Miguel Calderón Campos

Spanish Corpora: Big (Quality) Data?

1 Introduction

In Linguistics, reference to Big Data entails the reference to corpora and to their ensuing size, type, representativeness and sample selection. Figure 1 shows the tendency towards bigger and bigger Spanish corpora, from the early RAE projects of over 100-million words (CREA) to the macro-corpora of project *TenTenCorpora* aiming at over 10,000 million words. In the latter, the Spanish corpus, *EsTenTen18*, is close to 17,000 million words.

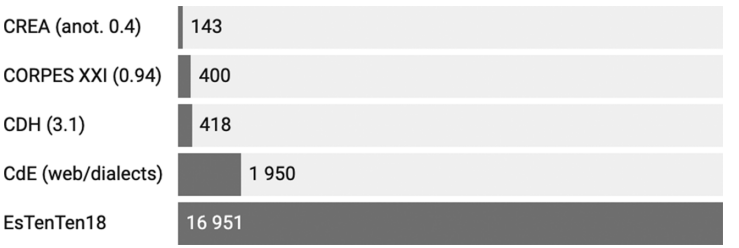

Figure 1: Spanish corpus size in millions of tokens.

The sizes shown in Figure 1 might give the impression that these resources are already beyond the minimum necessary for the exhaustive description of any question in Linguistics. Yet, the endless universe of the web and of social networks is still searched for new data, as if the big size of corpora were not enough for the description of some words' constructional or diachronic, stylistic or social variation profile. Equally paradoxically, small corpora are built more and more frequently to fill the gaps left by bigger, general corpora.

Computational Linguistics thus currently works on three fronts: the compilation of macro-corpora reference corpora, the annotation of highly specific small

Note: This work is part of the ALEA XVIII Project, funded by FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades/Reference Project P18.FR.695. It is also part of the ALEA XVIII. Oriental, financed by FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades, reference project A-HUM-116-UGR20.

https://doi.org/10.1515/9783110753523-008

corpora, and the improvement of traditional Corpus Linguistics by means of the analysis of massive internet and social network data.

This paper is an overview of Spanish Corpus Linguistics. Section 2 reviews the synchronic and diachronic corpora available and points out limitations imposed by source quality and by the interfaces used (in general, RAE corpora offer better data selection and achieve a higher representativeness, whereas non-RAE corpora use more flexible and powerful search engines, as shown in section 2.1). Based on the analysis of the Colombian Spanish token *parce*, section 2.2 shows that inaccurate search results are closely related to low quality samples and geographic metadata. Section 3 uses massive corpora, internet data and social network data for improved results on the little evidence of the quantifier *algotro* ('some other') available in RAE corpora. Finally, section 4 compares Big Data sources with two specific diachronic corpora: *Post Scriptum* (P.S.) and *Oralia diacrónica del español* (ODE).

2 Spanish Reference Corpora and Massive Corpora

General corpora or reference corpora are corpora intended for the attestation of general properties of a language at a given period of its history. For Spanish, a general or reference corpus must contain all types of texts, of all the periods into which the timeframe intended for research can be divided, and from all the countries where Spanish is spoken as a first language.

The *Corpus del Español del siglo XXI* (CORPES XXI) and the *Corpus del Español* (CdE web/dialects) are the two commonly acknowledged reference corpora of contemporary Spanish. The *Corpus del Diccionario Histórico de la Lengua Española* (CDH) and the historical subcorpus of the *CdE* (CdE hist) are diachronic Spanish reference corpora. The basic properties of all four corpora are outlined in Table 1 below.

The latest versions of the two RAE corpora, CORPES XXI and CDH, amount to ca. 400 million words. The former contains samples produced since 2001 and is intended to increase by ca. 25 million words per year. Transcripts of spoken samples amount to 1%, some linked to audio files. The intended variety proportion is ca. 30% European Spanish and 70% American Spanish.

The CDH includes the samples of the first RAE corpora, CREA and CORDE, after descriptive annotation (lemmatization and morphosyntactic labelling), similarly to CORPES XXI. Unlike the four major types of samples in CORPES XXI (fiction, non-fiction, press, spoken), the samples of the CDH are classified by topic

Table 1: The Spanish reference corpora.

	Tokens (by million)	Spain	America	Period	Fiction	Non-fiction	Press	Spoken
CORPES XXI (0.94)	400	35%	65%	2001–21	28%	21%	47%	1%
CDH (3.1)	418	71%	29%	12th c.-2005	classified by topic			
CdE (web/dialects)	1950	22%	78%	2013–14	Blog (53%) / General (47%)			
CdE (hist)	100	Data not available		13th c.-20th c.	20th c.: 25%	25%	25%	25%

(i.e. arts, social sciences, science and technology, leisure and everyday life, politics and economy, and health).

The CDH can be divided into three subcorpora, each of which can be accessed separately: i) the CDH core subcorpus (*CDH nuclear*) is a 63-million-word representative collection of samples taken from the CORDE and CREA; ii) the CDH XII-1975 subcorpus is a 230-million-word collection of most of the contents of the old CORDE corpus; and iii) the CDH 1975–2000 subcorpus is a 125-million-word collection of the CREA contents not included in the CDH core subcorpus. The proportion of European vs. American Spanish for the period from 1492 onwards in the CDH is 71% vs. 29% respectively.

The *CdE web/dialects* corpus is a reference macro-corpus (nearly 2,000 million words) of web samples of the period 2013 and 2014. It is arranged as two large sets (blogs vs. general) and is representative of the 21 Spanish-speaking countries.[1] The CdE's historical subcorpus contains samples from the 13th c. to the 20th c. Query results can be sorted by century and, for the samples of the 20th c., also by sample type (note that the 20 million words of the 20th c. are evenly distributed over the four sample types shown in Table 1).

At 16,951 million words, *EsTenTen18* is the biggest among the so-called massive corpora of Spanish. The samples were extracted automatically from internet sources and can be searched with Sketch Engine. Structured by subdomains (European Spanish domain.es, Mexican domain.mx, Chilean domain.cl, etc.), it allows to combine searches by descriptive and geographic data (see section 2.2).

1 21 countries including the United States, 22 including Equatorial Guinea.

2.1 Sample Quality vs. Interface Versatility

The main difference between the above corpora runs along the lines of Mair's (2006) contrast between 'big and messy' corpora vs. 'small and tidy corpora': the bigger the corpus, the lower the quality, the representativeness, and the accuracy of sample classification and annotation (both descriptive and presentational); by contrast, smaller corpora lend themselves to manual annotation and, therefore, achieve comparatively better sample selection and higher annotation accuracy.

RAE corpora are annotated and lemmatized remarkably accurately. Also, their samples are selected according to representativeness and are annotated with more accurate geographic, chronological and thematic metadata than non-RAE corpora (Rojo 2010). By contrast, non-RAE corpora rely on a more flexible and powerful search interface than RAE corpora, and count on bigger sizes: compared with CORPES XXI, CdE web/dialects is five times as big, and *EsTenTen18* is nearly fifty times as big.

While the quality of CORPES XXI's samples is praised on the CdE's website, it is also stated that '[. . .] it uses a fairly rudimentary web interface, which really limits what can be done with concordances, collocates, and frequency lists. In other words, the good textual data is "trapped" behind a poor interface, and is inaccessible to end users'.[2]

EsTenTen18 is praised for its size, for the collocate-based 'word sketches' and for the possibility to submit queries with CQP. By contrast, it is criticised for the poor lemmatization and for the amount of wrong or inaccurate annotation. Indeed, *EsTenTen18* becomes unbeatable for its size and for its powerful, user-friendly interface, when it comes to finding the combination profile (word sketch) of highly frequent words. Graphical representations of a given token's profile are easily generated, as in the adjective *severo* 'severe' of Figure 2. Remarkably, the same figure exposes one of the main shortcomings of this type of macro-corpora too, namely their poor morphosyntactic annotation: funny enough, the most frequent collocate for the adjective *severo* is Spanish Nobel prize winner's surname *Ochoa* (thus, *Severo Ochoa*).[3]

CdE web/dialects stands out for the possibility to research dialectal differences across the 21 Spanish-speaking countries. Thus, a single query for the adjectival suffix *-oso* returns Argentinian Spanish adjectives like *ochentoso* 'eighty-like', *noventoso* 'ninety-like', *criterioso* 'sensible', *modernoso* 'modern' or *culposo* 'guilty' vs. Euro-

2 https://www.corpusdelespanol.org/compare.asp (17–12-2021).

3 CdH yields the same wrong annotation. Wrong annotation can be revised only manually.

Figure 2: A graphical representation of the collocates of the adjective *severo* 'severe' in *EsTenTen18*.

pean Spanish adjectives like *lioso* 'messy', *cantoso* 'flagrant', *picajoso* 'fussy', *pasteloso* 'soppy' or *patoso* 'clumsy'.

The option *Chart* allows to obtain a very telling overview of well-attested general usage. Thus, the query "re _J*"[4] yields a chart comparing the normalized frequency of "*re*+adjective" in all the Spanish-speaking countries, and significant contrasts can be noticed: the highest frequencies occur in the varieties of Argentina (17.94 per million words), Chile (8.06 wpm) and Paraguay (5.58 wpm). Frequencies below 3.20 wpm (Mexico) are attested in the remaining varieties. The adverbial counterpart with *re-* (e.g. *rebién* 'very well', *remal* 'very bad', *retarde* 'very late', etc.) shows a similar distribution across varieties: Argentina attests 3.10 wpm, Chile 1.56 wpm and Uruguay 1.26 wpm. Guatemala attests a similar result as the south American countries: 1.07 wpm.

RAE corpora do not rely on search engines capable of rendering visual results as in Figure 2. CORPES XXI and CDH present quantitative results as absolute and relative frequencies by country. Surprisingly, the pie charts generated automatically only give results of absolute frequencies, and this may severely distort the picture. For example, the well-known American Spanish preference for *computadora* 'computer' vs. European Spanish *ordenador* 'computer', is confirmed by the

4 I.e. *re-* prefixed to an adjective for intensification, e.g. *rebueno* 'very good', *relindo* 'very nice', *reloco* 'very crazy', etc., NGLE 10.9j.

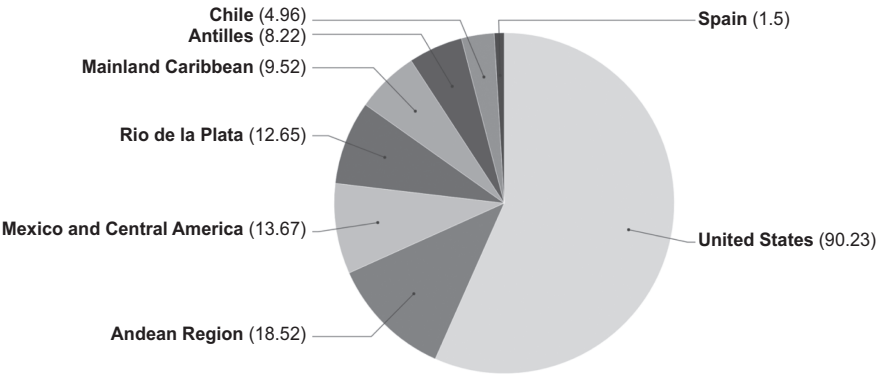

Figure 3: The frequency of *computadora* 'computer' in the CDH (wpm values).

CDH data: at 1.5 wpm, the relative frequency of *computadora* in European Spanish ranks lowest among the Spanish-speaking countries (cf. Figure 3, generated by the author, based on the CDH's wpm frequencies for this query).

Contrarily, based on absolute frequencies, the CDH's graphical representation (see Figure 4),[5] stands in sharp contrast with Figure 3 above, and is therefore misleading: as European Spanish amounts to 71% of the samples in the CDH, the absolute frequency of *computadora* for European Spanish (402 occurrences) is the highest in the corpus. This is a serious weakness of the concordancer's data management, and also one that could be easily overcome by linking pie chart generation to wpm frequencies instead of to absolute values.

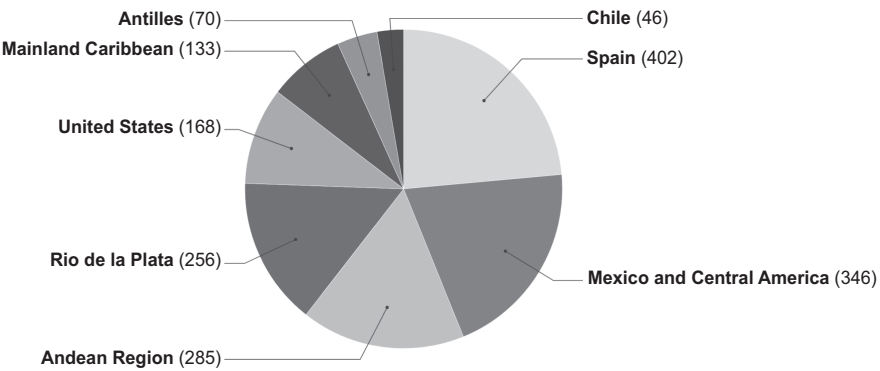

Figure 4: The frequency of *computadora* 'computer' in the CDH (absolute values).

5 This figure, generated by the author, is a copy of the figure generated by the CDH.

The lower flexibility of the interface is compensated in RAE corpora with a better sample selection and with more accurate descriptive (linguistic) and non-descriptive (presentational) annotation. CORPES XXI and CDH therefore allow a much more precise chronological and geographic identification of language data. Take, for example, the adjective *severo* again, originally used in Spanish to mean 'severe' (as in *castigo severo* 'severe punishment' or *crítica severa* 'severe criticism', etc.). A later development under the influence of English extended the semantic range to mean severity of illness (*depresión severa* 'severe depression', *discapacidad severa* 'severe handicap', *traumatismo severo* 'severe injury' etc.). Neither the CdE corpus nor the *EsTenTen18* corpus identify the earliest record of *depresión severa*, which CDH attests earliest in a sample of Venezuelan Spanish of 1976 (*sentimientos de culpabilidad y lo suficientemente severos* 'guilt feelings severe enough'). Something similar applies to the collocation *traumatismo severo*, first attested in Argentinian Spanish in 1988.

2.2 Precision and Recall. Corpus Evidence on Colombian Spanish *Parce*

'Precision and recall' are defined by Stefanowitsch (2020: 111–116) according to sample quality and annotation accuracy. Data retrieval is accurate whenever a query returns only exact matches. Thus, research on imperative verb forms ending in *-lde* (*dezilde* 'you tell him', *dalde* 'you give him', *enbialde* 'you send (to) him', etc.) in a non-annotated corpus of the 16th c. will retrieve both imperative forms and false positives, the latter as a result of the retrieval of nouns with the same ending, e.g. *alcalde* 'mayor', *balde* 'bucket' or *molde* 'cast'.

Exhaustive data retrieval ('recall') is achieved whenever every possible match is retrieved. This is especially difficult to attain in historical research, for the many orthographic variants that a token may display. Thus, the following variants are attested for the token *trébedes* 'trivet' in the ODE, some of which are quite unpredictable: *trevedes*, *trebedes*, *treuedes*, *treodes*, *trévedes*, *trebes*, *estrebes*, *esttreores*, *extrevedes*, *estrebedes*. These forms cannot be retrieved under the same query and are thus a major source of data loss during data retrieval (as 'false negatives').

This section assesses the degree of precision and recall of CORPES XXI, CdE and *esTenTen18* according to their samples and linguistic annotation. The source of the samples of RAE corpora is mainly publications, including revised editions. This reduces to a minimum the amount of typographical mistakes and inconsistencies, in contrast with corpora built with samples collected from blogs and non-institutional websites. This can be illustrated with the Colombian addressing term *parce* 'friend, pal'. A shortened form for *parcero*, it is used among the younger

speakers as an addressing term to express comradeship or conviviality. The term comes from Portuguese *parceiro* 'friend, pal'. It was allegedly used first in the lower class quarters of Medellín in the 1980s, and then spread over the rest of the country (Castañeda Naranjo 2005: 67).[6]

CORPES XXI contains enough evidence to describe the usage or the geographic distribution of *parce*: out of the 49 instances recorded in the corpus, only 8 are not from Colombian Spanish (4 are typographical mistakes[7] or foreign words[8] used in European Spanish samples; the other 4 are the vocative form used by Colombian Spanish speakers in literary works or journal articles).[9] The evidence available in this RAE corpus thus confirms that *parce* is associated with Colombian Spanish, and illustrates not just its combinatory possibilities and its meaning (1), but also its origin (1) and its chronological development (2):

(1) Parce!!! (de *parcero*, que en Colombia es amigo) Hermano!!! («Miguel Bosé se ofreció a mediar con las FARC al recibir nacionalidad colombiana». El Comercio. pe. Lima: elcomercio.pe, 2010-03-17).

(2) Es 1994, todavía son pocos los que dicen *parce* (Castro, Samuel: *A la velocidad del byte*. Medellín: Fondo Editorial Universidad EAFIT, 2008).

Parce is nearly always used as a vocative, before or after a pause (3–6). It is also recorded as a noun meaning 'friend, pal' ("se trataba de un parce de ellos"). It is often used with the pronoun *usted* ('you [formal]'), except for one example with *vos* ('you [informal]' 6) and another with *sumercé* 'you [formal]' (5).

(3) —Parces, ¿alguno de ustedes tiene algo para la cabeza? (Martínez, Fabio: «Los ensayistas del Parque del Perro». *El escritor y la bailarina*. Cali: Escuela de Estudios Literarios de la Universidad del Valle, 2012).

6 The earliest attestation in the CDH dates back to 1994: "Un ejemplo: ¿Entonces qué, parce, vientos o maletas? ¿Qué dijo? Dijo: Hola hijo de puta. Es un saludo de rufianes" (Vallejo, Fernando, La virgen de los sicarios [Colombia] [Santafé de Bogotá, Alfaguara, 1999).

7 *Parce* for *parece*: "parce que van dejando . . .".

8 The Latin formula "Parce nobis, Domine", or the French causal conjunction "parce que" 'because': "Hay una frase recurrente durante la película: parce que moi je rêve . . .".

9 The Mexican example is by a Colombian character in a play ("Cuántos años tenemos de parces, de amigos"). Two Ecuadorian examples are a news article about Colombian hit men ("acá lo cogemos, parce, y le damos paila"). The Bolivian example refers to Colombian singer Juanes' album *P.A.R.C.E.*

(4) – Si su mujer le puso los cuernos, parce, yo no tengo la culpa, la culpa la tiene usted (López, Andrés; Ferrand, Juan Camilo: *Las muñecas de los narcos*. Madrid: Aguilar, 2010).

(5) -Hum, parce, sumercé anda desactualizado (Álvarez, Juan: *C.M. no récord*. Bogotá: Alfaguara, 2011).

(6) —¿Querés, parce? (Franco, Jorge: *El cielo a tiros*. Bogotá: Penguin Random House Grupo Editorial, 2019).

The dialectal distribution of *parce* according to the chart based on CdE data runs against the data available from CORPES XXI, where the vocative is recorded in other varieties of Spanish too: Colombian (2.26 wpm), Salvadoran (2.03 wpm), Ecuadorian (0.88 wpm), Costa Rican (0.78 wpm) and Panamanian (0.67).

The quality of these varieties is, however, low. The use of the addressing term *parce* is well attested in the concordances of Colombian Spanish in the CdE,[10] even if it is fraught with false positives as a result of typographical mistakes. This is not always the case in the other subcorpora: all the occurrences in Salvadoran Spanish are a mistaken form for *parece* (“parce cada día más vacía”, “me parce muy interesante el comentario”, etc.); in Ecuadorian Spanish, 21 occurrences are for the name *Patricio Parces*; Panamanian Spanish contains 15 occurrences, 5 of which are typographical mistakes and the remaining 10 are vocatives but do not really evidence actual use in this variety: 2 occurrences come from a Colombian website (colombiatvglog.com), 4 are from a staged interview with a footballer from Barranquilla (Colombia), and the remaining 4 are comments on the Colombian TV series *El cartel de los sapos*.

The results available from *EsTenTen18* are unreliable too: at 4721 occurrences, *parce* has a frequency of 0.24 wpm, but most are typographical mistakes. Even more, only 66 occurrences of *parce* out of 217 in the Colombian section (.co) are vocatives. This means that, as the nominal form *parce* is virtually confined to Colombian Spanish, the true positives out of the original 4721 occurrences in the corpus must amount to slightly over 66.

Typographical mistakes mislead annotation and lemmatization to the extent that a high degree of inconsistency can be noticed: *parce* ‘parece’ is sometimes annotated rightly as *VMIP3S0* (i.e. the third person singular of the present tense, indicative mood) but is wrongly ascribed to the lemma *parce*; the opposite, i.e.

10 E.g. “parce, vos tenés que callarte”; “Buenos días, parce, hágame un favor”; “mis parces no se pierden ni un capítulo”; “quiubo, parce”; “vamos palante, parce, sintetiza el taxista”; “¿Parce, y la pasaste bien? Sí, güevón, super chimba”.

parce 'friend, pal' annotated as *VMIP3S0* ("decir parce, en vez de parcero"), is also recorded; some other times, both *parce* 'parece' and *Parce* 'friend, pal' are annotated as *NP* (proper noun), especially if the initial is upper case.[11] The precision of the corpus is, thus, remarkably low and, while it does not make it impossible to research specific cases in detail, data processing becomes significantly more demanding.

False negatives or misses (i.e. "fail[ure] to include instances of our phenomenon" Stefanowitsch 2020: 111) are a different case. In the example under study here, data may be missed, if the spelling associated with the realization of /θ/ in *parce* as /s/ (so-called *seseo*) were not considered. Lemmatization of the vocative does not attest such spelling, so additional queries are necessary for the form *parse* and its plural *parses*.

As in other examples described above, most of the instances retrieved are false positives: the technical term *parse* (meaning 'syntactic analysis') prevails in *EsTenTen18*,[12] and *parse* as a typographical mistake for *parte* 'part' ("parse integrante") distorts the frequency in the Puerto Rican subcorpus of CdE. The only relevant occurrences for this query are ca. 20 concordances taken from a blog about rock music where the author imitates colloquial speech ("eyos escuchan salsa y esa muciquita de regetoneros, parse, que paila que no aya tenido padres metaleros" (rockombia.com, CdE).

The above is intended to show how low data quality may lead to low quality query results and the latter, in turn, to wrong conclusions, e.g. if automatically-generated charts are taken at their face value, i.e. without concordance analysis. Awareness of the strengths and weaknesses of each corpus, i.e. of "the nature and composition of the corpus used" and "the kinds of linguistic information provided by automatic tools" is thus essential (Egbert, Larsson and Biber 2020: 1).

11 *Parce* is annotated as *NP* (*Nombre Propio* 'proper noun') in both "Parce, si usted puede" and "Parce 'parece' el problema de Linux".

12 This is clearly as a result of automatic data collection from computing blogs, which are of little interest for a general corpus of Spanish; even so, some useful concordances can be retrieved: "–No se me ahogue más en alcohol, parse; ya deje de chupar" (foroactivo.com, *EsTenTen18*); "así que pues le dejo ese consejito, parse alivien no se ponga a hacer afirmaciones tan absurdas" (prometec.net, *EsTenTen18*).

3 Beyond Corpora: The Web and the Social Networks

Octavio de Toledo y Huerta (2016) relied on systematic data gathering from online resources (Google Books, Google Scholar, and Google's search engine) to complete the insufficient lexicographical data and the little evidence of *algotro* 'some other' (from 'algún otro') available in RAE corpora. Additionally, he relied on the general archive of the *Real Academia de la Lengua Española* and on oral corpora (COSER). The data collected allowed to attest the origin of the abovementioned indefinite quantifier in Extremadura rather than in Andalusia. The data also allowed to identify the current distribution areas, namely El Salvador, Colombia, Mexico, Honduras, Guatemala, Argentina, Chile, Ecuador, Panama, Costa Rica and Peru (in order of decreasing frequency).

This section reviews the data collected by Octavio de Toledo on the reliability of CdE and *EsTenTen18* as regards research on low-frequency lemmas in RAE corpora. The value of additional evidence of *algotro* gathered from Twitter is then pondered as a qualification of the abovementioned corpus data.

The number of occurrences of *algotro* in RAE corpora is low but representative: 9 occurrences in the CDH between 1896 and 1954,[13] and 2 occurrences of Salvadoran Spanish in CORPES XXI. Figure 5 shows the wpm frequency of *algotro* in the *CdE web/dialects*. According to this figure, the quantifier's distribution by vari-

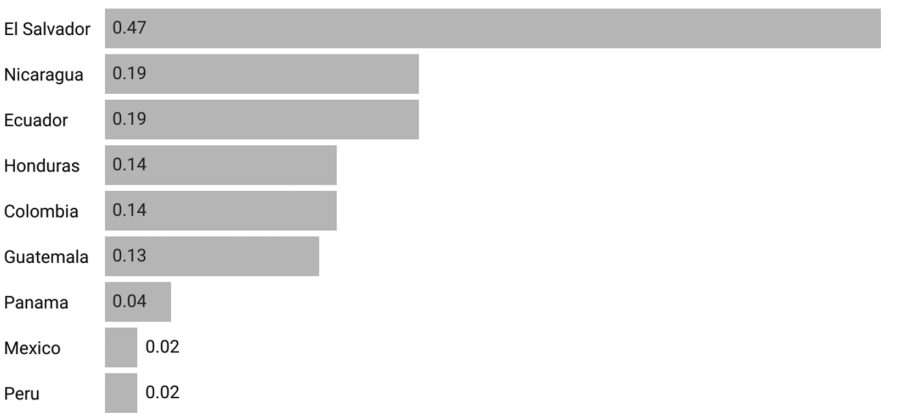

Figure 5: Wpm frequency of *algotro* in the CdE.

13 Of these, 4 are from Colombia, 2 from Honduras, 2 from Guatemala and 1 from Spain, specifically from Felipe Trigo's novel *Jarrapellejos* (1914), set in a village in Extremadura.

ety is close to Octavio de Toledo's claim, i.e. it is used mainly in Central America (El Salvador, Nicaragua, Honduras, Guatemala, and Panama), Colombia, Ecuador and, less frequently, in Mexico and Peru.

EsTenTen contains 132 occurrences of *algotro*, and the wpm frequency is therefore very low: 0.01. 50 concordances of *algotro* can be referred to American sources, 9 to Spanish, and the remaining 73 come from generic websites that cannot be ascribed to a specific variety. Among the above, the 50 American occurrences are distributed very much as described in the former paragraph:

– 5 from El Salvador (0.27 wpm),
– 1 from Guatemala (0.19 wpm),
– 2 from Honduras (0.15 wpm),
– 15 from México (0.1 wpm),
– 9 from Argentina (0.1 wpm),
– 8 from Chile (0.1 wpm),
– 6 from Colombia (0.1 wpm), and
– 4 from Nicaragua (0.1 wpm).

Twitter data reveal facts about *algotro* that are not evidenced by the above sources. The first fifty tweets containing the lemma *algotro* disclose the following distribution by country:

– Honduras (19 occurrences),
– El Salvador (10 occurrences),
– Colombia (8 occurrences),
– Mexico (6 occurrences),
– Guatemala (4 occurrences),
– España (2 occurrences), and
– Argentina (1 occurrence).

The most significant finding is that half the concordances are negative comments on the use of this quantifier. This is especially so in Hondurean Spanish, where 14 out of 19 tweets disapprove the use of this indefinite quantifier:

(7) Feliz día del idioma español . . . menos a los que dicen "haiga" "algotro" "embeces" . . . etc . . . no a ellos no! (Honduras).

(8) En una clase de la U un compañero exponiendo comete el terrible horror de decir algotro y la catedrática le hizo unos ojos que lo quemó y a todos nos quitó puntos por ese error, vieja cabrona (Honduras).

(9) ¿Qué flores se le compra a una dama que dice "haiga" y "algotro"? –Cilantro (Honduras).

(10) Valoro la creatividad de unir “algún otro” en “algotro”, pero no. No, por favor (El Salvador).

(11) Le dice mi esposa a mi hija de 3 años: –Me sorprende oírte decir la palabra algotro ya que es una palabra que cayó en desuso (México).

Twitter evidence stands out in other respects too. Remarkably, one of the examples of European Spanish confirms the current usage of this lemma in Extremadura (“Nos encanta “algotro”, que aún se usa en Extremadura, aunque el DLE no lo diga”). Otherwise, a tweet by a Mexican speaker illustrates the diastratic popular mark of *algotro* in Mexico:

(12) “Algotro lo tengo registrado en una de las entrevistas de mi tesis, de una mujer, de 20 y tantos, con estudios básicos, nacida y crecida en San Felipe, Guanajuato, México”.

Overall, the data available for *algotro* reveal the need for exhaustive procedures in low-frequency lemmas: RAE corpora are a reasonable starting point in that they supply fairly reliable chronological and geographic data. Three further sources may be used for additional evidence: i) CdE and *EsTenTen18* data, ii) Google searches, and iii) Twitter data. The resulting body of data allows the identification of the dialectal, combinatory and sociolinguistic profile of low-frequency lemmas.

4 Small Specific Corpora in the World of Big Data

Besides Big Data sources, small, specific corpora may widen the research data sources quite substantially. Specific corpora or complementary corpora are compiled according to a sample selection whereby the sources must share a given property that is relevant to the research objectives (Rojo 2021: 75). Thus, the sample may be by a given author, of a given literary or musical genre, of a given field of science, of a given period, etc.

Various specific corpora of Spanish are currently being compiled: diachronic corpora, like *Biblia Medieval*, CHARTA, CORDIAM, COREECOM, *CorLexIn*, etc., and spoken corpora, like COSER, ESLORA or PRESEEA. This section reviews two historical corpora managed with TEITOK both for language processing and for data selection and retrieval: *Post Scriptum* (Vaamonde 2017, 2018; Janssen and Vaamonde 2020) and *Oralia diacrónica del español*, ODE (Calderón Campos & Vaamonde 2020).

P.S. is a corpus of private correspondence of the Modern Period (1500–1833). It contains two million words distributed over two surcorpora: one for Portu-

guese and one for mainland Spanish. ODE is a corpus of handwritten documents of the 16th c. to the 19th c. Unlike the P.S. corpus, compilation of the ODE corpus is currently underway. It covers two sample types: i) witness statements at trials, and ii) inventories of personal belongings. The target size of the ODE corpus is one million words, and the original scope of sources has been widened from the historical kingdom of Granada (today's provinces of Granada, Almeria, and Malaga) to the rest of Andalusia plus Extremadura and Madrid. The two corpora allow simple retrieval as facsimiles, as palaeographic samples, and as modern text. CQL searches and result mapping are also available.

At one and two million words respectively, these specific corpora are intended to overcome the dialectal and/or stylistic limitations of the bigger reference historical corpora available of over 400 million words. Their purpose is, therefore, to supply corpus evidence for research on historical dialectology or pragmatics that is otherwise unavailable from larger reference corpora (Calderón Campos & Díaz Bravo 2021).

Regarding dialectal variation, reference corpora limit themselves to the 21 or 22 Spanish-speaking countries (cf. note 1). These corpora allow retrieval of specific usage in European Spanish (e.g. *mogollón* 'a lot', *comerse un marrón* 'to own up to something', *pasteloso* 'cheesy', etc.), Chilean Spanish (*fome* 'boring', *pololo* 'boyfriend', *erís* '[you.SG] are', etc.), or Colombian Spanish (*sumercé* 'you [formal]', *chimba* 'cool, nice', *parce* 'pal', etc.), but not within their regional or local varieties.

Regarding diaphasic or stylistic variation, reference historical corpora rely mainly on formal language, e.g. literature, historical prose, essays, and scientific and legal texts. Informal spoken language is barely represented in the corpora, especially for the period before the 19th c. As a way of example, *vos* 'you' is recorded 668 times in the CDH core subcorpus (European Spanish, 19th c.), most of them from samples of historical novels. Occurrences can be found in other genres too, e.g. 17 occurrences in romance novels like *Eumenia o la madrileña*. Precisely, example 13 illustrates the literary style of this genre, pompous ("tributó lágrimas a los quebrantos de Eumenia") and archaic (as evidenced by the use of *vos* 'you' as an addressing form), but barely representative of informal Spanish of the 19th c.[14] and of addressing terms:

14 Except for what regards the author's *laísmo*, i.e. the use of the feminine form of the pronoun *la* 'her' for the masculine or neuter *lo* 'him' or 'it', or for the gender-unspecified form *le* 'to him/her/it'.

(13) Tributó esta muger amable algunas lágrimas a los quebrantos de Eumenia, diciéndola: "Vos habéis sufrido mil penas, hija mía; lloráis aún la ausencia de un esposo, pero ¿qué sería si os hubiera abandonado antes de serlo, después de sediciros y deshonraros?" (1805, Zavala y Zamora, Gaspar, *La Eumenia o la madrileña*, teatro moral).

By contrast, the samples collected for P.S. and ODE are substantially different from those of reference corpora: not only are they more representative of spoken language, but they have also been transcribed according to the original spelling and thus make available data that would have been missed, if the present-day counterpart of the original samples had been used.

(14) Dijo a uisto y reconozido a la persona de Manuel Rodriges vezino de este dicho lugar, la que reconozida, le hallo vna herida en el vrazo disquierdo en la parte alta de el molleo, echa con instrumento cortante y punzante, como nabaja o cuchillo, y por los accidentes que pueden acadezer, tiene peligro de muerte (ARCHGR, Serie de pleitos, 5233/022, 1753, Cúllar Vega, Granada, ODE).

Example 14 shows how intervocalic *d* was frequently lost in the Spanish spoken in Granada in the 18th c.: *molleo* (referred to an arm) actually meant *'el molledo o bíceps'* 'the lean muscle or biceps' after *-d-* elision. Later hypercorrection is even more significant, as *-d-* was inserted between vowels, as in *acadecer* (for *acaecer* 'happen'). Neither *molleo* nor *acadecer* are recorded in the CDH, whereas 105 occurrences of the full form *molledo* are attested.

The samples compiled for the ODE were selected according to their value as evidence of informal, spoken language, and for the best possible exemplification of the language spoken (and pronounced) in Granada in the Modern Period. Similarly, P.S. contains transcripts of private correspondence, so the language spoken in mainland Spain in the same period can be analyzed:

(15) thio mio con la ocasion de hallarnos muy apurados de dinero ni tener donde cobrar por aber puestole a Dn Balthasar un pleitto las monjas de la conzepzion y aberle enbargado todas las renttas donde abia de cobrar y asta que se concluya no poder cobrar nada cansamos a Vm pidiendole que por amor de Dios nos aga gusto de darnos quatro o cinco mill Rs (1702, P.S.).

Example 15, taken from P.S., is a passage of a letter sent by Catalina Señor to her uncle, Pedro Señor y Angulo. A mother of seven children, Catalina Señor requests funds for child maintenance in her letter. The tenor is thus respectful, with use of the abbreviation *V.M.*, which the corpus editors rightly do not spell out. As the

P.S. corpus contains 9 letters sent by Catalina Señor to her uncle, other letters of the same collection reveal the meaning of the abbreviation: "en casa todos estamos buenos para lo que *usted* nos quisiere mandar que le obedezeremos con la boluntad que Vm sabe", and "de corazon reciui la de *vuersa merced* y siento mucho que mi tia aya malparido".

These passages thus reveal that the full form *vuesa merced* 'your honour' was still in use in the early 18th c. alongside the formal pronoun form *usted* 'you', which by then had become fully grammaticalized.

The letters reveal significant properties of the scribes' language and, by extension, of the lexical resources of that period. The image copies of the documents evidence two different handwritings: one by a scribe who used *seseo* (*resetado* for *recetado* 'prescribed') and *yeísmo*, i.e. the pronunciation of the digraph *ll* as the grapheme *y* (*áyome* for *hállome* 'I am', *ayarme* for *hallarme* 'to be', *aller* for *ayer* 'yesterday'); another by a scribe who used *leísmo*, i.e. the use of gender-unspecified *le* '(to) him/her/it' for masculine *lo* '(to) him' or feminine *la* '(to) her' ("si Vm tubiere un capote que no le sirva me *le* embiara; no canso mas a Vm si no es que me *le* gde Dios") and *laísmo* ("por no dar*la* pesadumbre le digo que no lo se y se me haze escrupulo el que aquella alma pierda las oraziones o misas que *la* puedan dezir").

These letters are also useful for attestation of everyday words that are barely recorded in general corpora. Thus, Catalina, anxious about the cold in Madrid, repeatedly requests from his uncle "2 cargas de *arrax* porque los frios por aca an entrado" ('two loads of [arrax] because the cold set in here'), i.e. two loads of "carbón de huessos de azeituna con que se hace un fuego mui apacible y durable para los braseros" (*Aut*.) ('brazier coal made of olive pits for a very comforting and lasting fire'). This variant form of *errax*, originally from Arabic, was rare as late as the 18th c. and is recorded once in the CDH.

All in all, the above shows that specific questions need both specific *ad hoc* corpora to fill the gaps of general corpora, and the ensuing data analysis and interpretation, which go beyond mere large-scale data collection.

5 Conclusions

Review of the strengths and weaknesses of RAE (CDH and CORPES XXI) and non-RAE corpora (CdE and *EsTenTen18*) reveal higher sample quality and more accurate descriptive and presentational annotation in the former, and bigger size and higher interface flexibility in the latter.

Automatic sample collection from various websites and blogs increases corpus size and is less time-consuming and requires less effort during corpus compilation. Still, there is a downside:

1. sample selection is less precise and, as a result, the resulting corpus is less representative;
2. samples are collected from internet sources with poor geographical metadata, so a large number of examples cannot be ascribed to any language variety or are ascribed wrongly; and
3. sample quality is lower as a result of typographical mistakes (*parce* for *parece*, *parse* for *parte*, etc.) and of inconsistencies (passages in other languages, *parce que*); this results in wrong annotation and lemmatization and, therefore, the degree of precision and recall decreases.

Despite the above, the resulting picture is good, especially if the user is fully aware of the properties of their corpus and, especially, if complementary corpora can be added. The review of *algotro* illustrates the use of this *collaborative procedure* that runs from RAE corpora, goes through CdE and *EsTenTen18*, and reaches internet websites and social networks.

Small, specific corpora can supply data to address research questions that Big Data resources leave unanswered for their lack of highly specific samples and data analysis qualitatively different from their large-scale data collection.

Bibliography

Aut. = RAE (1726-1739): *Diccionario de Autoridades*. <https://apps2.rae.es/DA.html> (14-12-2021).

Biblia Medieval = Enrique Arias, Andrés: *Corpus Biblia Medieval* <http://corpus.bibliamedieval.es/> (12-12-2021).

Calderón Campos, Miguel & Díaz-Bravo, Rocío (2021): "An online corpus for the study of historical dialectology: *Oralia diacrónica del español*", in *Digital Scholarship in the Humanities*, 36, pp. 30–48.

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

Castañeda Naranjo, Luz Stella (2005): *Caracterización lexicológica y lexicográfica del parlache para la elaboración de un diccionario*. Tesis Doctoral. Universitat de Lleida.

CdE= Davies, Mark: *Corpus del español*. <https://www.corpusdelespanol.org/> (15-01-2022).

CDH = Real Academia Española: *Corpus del Diccionario Histórico de la Lengua Española*. <https://www.rae.es/banco-de-datos/cdh> (16-01-2022).

CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <https://www.corpuscharta.es/> (22-12-2021).

CORDE = Real Academia Española: *Corpus diacrónico del español*. <https://www.rae.es/banco-de-datos/corde> (09-01-2022).

CORDIAM = *Corpus Diacrónico y Diatópico del español de América*. <https://www.cordiam.org/> (22-12-2021).

COREECOM = *Corpus Electrónico del Español Colonial Mexicano*. <https://www.iifilologicas.unam.mx/coreecom/index.php?page=inicio&men=1> (22-12-2021).

CorLexIn = *Corpus Léxico de Inventarios*.<https://corlexin.unileon.es/> (22-12-2021).

CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <https://www.rae.es/banco-de-datos/corpes-xxi> (11-01-2022).

COSER = *Corpus Oral y Sonoro del Español Rural*. <http://www.corpusrural.es/> (22-12-2021).

CREA = Real Academia Española: *Corpus de referencia del español actual*. <https://www.rae.es/banco-de-datos/crea> (09-01-2022).

Davies, Mark (2009): "Creating Useful Historical Corpora: a Comparison of CORDE, the *Corpus del español*, and the *Corpus do português*", in Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt: Iberoamericana/Vervuert, pp. 137–166.

Egbert, Jesse, Larsson, Tove & Biber, Douglas (2020): *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge: Cambridge University Press.

ESLORA = *Corpus para el Estudio del Español Oral*. <http://eslora.usc.es/> (22-12-2021).

EsTenTen18 = *Spanish corpus from the web*. <https://www.sketchengine.eu/estenten-spanish-corpus/> (31-12-2021).

González Sopeña, Inmaculada (in press): "Corpus de textos notariales extremeños (CORTENEX S. XVII). La edición de un corpus histórico-lingüístico en el ámbito de las humanidades digitales", in *Dialectologia*, n.º 31.

Jakubíček, Miloš, et al. (2013): "The TenTen corpus family", in *7th International Corpus Linguistics Conference CL*. Lancaster, pp. 125–127.

Janssen, Maarten & Vaamonde, Gael (2020): "Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK", in Rosario Álvarez & Ernesto González Seoane (eds.), *Calen barbas, falen cartas. A escrita en galego na Idade Moderna*, Santiago de Compostela, Consello da Cultura Galega (Ensaio& Investigación), pp. 271–292.

Mair, Christian (2006): "Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora", in Antoinette Renouf and Andrew Kehoe (eds.), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, pp. 355–376.

NGLE = RAE and ASALE (2009): *Nueva Gramática de la Lengua Española*. Madrid: Espasa.

Octavio de Toledo y Huerta, Álvaro (2016): "Sin CORDE pero con red: *algotras* fuentes de datos", in *RILI*, XIV, 29, pp. 19–47.

ODE = Calderón Campos, Miguel & García-Godoy, M.Teresa (dirs.): *Oralia diacrónica del español* <http://corpora.ugr.es/ode/> (16-01-2022).

PRESEEA = PRESEEA (2014-): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. <https://preseea.linguas.net/Corpus.aspx> (12-12-2021).

P.S. = *Post Scriptum. A Digital Archive of Ordinary Writing (Early Modern Portugal and Spain)*. <http://teitok.clul.ul.pt/postscriptum/> (05-01-2022).

Rojo, Guillermo (2010): "Sobre codificación y explotación de corpus textuales: otra comparación del *Corpus del español* con el CORDE y el CREA", in *Lingüística*, 24, pp. 11–50.

Rojo, Guillermo (2021): *Introducción a la lingüística de corpus en español*. London / New York: Routledge.

Stefanowitsch, Anatol (2020): *Corpus Linguistics. A guide to the methodology*. Berlin: Language Science Press.

TEITOK = Janssen, Maarten: *TEITOK, a Tokenized TEI environment*. <http://teitok.corpuswiki.org/> (12-12-2021).

Vaamonde, Gael (2017): *Userguide for Digital Edition of Texts in P. S. Post Scriptum*. <http://teitok.clul.ul.pt/postscriptum/> (15-12-2021).

Vaamonde, Gael (2018): "La multidisciplinariedad en la creación de corpus históricos: El caso de *Post Scriptum*", in *Artnodes*, 22, pp. 118–127.