Linguistic Corpora and Big Data in Spanish and Portuguese

Humanidades Digitales y Big Data en Iberoamérica

Digital Humanities and Big Data in Ibero-America

Editado por / Edited by
Ana Gallego Cuiñas y / and Azucena González Blanco

Volumen / Volume 4

Linguistic Corpora and Big Data in Spanish and Portuguese

Edited by
Miguel Calderón Campos and Gael Vaamonde

DE GRUYTER

Esta publicación es resultado de la Unidad Científica de Excelencia “Iber-Lab. Crítica, Lenguas y Culturas en Iberoamérica” (Ref. UCE2018-04) de la Universidad de Granada

ISBN 978-3-11-078145-8
e-ISBN (PDF) 978-3-11-078146-5
e-ISBN (EPUB) 978-3-11-078152-6
DOI https://doi.org/10.1515/9783110781465

Library of Congress Control Number: XXXXX

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

Cover image: as creative atelier/DigitalVision Vectors/Getty Images
Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

III Exploiting Portuguese Reference Corpora: The CdP and the CRPC Corpora

María Teresa García-Godoy

Big Data and Lexical History: Digital Newspaper Libraries in Spanish Diachronic Research

1 Introduction

The advent of the Royal Spanish Academy's (hereafter, RAE) first diachronic database in the late 20th c. has allowed free access to ten centuries of the history of Spanish texts in the form of a large annotated corpus ever since. A 250-million-word corpus covering from the earliest records to 1974, CORDE trailed the blaze for the big data resources that later became available for research on the history of Spanish and, thus, for the new paradigm of experimentally-based research that was to come. This is so much so that, all the RAE's databases, both of the earliest (CORDE, CREA) and of the latest generation (CDH and CORPES XXI) are considered reference corpora still today, and are essential for research on Spanish, whether present-day (CREA and CORPES XXI) or of the past (CORDE, CDH). Still, the RAE's databases are not enough for research on the history of diaphasic and diatopic variation, for two reasons: i) the genres where older colloquial forms are most likely to occur are underrepresented in these corpora; and ii) the geolocation of specific uses is not specific enough for research of intradialectal variation. The geographical classification of CORDE and CDH samples by country deny the possibility of research on regional or local forms. Thus, while the RAE diachronic corpora were the first big data resources available for diachronic research on Spanish, they are, paradoxically, also fairly limited as regards colloquial and dialectal diversity.

All the RAE corpora are based on a range of text types (fiction, notarial documents, chronicles, historiography, scientific treatises, etc.). Yet, not all the discourse genres are equally represented in the diachronic corpora, and this paper shows how the literary genres are primed over the non-literary ones. The resulting data unavailability can be noticed especially in documents of communicative immediacy, more likely to represent the spoken language in written (García-Godoy 2015). A lack of journalistic texts can also be felt in modern Spanish (18th and 19th centuries): Neither CORDE nor CDH cover a substantial amount of his-

Note: This contribution has been realized in the framework of Grant PID2022-136256NB-I00, funded by MICIU/AEI/ 10.13039/501100011033 and by 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/10.1515/9783110781465-006

torical press.[1] This neglect is particularly serious for the period 1750–1900, when the journalistic genre arose and grew in Spanish.

A lack of balance is also noticeable in the RAE diachronic databases as regards diatopic representativeness. Not all geolects are equally well represented in the CORDE and CDH corpora, e.g. American Spanish is represented by fewer documents than European Spanish and, in the latter, the variety of the *Meseta* prevails over other Spanish geolects. As the geographical classification in the diachronic corpora is by country, it is not possible to track down, e.g. the development of neologisms from their geographical origin. Language creativity and its geographical dissemination are easier to research from 1700 onwards, when the journalistic genre appeared. This is because, unlike other text types, journalistic texts, especially their sections of local news, always carry specific data on their day, month, year of publication, and on location, i.e. on key data for the geolocation of language variants.

Historical press is widely acknowledged as a major data source for research on diversity in modern Spanish, actually as one that is unparalleled by other text types in this regard. While the Spanish journalistic genre dates back to the 17th c., sparked by the spirit of the Enlightenment, journalism as we know it today developed from the first decree of press freedom issued by the Parliament of Cádiz (*Cortes de Cádiz*) in the 19th c. Since then, the Spanish press has proved a melting pot of texts of various types, topics and styles over the country. The early Spanish press discloses the inter(national), regional and local life of a specific time and thus depicts modern Spanish (i.e. of the 18th and 19th centuries) in ways that are not so noticeable in other genres. Current research thus relies on the digital newspaper libraries' search engines for retrieval of data unavailable from RAE corpora that may complete the picture of the language of the time, especially as regards diversity and the origin and development of colloquial forms in modern Spanish. The lack of experimental data of pre-20th c. spoken Spanish constrains diachronic research on diaphasic variation to the spoken language attested in written. As this chapter shows, journalistic texts from the 18th c. onwards have become data sources for colloquial forms and their geographical distribution which are hard to find outside newspaper libraries and, thus, largely outside of the genres recorded by reference diachronic corpora.

The relevance of texts for the account of language change is worth remarking. Their importance for research on the origins and development of specific cases has

1 The CORDE website lists the text types of the corpus. Fiction samples (poetry, prose, drama) amount to 40%, whereas press and commercial genres are only 8%. The CDH corpus, partly based on CORDE, classifies samples by topic instead of by genre, so the actual number of literary and press samples is hard to measure.

been underlined in the past decade by a number of authors (García-Godoy 2021). Epistemologically, journalism –and its range of topics– appears as a major genre where language change in the 18th and 19th centuries can be tracked. Due to the CORDE and CDH's neglect of newspaper texts, researchers turn to newspaper libraries of historical press for access to a wealth of evidence of pan-Hispanic language diversity. The digital newspaper section of the National Library (hereafter, HD) has become the most frequently used resource in this regard. It stores 2165 newspapers published nationally, regionally and locally in the Spanish speaking world of the period between the 18th and the 21st centuries (see section 3). Admittedly, a newspaper library's search engine cannot be compared with a computerized language corpus concordancer and the potential for data retrieval of its lemmatization, morphosyntactic annotation, or data size management, among others. At present, no newspaper library allows automatic data screening or analysis, so research on the evolution of Spanish based on digitalized press requires manual analysis of the data collected (see section 4.2.).

This chapter is intended to prove the relevance of the journalistic genre in the history of certain colloquial forms of modern Spanish (18th and 19th centuries), a crucial period for language standardization. Evidence of colloquial forms, specifically of the adverb *cabalito*, is compared between newspaper vs. other data sources, based on the HD and the CDH, respectively.

The chapter develops from this *Introduction* by way of four sections. First, a review of the state of the art on digital newspaper libraries for diachronic research on Spanish. Section 3 discusses corpora and methods. The history of *cabalito* then follows in Section 4 according to corpus evidence with regard to its diachronic, textual and diatopic properties, after its morphosyntactic description according to the grammars of reference. Section 5 deals with the formalization of *cabalito* in Spanish lexicographic practice over time, and then leads to the chapter's conclusions.

2 Digital Newspaper Libraries and Modern Spanish. State of the Art

In the past decade, large data bodies of historical press have raised the interest in diachronic research on Spanish, especially on modern Spanish. This is largely due to the development of online newspaper libraries, as they offer more and more evidence of the history of Spanish journalistic texts, so rare in Spanish corpora. Still, this resource is used for research on the evolution of Spanish in barely ten recent publications, mainly on lexical semantics and, less frequently, on the mor-

phosyntax of Spanish. The results obtained from digital newspaper libraries are worthy of special note for the following three processes of language variation and change: i) neology; ii) standardization; and iii) dialectalization.

In neology, the research on lexical and morphosyntactic change under the influence of the early journalistic genre stands out, especially in four sections of historical press: i) scientific dissemination; ii) crime and accidents; iii) advertisements; and iv) society. Campos Souto (2018) has proved that, for the former of these, HD is crucial for the compilation of the historical dictionary of Spanish, especially for the so-called specialty languages. Campos Souto argues that the digital contents of HD, whether generic or specialty, attest the earliest records of a large number of technical terms in the fields of medicine, music, or fashion between the 17th and the 20th centuries: Over 60 entries of these domains listed in the *Diccionario histórico del español* are first attested in the official press of the 18th c. (*gripe* 'flu', *guitarra* 'guitar', *violin* 'violin', *oboe* 'oboe', etc.), of technical and scientific journalism of the 19th c. (*gripal* 'flu-related', *saramponioso* 'infected with measles', *corsetería* 'corsetry', *cubrecorsé* 'camisole', . . .) and, particularly, of the press in general of the 20th c. (*griposo* 'infected with flu', *lepra*, 'leprosy', *violonchelista* 'cello player'). García-Godoy's (2015, 2017) lexical research also proves that lexical innovations currently in use date back to the political press between 1813 and 1823, e.g. *retaliación* 'retaliation', *retaliar* 'retaliate', *complotado* 'conspiror', *sufragar* 'vote', and neological senses of *departamento* 'department', *cantón* 'canton', *cantonal* 'cantonal', or *prefectura* 'prefecture' for administrative divisions. HD is also used for attestation of new morphosyntactic formations, e.g. *madama* 'Mrs.' / *madamita* 'Miss' under the influence of French in 18th c. society news (García-Godoy 2021), and the prepositional locution *a por* 'towards' with motion verbs in advertisements and in crime and accidents sections of the 19th c. (Company Company & Flores Dávila 2017, 2018).

In the references cited above, HD is used not only for the identification of lexical and morphosyntactic neologisms in journalistic texts, but also to assess their diatopic spread prior to standardization. The precise chronological dating of each neological use in journalistic texts by nature makes digital newspaper libraries a key data source for research on the convergence and divergence between European and American Spanish from the 18th c. onwards. Regarding European Spanish, Company Company & Flores Dávila (2017, 2018) and García-Godoy (2021) claim that *a por* 'towards' and *madama* 'Mrs.' / *madamita* 'miss' start out in the press of Madrid, wherefrom they quickly reached the rest of the country by imitation of the new formations of the Spanish of the court as a standard of prestige.

Regarding American Spanish, García-Godoy (2015, 2017) shows how the earliest attestation of public lexical divergence between the colonies and the metropolis can be found in the pro-independence press of the American emancipation period

across South America (Colombia, Mexico, Argentina, Venezuela). These divergences are today viewed as specifically American, century-old formations. Diatopically, these first instances of research based on HD evidence focus on European vs. American Spanish as two geolects separated by the Atlantic as an isogloss. Thus, geolocation of HD data help identify large-scale continental dialects taking the Atlantic as a boundary line, even if further research at a smaller scale is needed, and where HD may also supply earliest attestation data. In the latter respect, and despite the large amount of regional and local records available from the HD, the so-called local terms are in need of research with regard to the journalistic genre. Octavio de Toledo's (2016) "Sin CORDE pero con red: *algotras* fuentes de datos" pioneered this field by defining boundaries on the *terra incognita* of the historical map of dialects in the Iberian peninsula, based on online sources (e.g. specific subject field websites, data retrieval from *Google Books* by specific search engines), whenever the RAE corpora failed to supply the necessary data. This author does not use digital newspaper libraries for his research on *algotro* (and its morphological variants), but his methods are very similar to those reviewed in this section in that language traits unsupported from data available from reference corpora and metalinguistically marked in diatopically vague and contradictory ways in historical dictionaries are researched based on online resources searchable with specific engines. Against traditional dialectal lexicography, Octavio de Toledo thus proves that the quantifier *algotro* comes historically from Extremadura instead of Andalusia, and that it eventually was also used therefrom in American Spanish.[2]

3 Corpora and Experimental Methods

This chapter is based on three complementary databases: i) HD for the history of the journalistic genre; ii) CDH as a control corpus: as it covers a range of genres (except for press texts), it is used here to tell whether the journalistic genre made wider use of *cabalito* or not compared with HD; and iii) the *Nuevo Tesoro Lexicográfico de la Lengua Española* (hereafter, NTLLE) for the occurrence of *cabalito* in the history of Spanish dictionaries.

Our base corpus, the HD, focuses on press language and covers digitally available newspapers published between the 18th and the 21st centuries. The chronological frame is defined by HD's earliest and latest attestation of *cabalito* (1790–2022). The 2022 actualization of the virtual HD relies on 2,165 titles and over

2 Calderón Campos (2023: 119–121) uses the EsTenTen18 macrocorpus (Sketch Engine) and the social network X to analyze the current stigmatization of *algotro* in America.

seven million pages in digital format, with European press prevailing over American press, as mentioned above. This big data source of historical press offers the largest body of data on the history of *cabalito* currently available, namely 327 attestations between the 18th and the 21st centuries compared with 27 occurrences in the 3.5-million-word CDH corpus (specifically, 355,740,238 words) used for the RAE's historical dictionary.

The 322 years' span is divided into 30-year sections for identification of any neological stage and its subsequent evolution stages. Regarding the contents, the corpus is a scale model of the HD in that it accounts for the range represented in the HD too: Politics, satire, humour, science, religion, cultivated, leisure, sports, arts, literature, etc. This connection with the HD also shows in the bias towards Spanish press to the detriment of American press, and in the availability of regional and local press only for European Spanish.

As for data collection, the HD application allows a range of queries and access to the digital version of the original texts. Data collection is far from perfect, as it lets in undesired cases in the search for *cabalito*, e.g. *caballero* 'gentleman', *caballito* 'little horse', *cabellera* 'mane', etc., to be discarded manually. This limitation, largely inherent in newspaper libraries, precluded automatic compilation of the base corpus and required further data selection for identification and separation of so-called false positives. Even so, HD proved essential for the experimental data attestation used in this chapter.

On the other hand, the history of Spanish texts parallels in the main the RAE's diachronic corpora, namely CORDE and CDH. The latter shows major advantages by virtue of its three layers of documents, and they justify its use as a control corpus: a) a higher philological control of the sources; b) lemmatization; and c) morphosyntactic annotation.

The HD evidence is contrasted with CDH evidence from outside newspaper sources. Section 4 presents data analysis according to text type, diachrony, and diatopy. In the latter, geolocation data attested in historical press are more precise and homogeneous than in the rest of text types. This is because only newspapers data can count on exact production dates (day, month, year) and location.[3] The third corpus, of a lexicographic kind, covers Spanish dictionaries over time and relies on the NTLLE for the attestation of *cabalito* in (non-)academic lexicographic practice.

3 Early newspaper practice often published one and the same text on various dates and in various locations. For more accurate chronological attestation, the earliest date is recorded here for neologisms, even if the qualitative analysis takes account of all the published instances of the same use.

4 The History of *cabalito* 'exactly' in Corpora. Grammatical Status and Use Attestation

Some diminutive adjectives can be used as adverbs in present-day Spanish (*justito* 'just.DIM', *rapidito* 'fast.DIM'). In this context, the case of *cabalito* is a bit of a mystery for the following reasons: i) grammatically, this adverb has been neglected by both synchronic and diachronic research; and ii) the reference corpora do not attest examples for the 21st c., and even past evidence is scarce. The following sections show the sharp contrast between this lack of evidence of *cabalito* and the frequent use in historical press.

The following is intended to cast light based on the analysis of the evolution of *cabalito* in Spanish by identification of its morphosyntactic, diachronic and diatopic profiles. To that end, the state of the art is first reviewed and the data collected for *cabalito* from HD as the base corpus and from CDH as the control corpus are then contrasted and discussed.

4.1 Diminutive Adverbs in Spanish Grammar

According to the reference grammar (NGLE: § 9.2), the adjectives that can be used as adverbs often do so in their diminutive forms too. This has been described synchronically, and has been reported to be more frequent in American Spanish than in European Spanish. Table 1 shows the contrast according to dialect, here

Table 1: Diminutive adverbs in NGLE. Cross-dialectal examples.

American Spanish	American and European Spanish
ahicito	cerquita
ahorita	despacito
alrededorcito	poquito
allacito, allicito	prontito
apenitas	
aquicito	
antesito	
despuesito	
detrasito	
enantito	
nomasito	

illustrated with a dozen cases available only in American Spanish vs. three cases available in both American and European Spanish.

Gerhalter (2020: 190–194) confirms this dialectal contrast in her account of the paradigm of focusing adverbs, *cabalito* among them. In this paradigm, seven adjectives/adverbs are listed, and three of them have a diminutive form: *preciso* 'precise[ly]', *exacto* 'exact(ly)'/*exactito*' exact(ly)-DIM', *justo* 'just'/*justito* 'just-DIM' and *cabal* exact(ly)'/*cabalito* 'exact(ly)-DIM'. Gerhalter illustrates the use of the three diminutive adverbs (*exactito*, *justito*, and *cabalito*) with examples by South American literary authors recorded in the RAE corpora. Still, while the first two are illustrated with 20th c. examples, *cabalito* is illustrated with evidence of the 19th c. The suffix *-ito* in the adverbs *exactito*, *justito*, and *cabalito* adds an affective and emphatic nuance of meaning, but only in *cabalito* does Gerhalter underline the following difference: It is a statement marker in 19th c. use (dated 1875), as in examples by the Peruvian author Ricardo Palma (1).

(1) El padre Arce quedó un minuto pensativo; y luego, pegándose una palmada en la frente, como quien ha dado en el quid de intrincado asunto, exclamó: "¡*Cabalito*! ¡Eso es!" (Ricardo Palma, *Tradiciones peruanas*, third series, 1875. CDH, cited by Gerhalter 2020: 191).

All in all, the little evidence available of the adverb *cabalito* suggests that only this adverb developed specific pragmatic values, as attested in the American variety at least since the late 19th c. The following questions, unanswered so far, can then be asked to give shape to the following section on data analysis: Where, when and how did this functional peculiarity of *cabalito* arise? Is this grammatical profile of *cabalito* limited to the 19th c.? Is *cabalito* historically one of the few diminutive adverbs available both in European and in American Spanish, or is it only in American Spanish?

4.2 *Cabalito* in Corpora

This is a contrastive analysis of the diachronic use of *cabalito* in HD (corpus base) and CDH (control corpus), according to four variables: i) evidence of use and morphosyntactic profile; ii) chronological attestation; iii) text types; and iv) diatopy.

4.2.1 Evidence of Use and Morphosyntactic Status

Table 2 shows a much stronger attestation of *cabalito* in the base corpus (317 occurrences) than in the control corpus (27 occurrences). The former, 317 occurrences in HD, are dated between 1790 (example 2) and 2022 (example 3), whereas the latter, 27 occurrences in CDH, are dated between 1771 (example 4) and 1977 (example 5).

HD attests four early examples of *cabalito* in the 18th c. (1790–1799) by contrast with one example in CDH (1771). The latter is, incidentally, the earliest attestation available. Thus, the 18th c. examples of *cabalito* in the base corpus are four times as many as the ones in the control corpus. This lack of balance grows exponentially in the 19th and 20th centuries. The two sources also differ widely as regards current use: *cabalito* is attested for the 21st c. in HD, but it is not in the RAE corpora.

The contrast in the number and in the chronological distribution of examples suggests that HD and CDH are two entirely different accounts of *cabalito*, as will be shown below.

Table 2: Evidence of *cabalito* in the base corpus vs. the control corpus (18th-21th centuries).

	BASE CORPUS (HD)	**CONTROL CORPUS (CDH +CORPES XXI)**
18th c.	4	1
19th c.	160	19
20th c.	127	7
21th c.	26	0
	317 ocurrences in 302 documents	CDH: **27** ocurrences in 14 documents

(2) 1790. Madrid. Señor Editor: he visto en el n. 403 que el Caballero A.C. escribe quejándose del Señor *Quiqondam* y de mí. ¡Válgate Dios, que nunca hemos de poder contentar a todos! [. . .] Unos lloran de lo que otros ríen *cabalito*: eso es el mundo (*Correo de Madrid o de los ciegos*, Letter to the editor, 27/10/1790, page 7).

(3) 2022. Toledo. No tengo memoria de mi río. Nací, justo, *cabalito*, aquel año que lo robaron (*ABC*, Toledo edition, 19/06/2022, editorial article on local topics, page 65).

(4) 1771. Madrid. En el Lavapiesillo, / por el verano, / de aquesta forma cantan / majas y majos: "Que si ronda mi calle / Paco el herrero, / no le importa a ninguno, *cabalito* / (ea, ea, ea, ea, ea, ea), / segurito / (ea, ea, ea, ea, ea, ea.). No le importa a ninguno / y yo requiero" (Anonymous, "El juego del burro. Tonadilla a tres" in *Tonadillas teatrales*. Madrid: Tipografía de Archivos, 1932. In CDH: Fiction/poetry).

(5) 1977. Perú. "Los amores de un bebé y una anciana que además es algo así como su tía" –me dijo una noche la tía Julia, mientras cruzábamos el Parque Central–. "*Cabalito* para un radioteatro de Pedro Camacho" (Vargas Llosa, Mario, *La tía Julia y el escribidor*, Barcelona: Seix Barral. In CDH: Fiction, novel).

As for the grammatical status of *cabalito*, the earliest and latest attestations in the two sources are evidence of adverbial use alone, as this is the prevailing use in the two corpora. The quantitative analysis of all the occurrences shows that the adjectival use of *cabalito* is not attested in CDH and is rare in HD. Actually, only three out of the 317 examples of the base corpus are used as adjectives (see 6–8 below), and the rest are instances of adverbial use. Overall, the diminutive adverb *cabalito* is, typically, an adverb used during the course of ordinary spoken interaction. In these interactions, *cabalito* starts turn-taking and is often used for the expression of agreement ('language accuracy'), but it may also be used for disagreement in contexts like (4), ironical and jocular, where ordinary interaction takes place by the less privileged social strata in the area known as *Lavapiés*, in Madrid. Such syntactic and semantic specialization of *cabalito* becomes its main sign of identity and is a regularity of diachronic use attested in the two corpora. Besides language accuracy, and much less prominently, the adverb *cabalito* may also denote numerical or mathematical accuracy in the measurement of time, weight, etc. This semantic nuance of meaning, rarely attested in the corpora, is illustrated with example (3), where *justo* and *cabalito* are used as synonymous adverbs.

(6) 1884. Burgo de Osma. *Cosas y casos*. [. . .] Caballeros, les advierto a ustedes que soy ilustre [. . .] y aun cuando juzgan que soy el loco de la casa, no dejo de tener mi juicio muy *cabalito* (*La Propaganda: revista quincenal de intereses materiales, ciencias y literatura*, local news, 14/11/1884).

(7) 1966. Madrid. [Informaciones de espectáculos]. Tercera de la Feria granadina. Falleció esta madrugada don Ricardo Calvo, dos toreros y un toro. Granada. El conde de la Corte envió un encierro *cabalito* de peso. Hubo uno de 435 kilos y otro de 437. Seguramente dos perdieron 100 gramos en el

viaje y no pudieron pasar. ¡Estos toritos que ponen tan poco de su parte! (*Informaciones*, bullfight news, 13/06/1966, page 13).

(8) 2008. Sevilla. Si Pepín Liria se despide de El Puerto con un indulto, el público *cabalito* que conservó sus entradas, aunque no cayó nada bien la sustitución que anunció la empresa el mismo día (*Diario de Sevilla*, bullfighting, culture and leisure, 16/08/2008,[4] page 50).

4.2.2 Diachronic Evidence: Stages in the Evolution of *cabalito*

The lifespan of *cabalito* is over two centuries according to the base corpus, specifically 232 years from the earliest attestation, in 1790, to the latest in 2022. As described above, this period has been divided into 30-year segments of time, and each segment counts at least on ten attestations. By contrast, ten occurrences per segment is the highest record in the control corpus, and it is available only in one of the eight chronological segments (1821–1851). Even more, no attestations of *cabalito* are available from the earliest and latest chronological segments of CDH, namely 1790–1820 and 2007–2022, respectively.

Table 3: HD and CDH diachronic evidence of *cabalito*: Chronological distribution.

	Base corpus (HD)	Control corpus (CDH)
1771	0	1
1790–1820	10	0
1821–1851	35	9
1852–1882	68	9
1883–1913	95	2
1914–1944	62	1
1945–1975	12	4
1976–2006	18	1
2007–2022	17	0

4 This bullfight piece of news was published the same day in the *Diario de Cádiz* (page 42) and in the *Diario de Jerez* (page 39).

Table 3 shows four landmarks in the evolution of the use of *cabalito* according to the base corpus: i) the earliest records (1790–1830); ii) the earliest evidence of stabilization (1831–1851); iii) standardization (1852–1952); and iv) obsolescence from the mid-20th c. up to present-day. These stages are described in detail below.

Cabalito is recorded as a neologism in the base corpus between 1790 and 1830, as this is the segment where the earliest ten examples are attested in ten sources. As mentioned above, the earliest records date back to the last decade of the 18th c. During this period, this use started to seep into written Spanish, even if it must have been in the spoken language at least since the mid-18th c.

The number of attestations in the base corpus increased sharply in the second stage (1831–1851), namely more than three times as many compared with the former segment of time: From 10 to 35 instances. This record shows that *cabalito* is no longer a neologism and has become stable in the history of Spanish texts by the mid-19th c.

The third stage spans a century of general change (1852–1952), during which the use at issue becomes standardized. This period at the turn of the 20th c. attests the highest number of examples of *cabalito*: Compared with the first segment above, the number of occurrences is between 6 and 9 times as high: Nearly 100 of the 317 records of the base corpus date from this period and, as a result, standardization can be said to peak during this period in the evolution of *cabalito*.

The last segment covers from the mid-20th c. up to present-day. The figures decrease markedly and fall below 20 instances, i.e. between 4 and 8 times less than in the former stage.

The picture obtained from the control corpus is quite otherwise. Unlike the base corpus, where the case under study is attested by over 300 authors, the control corpus gives evidence of only 11 occurrences. According to the CDH, only one case is attested in the 18th c. and another in the first half of the 19th c., specifically in 1832, so the use of *cabalito* in these two periods is quite marginal. Still, the second half of the 19th c. attests 11 occurrences. Thus, while the control corpus suggests a marked decrease in the use of *cabalito* at the turn of the 20th c., the base corpus, remarkably, suggests the opposite and gives evidence of the widest spread recorded. No attestation is available from the RAE corpora (CDH and CORPES XXI) for the period 1971–2023, so *cabalito* appears to be falling out of use, even if the base corpus confirms the sustained occurrence from the 18th c. up to present day.

Finally, the morphosyntactic status of *cabalito* develops differently in the two corpora too. All the CDH occurrences are adverbs, whereas the base corpus shows that *cabalito* was used also as an adjective from the third stage onwards, which is when it became most widely used: Four HD records dated from 1884 onwards evidence an occasional use as an adjective, even if the use as an adverb prevails and is attested in all the four stages considered here. All in all, CDH data, scarce and

chronologically sparse, attest only the adverbial use, and a rare, peripheral use as an adjective. As shown in Table 3, the base corpus shows the opposite.

It should also be underlined that the data of *cabalito* in the base corpus do not allow, strictly speaking, the identification of evolutionary cycles according to relative frequency. The technological resources of newspaper libraries do not offer statistical accounts of *cabalito* over time based on accurate quantitative methods. The results presented here are based on manual counts of absolute frequencies after due data selection. Despite the difficulties inherent in the use of HD as a source of unencoded data, the historical press reveals a longer lifespan and a wider distribution of *cabalito* than the RAE corpora suggest. Note that the morphosyntactic tagging used in the latter corpora does not allow quantification of diachronic tendencies: Each occurrence of *cabalito* must be marked as adjectival or adverbial manually, because the CDH tagger marked each occurrence of *cabalito* both as an adjective and as an adverb. As a result, only after manual analysis can it be ascertained that all the 27 occurrences of *cabalito* in the CDH verify the adverbial use, and that no evidence of the adjectival use is available.

4.2.3 Textual Evidence

Table 4 shows the journalistic genre's major role in the development of *cabalito*, as it is attested historically in two text types: i) HD newspaper texts; and ii) literary texts, extensively covered by CDH. Most of the occurrences of *cabalito* in CDH are recorded in fiction texts, whether poetry or prose. Out of 27 instances, only 4 are from non-literary sources: They are from a volume on bullfight news written in 1970 by Díaz-Cañabate. As the news had already been published in specialized

Table 4: The occurrence of *cabalito* in two text types.

	Newspaper	Literary texts	Total
1771	0	1	1
1790–1820	10	0	10
1821–1851	35	9	44
1852–1882	68	9	77
1883–1913	95	2	97
1914–1944	62	1	63
1945–1975	16	0	16
1976–2006	18	1	19
2007–2022	17	0	17
Total	**321**	**23**	**343**

forums (sports publications), these four instances are listed under the journalistic genre (Table 4, row 7), whose quantitative relevance compared with literary texts appears at the foot of the table: 321 occurrences in newspapers vs. 23 in fiction texts, i.e. nearly 14 times as many.

Diachronically, the journalistic genre influences the four stages of the evolution of *cabalito* above especially heavily during the stages of innovation (1790–1830) and standardization (1852–1952). Most of the occurrences of *cabalito* in the two text types under consideration occur between 1852 and 1944: 225 instances are from the press, whereas CDH attests 12 occurrences by six authors (Ascasubi, Gaspar Enrique, José María de Pereda, Ricardo Palma, Jacinto Octavio Picón, and Eduardo Blanco). While the latter CDH data may suggest that *cabalito* was specific of literary texts, the former evidence proves that the press was the true catalyst for the standardization of *cabalito*. The journalistic genre also comprehends a number of text subtypes, so *cabalito* is recorded in news on a range of topics: Parliamentary news (9), opinion articles (10), letters to the editor (13), editor's releases (11), and even bullfight news (14).

(9) 1838. Madrid. [Intervención acalorada del Sr. Sancho, diputado por Valencia] El Sr. SANCHO: "Pues ahora contesto que ese cálculo es monstruosamente exagerado [. . .] El Sr. Martínez de la Rosa dice, que si no damos el diezmo, el clero se queda sin comer; pues yo digo lo contrario, si damos el diezmo el clero se queda sin comer. *Cabalito*. [. . .] Para mí es inconcebible" (*El Correo nacional*, 31/5/1838, page 3).

(10) 1845. Madrid. [Artículo de opinión] Conociendo el poco valor de semejante testimonio, se apresuró a pronosticar que los que habíamos dicho que la carta anterior era falsa, diríamos que también lo era la nueva. *Cabalito*: la misma fuerza nos hace la una que la otra (*La Esperanza*, Madrid, 20/11/1845, page 1).

(11) 1882. Burgos. [Nota del editor en respuesta a la de dos lugareños que se denominan "cándidos"] si ustedes son cándidos hay que meter en la cárcel a los innumerables mártires de Zaragoza, *cabalito* (*El Papa Moscas: periódico satírico*, Burgos, 04/06/1882, page 2).

(12) 1891. Madrid. [Correspondencia particular] Si son pocos los números que le faltan, tal vez podamos remitírselos. Y eso sería lo mejor. *Cabalito* (*Madrid Cómico*. 21/02/1891, page 7).

(13) 1894. Soria. [Carta al director] ¿Construcciones o ruinas? [. . .] Ya veo que se me va a hacer una pregunta suelta: ¿Y la Sociedad de socorros mutuos de Soria qué piensa de construcciones? *Cabalito*, ciudadanos. Formada esa sociedad obrera de pobres y ricos; confundidos en ella los de chaqueta con los de levita; hermanadas las ideas y los propósitos, deberíamos todos tomar un nuevo rumbo. [Firma el Pobrete de la clase] (*El Noticiero de Soria*, 03/03/1894, page 2).

(14) 1952. Logroño. [Taurinas] En Madrid sale en hombros un logroñés. Fenómeno habemus. . . Ya tiene La Rioja su torero. . . De pocos días data el sensacional descubrimiento. . . pero el hecho es cierto, *cabalito*, sin lugar a la más ligera duda (*La Rioja. Diario político*, Logroño. 23/11/1952, page 3).

Still, most of the occurrences of *cabalito* in newspaper texts come from the section on society news about various events. This highly popular section started in the 19th c. to cover local news, gossip, tales, humour and all kind of reviews about current issues, all of which were key to a newspaper's commercial viability. From 1840 onwards, the section also covered serial fiction, with novels in serial form. In the base corpus, *cabalito* is attested in these within excerpts of simulated orality, similarly to others in CDH (see examples under 1, 4 and 5 above). From the mid-19th c. onwards, the highly successful serial fiction takes over the contents of the society news, so the news on current issues are diverted to smaller sections of variety, gossip, and the like. Thus, the data on newspaper texts from the mid-19th c. onwards of Table 4 may also include excerpts of novels. This section contains 12% of the occurrences of *cabalito* in press texts, remarkably both literary and most of the non-literary too. The authors of this section in the 19th c. are typically fond of lexical fashions and of the most colloquial registers. Whether literary or not, the contents of the society news display features of less elaborate, rushed writing than in other newspaper sections, they forerun yellow press, and they also use local vocabulary that is eschewed in other text types. Thus, the society section discloses, from the mid-19th c. onwards, a more representative gamut of colloquial forms of the time than the literary canon does.

Overall, only HD offers evidence of *cabalito* both in fiction and non-fiction. For over two centuries, the attestation of this diminutive has been limited to the production by a small group of 11 authors, but the newspaper corpus proves that *cabalito* was not just a stylistic device of the poets and novelists of the literary canon since the 18th c. The journalistic genre also has an added value in that it gives evidence of the language proper to press language, but also of fiction texts intended for the masses. As will be shown below, the authors of such sections write informally, for a wide readership, often under a pen name, and become

both major participants and unique informants as regards language diversity, especially of the diaphasic and diatopic kind.

4.2.4 Diatopic Evidence

Diminutive adverbs are described above as more frequent in American than in European Spanish. Their geolectal nature is often underlined in the literature, and specific forms are recorded only for American Spanish (Table 1). Still, the history of the adverb *cabalito* seems to show that specifically European forms may have been available too. The diatopic indicators of the base corpus suggest that *cabalito* was a divergence in European Spanish since its earliest occurrences in the 18th c. until its standardization at the turn of the 20th c., as shown in Table 5.

Table 5: Diachronic interdialectal distribution of *cabalito*, where the first figure is the number of American examples and the second figure is the number of European examples.

	Base corpus (HD)	Control corpus (CDH)
1771	0/0	0/1
1790–1820	0/10	0/0
1821–1851	0/35	2/7
1852–1882	2/66	6/3
1883–1913	0/95	1/1
1914–1944	2/60	0/1
1945–1975	0/12	0/4
1976–2006	0/18	1/0
2007–2022	0/17	0/0

Pan-hispanically, the geolocation of the HD data refer virtually all the occurrences of *cabalito* to European Spanish (313 out of 317 occurrences), and only four to American Spanish (Cuban and Argentinian Spanish as in 15–17). At this point, it is in order to underline that *cabalito* is not recorded in the *Corpus Diacrónico y Diatópico del Español Americano* (CORDIAM). As this corpus covers a wide range of genres (letters, legal and administrative texts, literature, press) dated between 1494 and 1905, *cabalito* then becomes more strongly associated with European Spanish, as suggested by the base corpus data. Table 5 shows that the HD data suggest an extensive use in European Spanish press and quite the opposite in American Spanish press. Both CORDIAM and HD data suggest that *cabalito* has

little history American Spanish: The four occurrences in American Spanish attested in the base corpus are dated between 1860 and 1928. *Cabalito* must have been used rarely outside Spain during this period, with occasional instances in the literary section of South American press. The examples 15–17 show that the fiction dialogues of some American newspapers have used *cabalito* occasionally since 1860, i.e. 70 years after the earliest records of the European Spanish press.

Unlike CORDIAM and HD, CDH data reveal a well-balanced distribution of *cabalito* between American and European literary authors in the control corpus, so a more detailed interdialectal analysis is in order. All the instances contained in CDH are by eleven authors, five American (Ricardo Palma and Vargas Llosa from Peru, Manuel Eduardo de Gorostiza from Mexico, Eduardo Blanco from Venezuela, and Hilario Ascasubi from Argentina) and six European (Mariano José de Larra, Ayguals de Izco, Antonio Díaz Cañabate, Enrique Gaspar, Jacinto Octavio Picón, and José María de Pereda). Thus, out of the 27 CDH occurrences, 10 are American and 17 are European. Of these, the former also suggest the opposite evolution to the one obtained from the base corpus data: The latest attestation is recorded in Peru in 1977, and even for the period 1852–1882 the RAE corpus lists more occurrences of American Spanish (6 occurrences) than of European Spanish (3 occurrences). By contrast, the base corpus points in the opposite direction based on six times as many examples than in CDH: Only 2 out of 68 occurrences of *cabalito* are from American Spanish. The 66 records of European Spanish are thirty times as many as those of American Spanish, and they grow steadily in the base corpus too. The conflict between such opposite interdialectal data may be as a result of CDH's design, whereby fiction by canonical authors is primed and the journalistic genre, which is where the form under study is more likely to occur, is largely overlooked. The balance between American and European authors in CDH may be due to the literary connections between educated authors across the Atlantic, such that their exchange of texts may have been an elite channel for the specific dissemination of lexical fashions at various points of the Spanish speaking countries.

(15) 1860. Cuba. *Memorias de una viuda. Mi segundo marido* (continuación) [. . .] Ah! Boca de serafín. . . Dios te guarde, pico de oro. . . acertaste. . . *cabalito, cabalito*. Estuve en la gloria con aquel bribonzuelo (*El Moro Muza*, La Habana, 08/01/1860, page 3).

(16) 1869. Cuba. ¡*Un artículo de punta*! [editorial] No sé si lo de *punta* querrá decir artículo agudo, ó tal vez *punta* que hiera. [. . .] *D. Pacuato* quiere hacer milicianos nacionales, se insurreccionan combatiendo la nacionalidad

española, para lo mismo que en Alcolea *¡cabalito!* . . . para dar la libertad al país (*El Moro Muza*, La Habana, 09/05/1869, page 1).

(17) 1928. Argentina. *Los regionales. Pava y Varita*, por Fausto Burgos. [. . .] En clase, a hurto del catedrático, sacaba yo los billetes para contarlos y recontarlos. "Te ha dado justo, ¿che?", preguntaba Chumbo. "*Cabalito*". No veíamos la hora de llegar a la calle, de llegar a nuestra casa de huéspedes [. . .] "¿Te dieron de más?". "¡*Cabalito*!". Y echábamos cuenta: 200 x 400 = 80.000 pesos (*Caras y Caretas*, Buenos Aires, 01/12/1928, pages 150–151).

As the for the intradialectal properties of *cabalito* in European Spanish, the questions arise: When did this innovative use start, and how widespread did it become in European Spanish? As noted above, only the base corpus can supply geolocation by country, province, and specific location, in addition to the date and place of publication. The analysis of the diachronic diatopic data shows that the earliest records date from the last decade of the 18th c. and refer to publications in Madrid. Certainly, the highest number of attestations are from Madrid throughout the eight segments of time under consideration here. Newspapers from other places can be added at the turn of the 19th c. as early records, i.e. as records from before 1815: (18) from Zaragoza, (19) from Cádiz and (20) from Seville. From 1815 onwards, the number of publication places increases steadily until widespread occurrence over the Spanish press in the first decade of the 20th c. According to the base corpus, the dissemination amounted to newspapers published in up to 43 locations. The diatopic distribution of the occurrences of the base corpus is shown by region in the tables of Appendix 1. Their data are graphically represented in the following Map 1.

(18) 1798. Zaragoza. *Carta publicada en los Diarios de Madrid números 199 y 200 del miércoles y jueves, 18 y 19 de julio de 1788* [. . .] Por qué el adverbio *souvent* le (sic) traduce ahí con frecuencia y vele (sic) aquí, quatro líneas más abaxo, de *quando en quando*? "*Cabalito*", dixo él: "souvent lo mismo es que freqüentemente"(*Semanario de Zaragoza*, Zaragoza, 13/08/1798, page 3).

(19) 1813. Cádiz. ¡Qué poco calcula el que tal cosa propone! Peor lo habíamos de pasar: *cabalito*; peor (*El Duende de los cafées*, Cádiz, 08/08/1813, page 6).

(20) 1814. Sevilla. [. . .] ¡Hombre necio! Si la casa se está quemando! Piensa en socorrerla, y luego en adornarla. *Castaña*. ¡*Cabalito*! Pero agregue usté a eso las otras dos disposiciones (*La Tía Norica*, Sevilla, 1814, no. 20, page 4).

Map 1: *Cabalito* in Spanish newspapers (HD, 1790–2022). Intradialectal diatopic distribution.

AU: Please mention Map 1 in the text.

Attestations of *cabalito* can be found in the local press during the standardization stage, e.g. in newspapers of Jerez (province of Cádiz), Orihuela (province of Alicante), Burgo de Osma (province of Soria), Gandía (province of Valencia) and many others, both in reprints of news of Madrid and also in local news. These records bear witness to the wide spread of this colloquial term in Spain until the mid-20th c. From this point onwards, its distribution in the press becomes much more limited as *cabalito* becomes gradually obsolescent. By the 21st c., the attestations in the base corpus are occasional and mainly limited to the region of Castilla-La Mancha. Further research may reveal whether *cabalito* remained only in this geolect and fell out of use elsewhere.

The final picture emerging from the base corpus presents *cabalito* as an European singularity arisen in Madrid in the last decade of the 18th c. and used in the rest of the country one century later after dissemination between 1850 and 1950. According to the HD, *cabalito* receded mainly in the 21st c., as it seems to have become a dialectal form used in Toledo and nearby areas. This adverb therefore evolved from a catchy colloquial form of Madrid in 1790 to a frequent term in the press of the country for two centuries and, finally, probably a mere diatopic form of La Mancha in the 21st c.

5 Encoded Use: *Cabalito* in the History of Spanish Dictionaries

An innovative use's relevance for metalinguistic analysis and its full acceptance by a language's reference codes are clear evidence of standardization. This section researches the encoding stage of *cabalito* in the history of Spanish dictionaries. The corpus of dictionaries used here covers all the academic and non-academic references contained in the NTLLE, i.e. over 90 titles published between the 15th and the 20th centuries. The 21st c. is here accounted for by academic and non-academic references too, namely the electronic editions of the RAE Dictionary (DLE 2023) and of the *Diccionario del español actual* (DEA 2023). The latter's diatopic coverage is limited to European Spanish.

The aim is to find out when and how the term *cabalito* was recorded in the history of Spanish dictionaries according to the gloss type, grammatical classification, use marks and examples. Table 6 shows in chronological order (1853–2023) how *cabalito* was recorded in academic and non-academic dictionaries, and classifies the four variables listed above as columns, where the leftmost column shows the models recorded by the dictionaries as in examples 21 through 23.

Table 6: *Cabalito* in the history of Spanish dictionaries.

Year, author	Gloss	Grammatical category	Usage	Occurrences
1853. Domínguez, s.v. *cabalito*. (idem Domínguez 1869)	Ironical reply to express negation, mockery, confirmation. Similar uses are also reported	Diminutive adjective for *cabal*	Unmarked	3 use patterns (examples under 21)
1917. Alemany, s.v. *cabalito*	Cabalmente	Diminutive for *cabal*. Masculine adverb	Informal	No examples
1936. RAE, s.v. *cabalito*	Cabal. Cabalmente	Masculine adverb	Informal	4 use patterns (examples under 22)
2023. DEA, s.v. *cabal*. (idem DEA 1999 and DEA 2007).	Exactamente. Dicho para asentir a lo que acaba de oírse 'expressing agreement'	Adverb	Colloquial	1 use pattern (example under 23)

(21) 1853 Domínguez (idem in Domínguez 1869). [Para la negación] "*Me hará vd. gusto*". "*¡Cabalito!*"; [para expresar burla] "*Cabalito, que me gusta mucho*"; [para expresar la decisión] "¿Irás a verle?". "*Cabalito*".

(22) 1936. RAE (*Diccionario histórico*) "Empiece usted por su casa / a corregir el exceso" / "¿Por mi casa?". *"Cabalito*" (Ramón de la Cruz, 1731–1824); Que sois mi suegro, / *cabalito*, en dos palabras (Leandro Fernández de Moratín, 1760–1828); "¿Le amáis por fe?". "*Cabalito*" (Juan Eugenio Hartzenbusch, 1806–1880); *Cabalito*. Eso quiero, que gastes de lo tuyo (Jacinto Octavio Picón, 1852–1923. *La honrada*, ed. 1924).

(23) 2023. DEA (ídem en DEA 1999 y en DEA 2007). "Esa es la fuente de vino que dicen ¿no?", le preguntaron. "*Cabalito*" (Ángel María de Lera. 1912 Baides-Castilla la Mancha-1984, Madrid; *La boda* [1959], in *Novelas*. 1966).

As can be seen from the above, *cabalito* is recorded in few dictionaries, current or past. Two major points from Table 6 can be underlined: i) the RAE has not recorded *cabalito* in its official dictionary of Spanish in three centuries, but it did in the first diachronic dictionary (RAE 1936); and ii) the earliest record is by non-academic lexicography (Domínguez in 1853), and it also records this form in present-day (DEA 2023 is the only 21st c. dictionary to cite this diminutive in current Spanish use). Table 6 is inspected in further detail below.

Cabalito is recorded only in four dictionaries during the period 1853–2023: i) the two editions of *Diccionario nacional* by Domínguez (1853, 1869); ii) Alemany's (1917) dictionary; iii) the RAE's first historical dictionary (RAE 1936); and iv) the three editions of the dictionary by Manuel Seco, Olimpia Andrés and Gabino Ramos (DEA 1999, DEA 2007, DEA 2023). In all cases, the same information is given across editions of the same dictionary. Column 1 shows that this diminutive is not recorded the same in all dictionaries. Domínguez was first to record *cabalito* and also to list it as a separate entry in his 1853 dictionary. Alemany (1917) and the RAE (1936) followed this convention in the 20th c. The relevance of *cabalito* in the 1853 and 1917 general dictionaries may stem from the relevance of the term at the time, which is also the period of highest frequency in the base corpus. At the turn of the 21st c., the DEA differs from the above and, following more orthodox criteria, lists *cabalito* as a variant form of *cabal*.

Table 6 evidences disagreement in the glosses. While Alemany (1917) and the RAE (1936) merely point out the semantic equivalence of *cabalito* with *cabalmente* and *cabal* 'upright' within the same word family, Domínguez (1853 and 1869) and the DEA (1999, 2007 and 2023) describe its use as a colloquial form. Notably, only the earliest and latest lexicographic records highlight the pragmatic function of

cabalito in conversation: Domínguez describes it as an ironical reply to express negation, mockery, confirmation, and the like ("estribillo irónico que manifiesta: una negación, la burla, la decisión; tiene otros usos del mismo tenor"), and the DEA dictionary lists *cabalito* (as a variant of *cabal* 'upright') to express agreement ("para asentir a lo que acaba de oírse") meaning 'precisely' ("exactamente"). The semantic range is wider according to Domínguez than to the DEA dictionary, maybe for the more frequent use of *cabalito* in the mid-19th c. than at present, as shown by the newspaper corpus used here.

As for grammatical status, *cabalito* is first described as an adverb in the 20th c. dictionaries. Domínguez just mentions the morphological form as the diminutive adjective from *cabal* ("adjetivo diminutivo de cabal"), and Alemany (1917) and the RAE dictionary (1936) refer to it as a masculine adverb ("adverbio masculino"). As mentioned above, the DEA (1999–2023) equates *cabalito* and *cabal* when the latter occurs as an adverb.

Table 6 also shows that the 20th c. dictionaries mark the term *cabalito* as proper of *informal* use (Alemany 1917, RAE 1936) and later as *colloquial* (DEA 1999, 2007, and 2023). The *Diccionario nacional* (1853) does not add an explicit categorization, but Domínguez links up *cabalito* with colloquial register and, thus, describes it as a deeply ironic feature of spoken language. In either case, dictionaries explicitly or implicitly associate *cabalito* with the most informal registers. Colloquial forms more proper of spoken than of written language are typically not recorded in dictionaries, even if they have become colloquial pet phrases, as Domínguez suggests for *cabalito*. Actually, as mentioned above, this diminutive, first attested in 1760–1770 and widely used in the 19th c. press according to corpus evidence, was never recorded by the RAE's general dictionary.

Examples 21 through 23 illustrate *cabalito* respectively as in Domínguez (1853), the first historical dictionary (RAE 1936), and the DEA 2023. Except for Domínguez (1853), who inserted his own examples in the gloss, all the examples are from literary texts: The RAE dictionary (1936) gives examples by four 18th and 19th century authors (Ramón de la Cruz, Moratín, Hartzenbusch, and Picón), and the DEA 2023 cites *La Boda*, a 1966[5] novel written by Lera. Whether fiction or non-fiction, all the examples are direct speech, and thus underline this diminutive's colloquial nature.

Finally, no dictionary record gives diatopic information of use, whether inter- or intra-dialectal. The list of literary authors cited may be used, though, as indirect

5 This is the only attestation in the three editions of the DEA: A 1966 occurrence in literary language. The wider chronological coverage of the third edition (1950–2023, according to the foreword) compared with the first two (European Spanish of the second half of the 20th c.) does not add any attestation of *cabalito*.

evidence of diatopic information, if their provenance is researched (see Table 7): Thus, all the examples cited in the historical dictionary (RAE 1936) are by four authors from Madrid of the 18th and 19th century, one of whom is also listed in the CDH corpus (Jacinto Octavio Picón). Remarkably, the 17 European occurrences of *cabalito* in the RAE are by five authors born in Madrid, and one in Cantabria. Overall, the entire list of examples is by ten authors, nine of whom are from Madrid. Even the 1771 anonymous record is referred to Madrid. As a result, *cabalito* may have been historically associated with the Court of the 18th to the 20th centuries. However, this dialectal bias is not supported by HD evidence, as shown in the previous section. Diatopically, *cabalito* arose in Madrid in the second half of the 18th c., spread over European Spanish in the 19th c., and seems to have retreated to Castilla-La Mancha in the 21st c. This is also the origin of the 1966 example of *cabalito* by the novelist cited in the DEA (1999, 2007 and 2023).

The RAE data may suggest that *cabalito* arose and disappeared as a form specific of the Spanish spoken in Madrid, even if the evidence of newspaper data shows that it reached the dialectal status of a typical European form in the Spanish-speaking countries.

Table 7: The diachronic use of *cabalito* according to RAE 1936 and CDH. Birthplace of Spanish authors.

Author /Title	Date	Country	Author's provenance	No. of occurrences
Anonymous	1771	Spain	Madrid	1 (CDH)
Ramón de la Cruz (Madrid 1731-Madrid 1824), *Obras*	18th c.	Spain	Madrid	1 (RAE 1936)
Leandro Fernández de Moratín (Madrid 1760-París 1828) *Obras*	18th c	Spain	Madrid	1 (RAE 1936)
Mariano José de Larra (Madrid 1809-Madrid 1837) *Traducción de Roberto Dillón*	1832	Spain	Madrid	1 (CDH)
Juan Eugenio Harztzenbusch (Madrid 1806-Madrid 1880), *Obras*	19th c.	Spain	Madrid	1 (RAE 1936)
Gorostiza *Contigo pan y cebolla. . .*	1833	Mexico	Madrid	2 (CDH)
Ayguals de Izco *La Bruja de Madrid. . .*	1850	Spain	Madrid	7 (CDH)

Table 7 (continued)

Author /Title	Date	Country	Author's provenance	No. of occurrences
Ascasubi *Paulino Lucero* (1852) *Aniceto el Gallo* (1872)	1852, 1872	Argentina		1 (CDH) 3 (CDH)
Gaspar Enrique (Madrid 1852-Francia 1902) *La Chismosa*	1868	Spain	Madrid	1 (CDH)
José María de Pereda (Polanco 1833, Santander 1906) *Tipos y Paisajes*	1871	Spain	Cantabria	2 (CDH)
Ricardo Palma *Tradiciones peruanas*, 3ª serie, (1875) *Tradiciones peruanas* 4ª serie, (1877)	1875 1877	Peru		1 (CDH) 1 (CDH)
Jacinto Octavio Picón (Madrid 1853-Madrid 1923) *La honrada*	1890	Spain	Madrid	1 (RAE 1936) ídem (CDH)
Eduardo Blanco *Tradiciones épicas y cuentos viejos*	1912	Venezuela		1 (CDH)
Antonio Díaz-Cañabate (Madrid 1987-Madrid 1980) *Paseíllo por el planeta de los toros, crónicas taurinas*	1970	Spain	Madrid	4 (CDH)
Vargas Llosa *La tía Julia . . .*	1977	Peru		1 (CDH)

6 Conclusions

This paper underlines the need for a combined use of diachronic corpora (linguistically annotated, big data) and digital newspaper libraries (linguistically non-annotated, big data) in the pan-Hispanic research on certain issues of the evolution of colloquial Spanish, e.g. the adverb *cabalito* for the low frequency in diachronic reference corpora. This paper relies on HD as the base corpus and on CDH as a control corpus, and discloses to a large extent the evolution of this adverb on ac-

count of its diachrony evidence, its occurrence across text types, its diatopic evidence, and its grammatical status.

In the former respect, *cabalito* is an innovative form of the mid-18th c. It became widespread and reached full standardization in the late 19th c. In the 20th and 21st centuries it fell out of use. These radical changes can be found only in the base corpus, for two reasons: i) *cabalito* is attested for over two centuries (1790–2022) only in the base corpus, whereas it is attested in CDH only continually with major chronological gaps; and ii) the number of occurrences in the base corpus is 12 times as high as in the CDH corpus (317 vs. 27 occurrences, respectively).

Regarding text types, the development of *cabalito* was heavily influenced by the journalistic genre. The term is attested more frequently in rushed writing and in conditions of communicative immediacy, and is much more open to informal language: *Variety news* (local events, gossip) and *serial publications* (literary and non-literary). In these subtypes, *cabalito* is always recorded in direct speech and starts a turn. This is evidence of this diminutive's association with the spoken mode.

Pan-Hispanically, *cabalito* is specific of European Spanish. Interdialectal distribution shows that the term was attested in Spain but not in America for over two centuries. Interdialectally, it started in Madrid and spread from the geolect of the *Meseta* to the rest of diatopic varieties.

Morphosyntactically, the adverb *cabalito* expresses agreement with a previous statement, or disagreement, if used ironically. Under this morphosyntactic and pragmatic specialization, it is highly frequent, even if it is rarely recorded in Spanish dictionaries.

Appendix

Spanish Historical Press, with Attestations of Cabalito *by Region*

Andalusia. Journalistic examples of *cabalito* (1809–1944)

Place	Newspaper. Date, Page
Almería	*El Radical.* 16/07/1906, p. 4
Almería	*Yugo.* 09/07/1944, p. 7
Almería/Vélez Blanco	*La Opinión.* 14/11/1895, p. 1
Almería/Vélez Rubio	*El Loro*. *01/12/1913, pág. 1*
Cádiz	*El Duende de los cafées*. 08/08/1813, p. 6

(continued)

Place	Newspaper. Date, Page
Cádiz	*El correo de Cádiz.* 04/08/1920, p. 1
Cádiz/Jerez	*El Progreso*. 21/04/1870, p. 2
Cádiz/Jerez	*El Guadalete.* 12/08/1920, p. 1
Córdoba	*El amigo católico*. 17/06/1875
Córdoba	*La voz*. 17/05/1924, p. 16
Córdoba/Pozoblanco	*El Cronista del Valle*. 28/05/1960, p. 2
Granada	*El Defensor de Granada.* 03/04/1890, p. 2
Granada	*El Defensor de Granada.* 23/12/1924, p. 2
Málaga	*Atalaya patriótico de Málaga.* 01/04/1809, p. 9
Málaga	*El Folletín.* 28/02/1875, p. 3
Sevilla	*La Tía Norica*. 1814, nº 20, p. 4
Sevilla	*ABC-Sevilla.* 22/04/1942, p. 9

Aragon. Journalistic examples of *cabalito* (1798–1927)

Place	Newspaper. Date, Page
Teruel	*Guía del magisterio.* 30/01/1881, p. 5
Teruel	*El Mercantil.* 09/01/1914, p. 3
Zaragoza	*Semanario de Zaragoza.* 13/08/1798, p. 3
Zaragoza	*Heraldo de Aragón.* 14/01/1927, p. 8

Asturias. Journalistic examples of *cabalito* (1886–1926)

Place	Newspaper. Date, Page
Gijón	*Gijón cómico.* 14/09/1889, n. p.
Gijón	*Páginas escolares.* 01/11/1926, p. 2
Oviedo	*La cruz de la victoria.* 25/11/1886, p. 2

Balearic Islands. Journalistic examples of *cabalito* (1871–1914)

Place	Newspaper. Date, Page
Mallorca	*El genio de la libertad.* 24/11/1839, p. 1
Mallorca	*El juez de paz.* 21/09/1871, p. 4
Mahón	*El bien público.* 8/03/1913, p. 2
Mahón	*La Alquitara.* 21/03/1914, p. 5

Canary Islands. Journalistic examples of *cabalito* (1908–1911)

Place	Newspaper. Date, Page
Tenerife/Sta. Cruz	*La gaceta de Tenerife.* 16/07/1910, p. 3
Tenerife/Sta. Cruz	*El Tiempo.* 05/12/1908, p. 2
Tenerife/Sta. Cruz	*Gaceta de Tenerife.* 19/08/1910, p. 4
Tenerife/Sta. Cruz	*El Progreso.* 22/02/1911, p. 3

Cantabria. Journalistic examples of *cabalito* (1844–1922)

Place	Newspaper. Date, Page
Santander	*La Verdad, diario de la mañana*. 08/01/1844, p. 2
Santander	*La Abeja montañesa: periódico de intereses locales*. 02/11/1865, p. 3
Santander	*La Abeja montañesa*. 1865, p. 3
Santander	*El Atlántico*. 28/08/1889, p. 2
Santander	*El Avisador*. 14/03/1897, p.2
Santander	*La Atalaya*. 27/03/1905, p. 3
Santander	*El pueblo cántabro.* 22/07/1922, p. 8

Castile and León. Journalistic examples of *cabalito* (1882–1932)

Place	Newspaper. Date, Page
Burgos	*El Papa-Moscas: periódico satírico*. 04/06/1882, p. 2
Burgos	*El Papa-Moscas: periódico satírico*. 22/02/1885, p. 1
León	*Diario de León*. 05/05/1908, p. 3
León	*Diario de León.* 02/07/1932, p. 4
Salamanca	*La Legalidad: revista de asuntos administrativos.* 30/01/1982, p. 3
Salamanca	*El Salmantino.* 05/02/1910, p. 2
Salamanca/Béjar	*La Victoria: semanario de Béjar*. 26/07/1902, p. 1
Salamanca/Béjar	*La Victoria: semanario de Béjar*. 24/02/1906, p. 2
Soria/Burgo de Osma	*La Propaganda.* 14/11/1884, p. 2
Soria/Burgo de Osma	*La Propaganda. 0*5/08/1891, p. 3
Soria	*El noticiero de Soria*. 03/03/1894, p. 2
Soria	*Ideal numantino*. 21/04/1911 p. 2

Castile-La Mancha. Journalistic examples of *cabalito* (1898–1930)

Place	Newspaper. Date, Page
Ciudad Real	*El Pueblo manchego*. 21/03/1911, p. 2
Cuenca	*El Catequista.* 31/01/1907, p. 3
Toledo	*La Aurora.* 18/10/1898, p. 4
Toledo	*El Heraldo toledano.* 29/06/1930 p. 10

Catalonia. Journalistic examples of *cabalito* (1881–1915)

Place	Newspaper. Date, Page
Barcelona	*El Mundo ilustrado. Biblioteca de las familias*. 1881–1883. Nº 152, p. 26
Barcelona	*Iris*. 18/4/1903. p. 9
Gerona	*La nueva lucha: diario de Gerona.* 190/02/1888, p. 1
Gerona/Olot	*El Eco de la montaña.* 16/05/1897, p. 2
Tarragona	*Diario del comercio.* 06/08/1898, p. 3
Tarragona	*La Cruz.* 10/03/1915, p. 1
Tarragona/Tortosa	*El estandarte católico.* 11/07/1899, p. 2
Tarragona/Tortosa	*El Restaurador.* 30/10/1908, pág.3

Extremadura. Journalistic examples of *cabalito* (1885–1908)

Place	Newspaper. Date, Page
Badajoz	*El Avisador de Badajoz*. 19/03/1885, p. 3
Badajoz	*Noticiero extremeño.* 17/06/1906, p. 3
Cáceres	*Revista de Extremadura. Ciencia y arte.* 01/09/1903, p. 42
Cáceres	*El bloque.* 14/06/1907, p. 3

Galicia. Journalistic examples of *cabalito* (1889–1927)

Place	Newspaper. Date, Page
Lugo	*El Lucense.* 20/04/1889, p. 1
Lugo	*El norte de Galicia.* 02/07/1906, p. 3
Lugo	*Acción social.* 15/02/1921, p. 8
Lugo	*El Progreso: diario liberal.* 08/02/1927, p. 3

La Rioja and the Basque Country. Journalistic examples of *cabalito* (1893–1933)

Place	Newspaper. Date, Page
Logroño	*La Rioja: diario político*. 06/09/1893, p.2
Logroño	*La Rioja: diario político.* 28/03/1933, p. 4
Álava	*Heraldo Alavés: Diario independiente*. 15/04/1902, p. 1

Murcia. Journalistic examples of *cabalito* (1900–1903)

Place	Newspaper. Date, Page
Murcia	*La Juventud literario.* 10/04/1900, p. 3
Murcia	*Heraldo de Murcia.* 10/4/1902, p. 3
Murcia	*El Liberal.* 31/01/1903, p. 31

Valencia. Journalistic examples of *cabalito* (1881–1910)

Place	Newspaper. Date, Page
Alicante	*El Graduador.* 18-08-1898, p. 1
Alicante	*La Voz de Alicante.* 28/04/1906, p. 1
Alicante/Orihuela	*La lectura popular*. 15/09/1884, p. 2
Alicante/Orihuela	*El nuevo alicantino.* 10-09-1897, p. 2
Alicante/Alcoy	*Heraldo de Alcoy.* 04/09/1902, p. 4
Alicante/Gandía	*Revista de Gandía.* 30/05/1908, p. 1
Valencia	*Periódico monárquico.* 21/12/1881, pág. 2
Valencia	*El Pueblo: diario republicano de Valencia.* 09/08/1910, p. 2

Bibliography

Alemany y Bolufert, José (1917): *Diccionario de la lengua española*. Barcelona: Ramón Sopena. Available in NTLLE.

Calderón Campos, Miguel (2024): "Spanish Corpora: Big (Quality) Data?", in Ana Gallego Cuiñas and Daniel Torres-Salinas (eds.). *Humanities and Big Data in Ibero-America. Methodological Issues and Practical Applications*. Berlin/Boston: De Gruyter, pp. 109–127.

Campos Souto, Mar (2018): "Bibliotecas y hemerotecas digitales en el NDHE", in *Cuadernos del Instituto Historia de la Lengua*, 11, pp. 237–255.

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

Company Company, Concepción and Rodrigo Flores Dávila (2017): "Género textual, diacronía y valoración de un cambio sintáctico", in BRAE XCVII-CCCXV, pp. 203–239.

Company Company, Concepción and Rodrigo Flores Dávila (2018): "El contraste *a por* vs. *por* con verbos de movimiento", in *Revista de Filología Española*, 98(2), pp. 281–318.

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/>.

CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <https://www.rae.es/banco-de-datos/corpes-xxi>.

CREA = Real Academia Española: *Corpus de referencia del español actual*. <https://www.rae.es/banco-de-datos/crea>.

DLE = Real Academia Española (2023): *Diccionario de la Lengua Española*. <https://dle.rae.es/>.
Domínguez, Ramón Joaquín (1846–1847/^{2}1853): *Diccionario nacional o gran diccionario clásico de la lengua española*. Madrid-Paris. Available in NTLLE.
García-Godoy, María Teresa (2015): "Political and Lexical Emancipation in Spanish America. The Nineteenth Century in the History of Americanisms", in *Nineteenth-Century Context*, 37(4), pp. 321–339.
García-Godoy, María Teresa (2017): "La diferenciación léxica del español de América. Anglicismos jurídicos e institucionales en la Colonia tardía", in *Hispania*, 100(1), pp. 65–78.
García-Godoy, María Teresa (2021): "De *madamas* y *madamitas*. Un tratamiento galicado en la historia del español moderno", in *Rilce*, 37(1), pp. 46–72.
Gerhalter, Katharina (2020): *Paradigmas y polifuncionalidad. Estudio diacrónico de preciso/precisamente, justo/justamente, exacto/exactamente y cabal/cabalmente*. Berlin/Boston: De Gruyter.
HD = Biblioteca Nacional de España: *Hemeroteca Digital* <https://hemerotecadigital.bne.es/hd/es/advanced>.
NGLE = Real Academia Española y Asociación de Academias de la Lengua Española (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.
NTLLE = Real Academia Española: *Nuevo Tesoro Lexicográfico de la Lengua Española*. <https://apps.rae.es/ntlle/SrvltGUILoginNtlle>.
Octavio de Toledo y Huerta, Álvaro S. (2016): "Sin CORDE pero con red: algotras fuentes de datos", in *Revista Internacional de Lingüística Iberoamericana* (RILI) 28, pp. 19–48.
Seco, Manuel, Olimpia Andrés and Gabino Ramos (1999/2007): *Diccionario del español actual*. Madrid: Espasa Calpe. 2 vols.