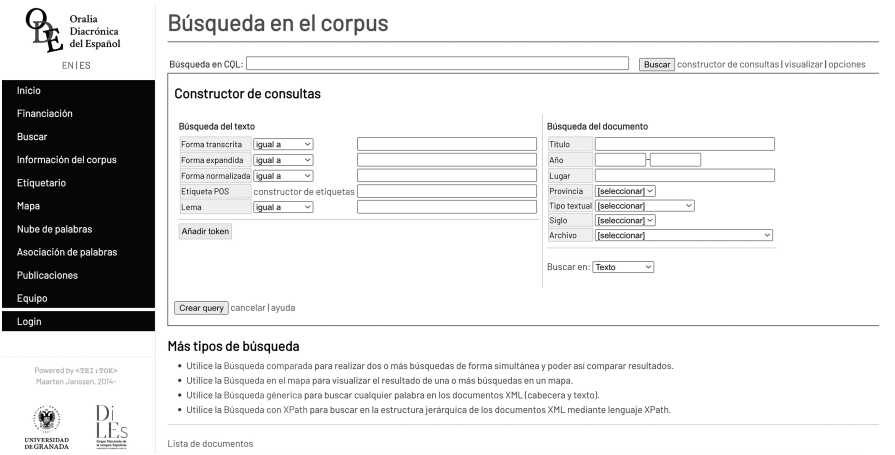

Figure 3: The ODE corpus query interface.

written as the scrivener would best understand in the communicative situation in question. Thus, a query for the Arabism *tahalí* 'sheath' (Figure 4) retrieves unexpected forms like *taali* or *taxali*:

context	de lana blanca bordados y	**taali**	. | Vna vasquiña y vn	1661	España, Badajoz, Badajoz
context	larga y vn frasco y	**taali**	en ochenta y ocho rs	1701	España, Jaén, Villacarrillo
context	olan de cristal y un	**tahali**	de vezero vordado de plata	1666	España, Badajoz, Badajoz
context	en quatroçientos rs. | Un	**tahali**	de cordouan con fluecos negros	1666	España, Badajoz, Badajoz
context	calzettas nueuas, mas dos	**tahalies**	, vno | de vaca y	1677	España, Cáceres, Cáceres
context	de a bara. | Vn	**talai**	| Vna bandola de vaqueta.	1704	España, Murcia, Lorca
context	con su mangas; un	**taxali**	de tela | parada plateada;	1661	España, Badajoz, Badajoz

Figure 4: *Tahalí* 'sheath' in the ODE corpus. Query by lemma.

A similar case can be cited for *jáquima* 'cord lead' (Figure 5), and the Granadan variant *xaquyma*:

Retrieval of Arabisms allows, for words like *alhaja* 'jewel' (Figure 6), not just the concordances and variants, but also the frequency of each spelling variant recorded, as in Figure 4:

Figure 7 shows a different visualization option for Arabisms with multiple spelling variants, both in the singular and in the plural number:

In the example of *guadamecí* 'garnished leather', the information retrieved includes spelling variants, but also instances of /θ/ spelt as *s*, vowel alternation, or the addition of a liquid consonant at the end of the word. This information, com-

context	. Un | caveson y una |	**jaquima**	. Dos | pleytas grandes.	1752	España, Sevilla, Osuna
context	seis. | Un caveson y |	**jaquimas**	en | dos reales. | Dos	1752	España, Sevilla, Osuna
context	cuchillos de coçina. | Una	**xaquima**	nueba. | Abriosse otro	1663	España, Cáceres, Cáceres
context	rastrillo. | | Dos cadenas de	**xaquima**	de mula. | Vna guarniçion	1564	España, Cáceres, Cáceres
context	de la brida. | Vna	**xaquima**	vieja de cañamo.	1576	España, Cáceres, Cáceres
context	de cavallo viejo. | Tres	**xaquimas**	nuevas de cañamo. | Dos	1564	España, Cáceres, Cáceres
context	tres o [...]. | | Yten tres	**xaquimas**	de cañamo guarneçidas de	1580	España, Sevilla, Sevilla
context	muerto naturalmente, | desatada la	**xaquyma**	, que se abia | sacado	1578	España, Granada, Iznalloz

Figure 5: *Jáquima* 'cord lead' in the ODE corpus. Query by lemma.

Corpus Distribution

Search Query	Lemma = **alhaja**
Group query	Expanded form
Total	129
Reference size	1210347

Graph: Table | Count: Count | Download: [select]

Group	Count	WPM	Percent
alajas	46	38.01	35.66
alaxas	33	27.26	25.58
alhajas	23	19	17.83
Alajas	8	6.61	6.2
alhaxas	6	4.96	4.65
Alaxas	3	2.48	2.33
Alhajas	3	2.48	2.33
alagas	2	1.65	1.55
âlaxas	1	0.83	0.78
alaxittas	1	0.83	0.78
alajitas	1	0.83	0.78
Alhaxas	1	0.83	0.78
alfajas	1	0.83	0.78

Figure 6: The frequency of the variant forms of the lemma *alhaja* 'jewel' in the ODE corpus.

bined with metadata of geographical location, helps track down the evolution of such processes over time.

Yet, research on diachronic and diatopic lexical variation, especially on Arabisms, requires multilayer encoding: XML-mark-up, TEI-compliance, TEITOK access. Specifically, corpus lemmatization is a key requirement, even more in the

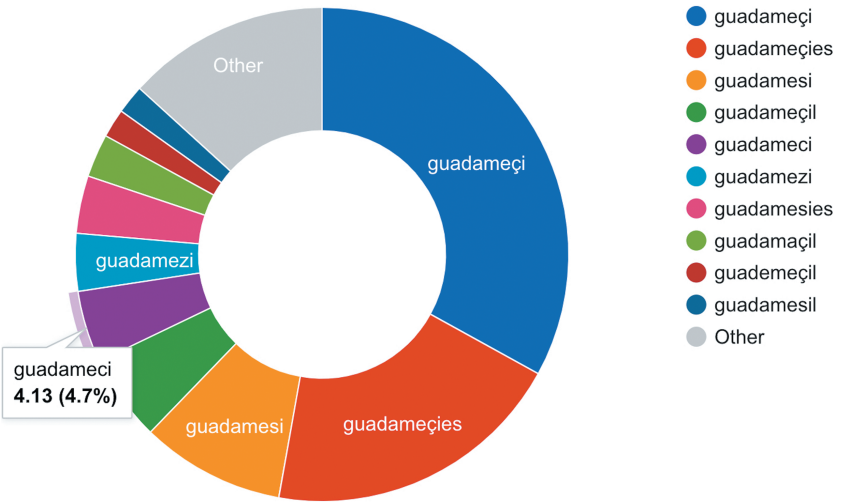

Figure 7: The frequency of the variants of *guadamecí* 'garnished leather' in the ODE corpus.

ODE corpus, because it contains samples that are particularly close to spoken language.

Despite the progress in the research of Arabisms based on a corpus built as described above, there is still room for improvement, e.g. by use of TEI tags (<etym>) for dictionary-making, even if this entails manual tagging of each Arabism during text conversion. Semantic annotation by lexical field[24] may mean a major step forward in the classification of such a versatile component as a language's lexis, but identification of Arabisms requires human intervention too. Finally, XML's versatility allows use of specific, customizable scripts in TEITOK corpora and, again, retrieval of Arabisms relies on the availability of a list of all the cases contained in the corpus and on their subsequent manual retrieval.

5 Conclusions

Diachronic Spanish corpora are still to face the challenge of the implementation of the technical and digital methods and resources available for language research today. This chapter reviews some of the advantages of textual and meta-

24 Consider proposals like USAS (UCREL Semantic Analysis System) <https://ucrel.lancs.ac.uk/usas/>.

textual digital encoding of a corpus samples, as well as issues arising from the use of various taggers and lemmatizers, which are based on present-day data and neglect the language before the 20th c.

Retrieval of normalized or lemmatized forms of Arabic words alongside all their spelling variants is a breakthrough in the research of Arabisms. Other retrieval options by use of specific tags during text conversion still need improving, after careful planning of the corpus design and purpose.

While the methods described for diachronic corpora based on XML mark-up, on the use of the TEI *Guidelines*, and on use within the TEITOK platform bring obvious advantages, other factors must be taken into consideration. Thus, the above specifications are ideal for medium-sized, specialized corpora, but they demand heavy investment in terms of time, human resources, and careful work in the case of the reference corpora of millions of words cited at the beginning of the chapter. XML mark-up and the implementation of the TEI *Guidelines* are demanding too, require previous training, and may eventually result in substantial differences across projects, because they are not rigid directives.

Finally, the use of the methods presented for the ODE corpus and other corpora may help improve many dictionaries: They may carry exhaustive definitions, but they also find difficulties to cite consistent examples from a variety of text types for illustration of diatopic and diastratic variations in Spanish lexis.

Bibliography

Allés Torrent, Susana (2019): "Sobre la complejidad de los datos en Humanidades o cómo traducir las ideas a datos", in *Revista de Humanidades Digitales*, 4, pp. 1–28.

Calderón Campos, Miguel (2010): "Aspectos de la vida social granadina a través de diez arabismos de las actas del ayuntamiento y de las ordenanzas municipales (1492–1552)", in *Études romanes de Brno*, 2, pp. 179–192.

Calderón Campos, Miguel: (2015): *El español del reino de Granada en sus documentos (1492–1833). Oralidad y escritura*. Bern: Peter Lang.

Calderón Campos, Miguel (2019): "La edición de corpus históricos en la plataforma TEITOK. El caso de *Oralia diacrónica del español*", in *Chimera*, 6, pp. 21–36.

Calderón Campos, Miguel and Gael Vaamonde (2020): "*Oralia diacrónica del español*. Un nuevo corpus de la Edad Moderna", in *Scriptum digital*, 9, pp. 167–189.

CdE = Mark Davies (dir.): *Corpus del español*. <www.corpusdelespañol.org>

CDH = Real Academia Española: *Corpus del Diccionario histórico de la lengua española*. <https://www.rae.es/banco-de-datos/cdh>.

CHARTA = Belén Almeida Cabrejas (coord.): *Corpus hispánico y americano en la red: textos antiguos*. <https://www.corpuscharta.es/consultas.html>.

CODEA = *Corpus de documentos españoles anteriores a 1900*. GITHE, Universidad de Alcalá. <https://www.corpuscodea.es/>.

CODEMA = Carrasco Cantos, Inés (dir.): *Corpus diacrónico de documentación malagueña*, <http://www.arinta.uma.es>.
Colón Domènech, Germán (2002): *Para la historia del léxico español*. Madrid: Arco/Libros.
CORDE = Real Academia Española: *Corpus diacrónico del español*. <https://www.rae.es/banco-de-datos/corde>.
CORDIAM = Company Company, Concepción and Virginia Bertolotti (dirs.): *Corpus diacrónico y diatópico del español de América*, Academia Mexicana de la Lengua. <https://www.cordiam.org/>.
COREECOM = Arias Álvarez, Beatriz (coord.): *Corpus electrónico del español colonial mexicano*. <https://www.iifilologicas.unam.mx/coreecom/>.
CorLexIn = *Corpus Léxico de Inventarios*. Universidad de León. <https://corlexin.unileon.es/el-corpus/>.
CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <https://www.rae.es/banco-de-datos/corpes-xxi>.
Corriente, Federico (1977): *A gramatical sketch of the Spanish-Arabic dialect bundle*. Madrid: Instituto Hispano-Árabe de Cultura.
Corriente, Federico (1999): *Diccionario de arabismos y voces afines en iberorromance*. Madrid: Gredos.
CREA = Real Academia Española: *Corpus de referencia del español actual*. <https://www.rae.es/banco-de-datos/crea>.
DCECH = Corominas, Joan and José Antonio Pascual (1980–1991): *Diccionario crítico-etimológico castellano e hispánico*. Madrid: Gredos.
Díaz Bravo, Rocío (2018): "Las Humanidades Digitales y los corpus diacrónicos en línea del español: problemas y sugerencias", in Lidia Bocanegro and Esteban Romero Frías (eds.). *Humanidades Digitales aplicadas*. Granada/New York: University of Granada/Downhill Publishing, pp. 562–586.
Dozy, Reinhart Pieter Anne and Willen Herman Engelmann (1869): *Glossaire des mots espagnols et portugais dérivés de l'árabe*. Leiden: Brill.
Dworkin, Steven (2012): *A History of the Spanish Lexicon. A linguistic Perspective*. Oxford: Oxford University Press.
Eguílaz y Yanguas, Leopoldo (1886 [1974]): *Glosario etimológico de las palabras españolas de origen oriental*. Granada: La Lealtad.
Fontanella de Weinberg, María Beatriz (1993): *Documentos para la historia lingüística de Hispanoamérica. Siglos XVI a XVIII*, Anejo 53. Madrid: Boletín de la Real Academia Española.
García González, Javier (1998): "Clases de arabismos en los textos alfonsíes", in Claudio García Turza *et al*. (eds.). *Actas del IV Congreso Internacional de Historia de la Lengua Española*, vol. 2. La Rioja: Universidad de La Rioja, pp. 127–136.
García González, Javier (2007): "Una perspectiva sociolingüística de los arabismos en el español de la alta Edad Media (711–1300)", in Inmaculada Delgado Cobos and Alicia Puigvert Ocal (eds.). *Ex admiratione et amicitia: homenaje a Ramón Santiago*, vol. 1. Madrid: Ediciones del Orto, pp. 523–548.
Garulo Muñoz, Teresa (1983): *Los arabismos en el léxico andaluz (según los datos del Atlas lingüístico y etnográfico de Andalucía)*. Madrid: Instituto hispanoárabe de cultura.
Giménez-Eguíbar, Patricia (2015): "Dos casos de sustituciones léxicas: los arabismos alfayate y alfajeme", in Francisco Javier de Cos Ruiz and Mariano Franco Figueroa (coords.). *Actas del IX Congreso Internacional de Historia de la Lengua Española*. Madrid/Frankfurt: Iberoamericana/Vervuert, pp. 1413–1427.
Giménez-Eguíbar, Patricia (2016): "Attitudes toward Lexical Arabisms in 16th Century Spanish Texts", in Sandro Sessarego and Fernando Tejero-Herrero (eds.). *Spanish Language and Sociolinguistics Analysis*. Amsterdam/Philadelphia: Jonh Benjamins, pp. 363–380.

Giménez-Eguíbar, Patricia (2023): “La contribución del árabe al hispanorromance”, in Steven Dworkin *et al.* (eds.). *Lingüística histórica del español. The Routledge Handbook of Spanish Historical Linguistics*. Londres: Routledge, pp. 362–371.

Gómez Capuz, Juan (2004): *Los préstamos del español*. Madrid: Arco/Libros.

González Sopeña, Inmaculada (2017): “Arabismos y fiscalidad”, in *Dicenda*, 35, pp. 109–130.

Graham, Shawn, Ian Milligan and Scott Weingart (2015): *Exploring Big Historical Data. The Historian’s Macroscope*. London: Imperial College Press.

Gries, Stefan Th. (2009): “What is corpus linguistics”, in *Language and Linguistic Compass*, 3, pp. 1–17.

Janssen, Maarten (2012): “Neotag: a POS tagger for grammatical neologism detection”, in *Proceedings of the tenth international conference on Language Resources and Evalua- tion*. Istanbul, Turkey, pp. 1–7.

Janssen, Maarten (2016): “TEITOK: Text-Faithful Annotated Corpora”, in *Proceedings of the Language Resources and Evaluation Conference*. Portoroz, Slovenia, pp. 4037–4043.

Jódar Marín, Juan Ángel (2010): “La era digital: nuevos medios, nuevos usuarios y nuevos profesionales”, in *Razón y Palabra*, 71, pp. 1–12.

Lapesa, Rafael (1981): *Historia de la lengua española*. Madrid: Gredos.

Leech, Geoffrey (1992): “Corpora and theories of linguistics performance”, in Jan Svartvik (ed.). *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, pp. 105–122.

Maíllo Salgado, Felipe (1998): *Los arabismos del castellano en la Baja Edad Media*. Salamanca: Universidad de Salamanca.

Martínez-Gambia, Ricardo (2016): “Big data en humanidades digitales: de la escritura digital a la lectura distante”, in *Revista chilena de literatura*, 94, pp. 39–58.

Morala Rodríguez, José Ramón (2012a): “Arabismos en textos del siglo XVII escasamente documentados”, in *Revista de Investigación Lingüística*, 15, pp. 77–102.

Morala Rodríguez, José Ramón (2012b): “Léxico e inventarios de bienes en los Siglos de Oro”, in Gloria Clavería Nadal, Margarita Freixas, Marta Prat Sabater and Joan Torruella Casañas (coords.). *Historia del léxico: perspectivas de investigación*, pp. 199–218.

Neuvonen, Ereo Kalervo (1941): *Los arabismos del español en el siglo XIII*. Helsinki: Finnish Literature Society.

ODE = Calderón Campos, Miguel and María Teresa García-Godoy (2019-present) (dirs.): *Oralia diacrónica del español*. DiLEs, Universidad de Granada <http://corpora.ugr.es/ode>.

Oliver Pérez, Dolores (2004): “Los arabismos dentro de la historia del español: estudio diacrónico de su incorporación”, in Manuel Cecilio Díaz *et al.* (eds.). *Estudios dedicados a José María Fernández Catón*, vol. 2. León: Centro de Estudios e Investigación San Isidoro, pp. 1073–1095.

Parodi, Giovanni (2010): *Lingüística de corpus: de la teoría a la empiria*. Madrid: Iberoamericana.

Pocklington, Robert. (1984): “Nuevos arabismos en los textos alfonsíes murcianos”, *Miscelánea medieval murciana*, 11, pp. 261–295.

Real Academia Española (2014): *Diccionario de la lengua española*. Madrid: Espasa/Calpe.

Rojas, Elena (2008): *Documentos para la historia lingüística de Hispanoamérica (siglos XVI a XVIII)*, Anejo 61. Madrid: Boletín de la Real Academia Española.

Rojo, Guillermo (2016): “Corpus textuales del español”, in Javier Gutiérrez Rexach (ed.). *Enciclopedia de lingüística hispánica*, Vol. 2, pp. 285–296.

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

Romero Frías, Esteban (2014): “Ciencias sociales y Humanidades Digitales: una visión introductoria”, in Esteban Romero Frías and María Sánchez González (eds.). *Ciencias sociales y humanidades digitales: técnicas, herramientas y experiencias de e-research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social, pp. 19–50.

Sánchez González, María (2014): "El Big Data como herramienta para la *e-Research* en entornos infosaturados y complejos", in Esteban Romero Frías and María Sánchez González (eds.). *Ciencias sociales y humanidades digitales: técnicas, herramientas y experiencias de e-research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social, 131–161.

Sánchez-Prieto Borja, Pedro (1995): *Textos para la historia del español II. Archivo municipal de Guadalajara*. Alcalá de Henares: Universidad de Alcalá.

Schöch, Christof (2013): "Big? Smart? Clean? Messy? Data in the Humanities", in *Journal of Digital Humanities*, 2(3), pp. 2–13.

Spence, Paul (2014a): "Prólogo", in Esteban Romero Frías and María Sánchez González (eds.), *Ciencias sociales y humanidades digitales: técnicas, herramientas y experiencias de e-research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social, pp. 9–12.

Spence, Paul (2014b): "Centros y fronteras: el panorama internacional de las humanidades digitales", in Sagrario López Poza and Nieves Pena Sueiro (eds.). *Humanidades Digitales: desafíos, logros y perspectivas de futuro.* A Coruña: SIELAE-Janus, pp. 37–61.

Steiger, Arnold (1932): *Contribución a la fonética del hispanoárabe y los arabismos en el iberorrománico y siciliano*. Madrid: CSIC.

TEI Consortium (ed.) (2016): *TEI P5: guidelines for electronic text encoding and interchan- ge*. *Text Encoding Initiative Consortium*. <https://tei-c.org/guidelines/p5/>.

Torres Montes, Francisco (1996): "Nombres de medidas agrarias en la costa del antiguo Reino de Granada", in Juan de Dios Luque Durán (ed.). *Segundas jornadas sobre el estudio y enseñanza del léxico*. Granada, pp. 265–282.

Torruella, Joan and Johannes Kabatek (2018): *Portal de corpus históricos iberorrománicos* (CORHIBER). <http://www.corhiber.org/>.

Walsh, John (1967): *The Loss of Arabisms in the Spanish Lexicon*. Unpublished doctoral thesis. Virginia: University of Virginia.