The *Corpus del español del siglo XXI* (*CORPES XXI*): a tool for the study of syntactic variation in Spanish*

Cristina Buenafuentes de la Mata (Universitat Autònoma de Barcelona)

Carlos Sánchez Lancis (Universitat Autònoma de Barcelona)

(Published in: Buenafuentes de la Mata, Cristina and Carlos Sánchez Lancis (2021): "The *Corpus del español del siglo XXI* (*CORPES XXI*): A tool for the study of syntactic variation in Spanish", in Cerrudo, Alba, Gallego, Ángel J. & Francesc Roca (eds.), *Syntactic Geolectal Variation: Traditional approaches, current challenges and new tools*. Amsterdam / Philadelphia: John Benjamins (*Issues in Hispanic and Lusophone Linguistics* [IHLL], 34), pp. 319-346 (ISBN (hbk): 9789027210517; ISBN (ebk): 9789027259875; DOI: https://doi.org/10.1075/ihll.34.11bue).

Abstract:

In this work we review the multiple uses of the *Corpus del español del siglo XXI* (*CORPES XXI*) to facilitate the study of syntactic variation in Spanish. We present different kinds of searches and data sorting available from the corpus and we provide some examples of syntactic phenomena, most of them characteristic of American Spanish areas. In sum, we show that

* This research has been partially financed with the help of MICINN and FEDER (FFI2017-87140-C4-1-P and PGC2018-094768-B-I00) and CIRIT of Comissionat per Universitats i Recerca de la Generalitat de Catalunya (2017 SGR 634 and 2017 SGR 1251).

CORPES XXI can become an extremely useful tool to describe the syntactic variation of Spanish.

Keywords: syntactic variation, diatopic variation, corpus linguistics, American Spanish.

1. Introduction

Nowadays, the importance of developing big textual databanks in linguistic research is undeniable. Still, a significant progress has already been made in linguistic corpora, as evidenced, for instance, when comparing the two corpora of Modern Spanish of the Real Academia Española (RAE): *Corpus de Referencia del Español Actual* (*CREA*), which has its origins in the decade of the nineties of the 20th century, and *Corpus del español del Siglo XXI* (*CORPES XXI*), which has been launched in 2007 and is still in development phase. The differences between them reflect the progress made in these resources during the twelve years that separate them (cf. Rojo, 2016a), with a clear intention of improving this tool to make it equally valid in the extraction of both syntactic and morphological data.

This contribution aims at showing the usefulness of corpora in the analysis of syntactic variation and, in particular, the utility of *CORPES XXI*. Given that the corpus is still under construction, at present it merely

contains 312 million forms instead of the 400 million forms planned (*CORPES XXI*, beta version 0.92). Nevertheless, we will show that it is suitable for the study of syntactic variation in Spanish, paying special attention to its flexibility, since it is a tool that accommodates to the different needs of researchers.

For this purpose, we first describe the structural parameters of *CORPES XXI* (section 2), which determine the design of the corpus and, therefore, are essential when researchers make use of it. We next show how it can be used for the study of syntactic variation in Spanish (section 3). To do that, we expose the different possibilities of search offered by the corpus and we explain how they can be adjusted to obtain syntactic data. Moreover, we indicate how to make the results obtained more accessible and how to treat them from the standpoint of syntactic variation. In order to illustrate all the possibilities that *CORPES XXI* offers to researchers, we deal with a series of syntactic phenomena that, according to the *Nueva gramática de la lengua española* (*NGLE*) (cf. RAE-ASALE 2009), are characteristic of certain Spanish varieties.

The ultimate goal of this analysis is to demonstrate that *CORPES XXI* is a useful tool for the description of syntactic variation in Spanish.

2. Structural parameters of *CORPES XXI*

From a chronological point of view, *CORPES XXI* is a reference corpus that collects current Spanish texts. As pointed out by the RAE (www.rae.es), the main purpose of *CORPES XXI* is to be used as a tool to obtain the main characteristics that a language presents at a given moment in its history. In particular, the corpus is composed of texts from the year 2001 to the year 2020.[1] In fact, it is the continuation of *CREA* as the reference corpus of present day Spanish and, accordingly, it has become the main source of data for RAE's works (especially for its dictionary).

From a structural point of view, an important achievement of *CORPES XXI* is the attention it pays to the representativeness of the linguistic reality. For this reason, the percentage of texts from each geographical area has been determined according to the presence of Spanish speakers in each of these varieties. Attending to this criterion, the corpus gathers American Spanish texts and Peninsular Spanish texts in different proportion (70% American Spanish and 30% Peninsular Spanish). This is so because the number of Spanish speakers is higher in America than in Spain.[2] Each of the American linguistic areas has also been assigned a separate percentage.[3]

[1] *CORPES XXI* is conceived of as a semi-open corpus that will increase in the coming years, always following the representativeness granted to each of the parameters.

[2] It is worth mentioning that in *CREA*, for instance, the proportion in this parameter is 50%-50%. In this paper the version 0.83 (June 2016) has been used. The later versions of the corpus (0.9, June 2018; 0.91, December 2019 and 0.92, May 2020) maintain the same structure parameters and search parameters which are described and illustrated in the present study, so the only difference between these versions of the corpus consists in a greater number of forms and a wider range of years included.

[3] The RAE has conducted this distribution by geographical areas and it has labeled them by taking into account the traditional dialectal classification. The distribution by geographical areas and countries is the following one: Mexico and Central America: Mexico, Guatemala, Honduras, Nicaragua, Costa Rica and Panama; River Plate: Argentina, Uruguay and

Thus, Mexico and Central America, River Plate area and Antilles have a higher percentage of representation (19%, 13% and 12%, respectively), whereas Continental Caribbean, the Andean area, Chile and USA have less weight in the corpus (9% the former and 4% the latter). This distribution is based on different parameters, to wit, the population, the amount of publications, and the number of digital editions of newspapers and magazines, among others. Another remarkable fact of this corpus is that it contains documents from areas where Spanish is a co-official language as, Equatorial Guinea or the Philippines, for instance. However, the accuracy found in the geographical classification of the American Spanish texts is missing in the case of Peninsular Spanish: the latter are all classified as texts from Spain, and therefore it is not possible to know their dialectal origin, which would be very interesting for the analysis of the Peninsular varieties of Spanish.

Concerning the characteristics of the texts beyond their geographical origin and their temporal limits, *CORPES XXI* is characterized for encompassing a broad typology of texts that addresses different parameters. First of all, one of the advantages of *CORPES XXI* is that it brings together both written and oral texts, although the weight of each of them is significantly unequal, since only 10% of the documents are oral.[4]

Paraguay; Antilles: Cuba, Dominican Republic and Puerto Rico; Continental Caribbean: Venezuela and Colombia; Andes: Peru, Bolivia and Ecuador; Chilean: Chile; USA: USA.

[4] *CORPES XXI* allows listening the aligned sound of the oral texts (see section 3.2).

The written texts mainly belong to books (40%) and to periodic publications (40%),[5] but these are not the only formats that the corpus incorporates: another relevant contribution is the inclusion of Internet texts such as blogs, emails, or messages in networks or discussion forums (7,5%).[6] This makes the corpus a clear reflection of the new ways of communication derived from new technologies. Besides, the written documents are classified taking into account if they are fiction or non-fiction texts, and accordingly one can clearly distinguish literary documents (i.e., novels, theatre plays, movie scripts, short stories) from non-literary ones. Non-fiction texts represent the 75% of the written documents, which points towards a preference of the corpus for a less elaborate type of language than the literary one, the latter usually linked to fiction texts.

Finally, documents are classified according to other related parameters: topic and textual typology. Both criteria are of great relevance to relate certain linguistic uses to their specific topic or typology. Non-fiction texts are classified in accordance with six topics: "News, leisure and daily life", "Art, culture and shows", "Social sciences, beliefs and thought", "Sciences and technology", "Politics, economy and law" and "Health". These topics are represented in different degree within the corpus. Hence, "Politics, economy and law" and "Social sciences, beliefs and thought" represent, each one, a 20% of the texts, and the rest represent a 15%.

[5] Notice that the periodic publications included in *CORPES XXI* are from digital media.
[6] Other textual types like, for example, advertising brochures or medicine leaflets are classified upon the label "miscellany" (7,5%).

Non-fiction texts show different textual typologies depending on their being journalistic texts or not being so. Journalistic texts may be classified as "News", "Report", "Letter to the Editor", "Review", "Column", "Dissemination", "Editorial", "Interview", "Opinion" and "Academic"[7], whereas non-journalistic texts, specially books, are classified as "Academic", "Biography, memoirs", "Dissemination", "Legal-administrative" and "Text book". Textual typology is the only parameter that has not been assigned a percentage to measure its presence in the corpus. Therefore, there might be an imbalance among the journalistic texts corresponding to "Reviews" and those corresponding to "Letters to the Editor".

As for fiction texts, they are classified according to the usual textual genres, namely, "Novel", "Theatre", "Script" and "Short story". Novels and short stories represent, each one, the 40% of fiction texts, whereas the remaining 20% is represented by theatre and movie scripts.

Although the textual classification offered in *CORPES XXI* is very precise and takes into account the possible needs of the researcher, the level of formality of the texts is not considered. It is well known that the communicative situation has an influence on the type of discourse and that several factors between distance and communicative immediacy must be taken into account: the more or less public character of the communicative

[7] Given the difficulty of finding press news on certain topics (for example, "Science and technology" or "Health") in certain areas, the corpus also includes research articles from scientific journals that are classified as "Academic".

act, familiarity between the interlocutors, the degree of emotional involvement with regard to the interlocutor or the topic of the conversation, spontaneity, etc. (see Koch & Oesterreicher, 2007). These determining factors are difficult to deal with in a corpus that contains a large amount of words, as is the case of *CORPES XXI*. In fact, dealing with them would require a detailed examination of the different communicative situations found in each text, which would be impossible.[8] Despite this, thanks to some of the parameters that have been described, researchers can deduce the degree of formality of the documents from the communicative context provided when a particular example is chosen.

To sum up, *CORPES XXI*, because of its design, is an ideal tool to extract data that can be analyzed from the point of view of syntactic variation. In the following section we comment on the possibilities offered by *CORPES XXI* to get and to analyze data.

3. Description and possibilities of *CORPES XXI* in the analysis of syntactic variation in Spanish

Knowing the possibilities that a corpus offers to the researcher is essential to get the most of it. In the particular case of *CORPES XXI*, its

[8] However, other corpora, like *PRESEEA* (http://preseea.linguas.net/Inicio.aspx), classify texts according to different parameters associated with the communicative situation.

features focus on two aspects: first, on the different types of search that can be done to obtain the data, and, second, on the processing of the results obtained from the search. In what follows we describe how a researcher can work with *CORPES XXI* paying especial attention to these two fundamental functions. This will provide a concise idea about the usefulness of this corpus in the analysis of syntactic variation in Spanish. With this goal in mind, we deal with some syntactic phenomena that, according to the *NGLE* (2009), show variation from a diatopic point of view.

3.1. Simple search

The first type of search that the researcher can do is the simplest one, that is, the search for a specific lexical form (see Figure 1, number 1).

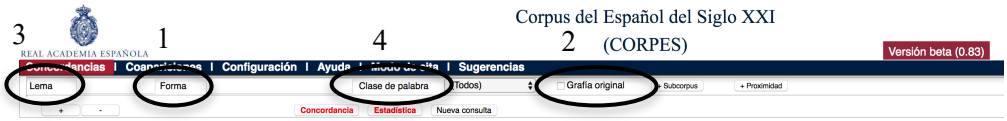


Figure 1. Main page of *CORPES XXI*.

Simple searches allow researchers to choose the option of keeping the original spelling of the searched item (see Figure 1, number 2). This is very useful because, if the researcher chooses this option, then the corpus distinguishes between capital and small letters (which helps differentiating, for example, common nouns from proper nouns), or between accented and

not accented words (which allows distinguishing, for instance, the adverb *aún* ‘still’ from the conjunction *aun* ‘even’).

Within this type of search, the corpus allows going further and doing searches by lemmas or by grammatical categories (*CORPES XXI* is a lemmatized and categorized corpus). Lemmatization makes it possible to obtain all the inflected forms of an element, whereas categorization allows searching by grammatical categories (noun, verb, adjective, adverb, quantifier, conjunction, etc.) (see Figure 1, number 3). In this last aspect, we can further refine the search by word class. For example, it is possible to indicate time and mode of conjugation or degree in the case of, respectively, verbs and adjectives (see Figure 1, number 4). In the field of linguistic corpora, lemmatization and categorization are a significant progress for the study of the morphology and the syntax of a language (see Rojo, 2014; Rojo, 2016b). In fact, extracting syntactic information from a corpus was difficult until relatively recently, given that investigators had to work with the data manually. The lemmatization and categorization of the corpus is a key factor that makes the task of gathering syntactic data more feasible. Although it is important to be aware of the fact that the researcher should always filter the results, the progress made in this regard is very important.

The following table summarizes the search possibilities based on grammatical categories as well as the degree of concretion that can be achieved.

Noun	• Number (plural, singular, syncretic) • Gender (common, masculine, feminine, neuter, triple) • Type (common noun, proper noun)	Possessive	• Number (plural, singular, syncretic) • Person (first, second, third) • Gender (common, masculine, feminine, neuter, triple) • Function (determiner, pronoun) • Possessor (ambiguous, singular, plural)
Verb	• Mode (imperative, indicative, subjunctive) • Tense (present, past, present participle, past participle, infinitive, imperfect, gerund, future, conditional) • Person (first, second, third) • Number (plural, singular, syncretic) • Type (imperfect subjunctive in *-ra*, imperfect subjunctive in *-se*)	Personal pronoun	• Number (plural, singular, syncretic) • Person (first, second, third) • Gender (common, masculine, feminine, neuter, triple) • Case (accusative, dative, unspecific, prepositional, unmarked)
Adjective	• Number (singular, plural, syncretic) • Gender (common, masculine, feminine, neuter, triple) • Type (qualifying, title) • Degree (comparative, superlative, positive)	Numeral	• Number (plural, singular, syncretic) • Gender (common, masculine, feminine, neuter, triple) • Type (partitive, ordinal, cardinal) • Function (determiner, pronoun)
Adverb	• Type (affirmative, deictic, intensifier, interrogative, negative, relative) • Degree (comparative, superlative, positive)	Conjunction	• Type (coordinating, subordinating)
Relative Interrogative Demonstrative Quantifier Amalgam	• Number (plural, singular, syncretic) • Gender (common, masculine, feminine, neuter, triple) • Function (determiner, pronoun)	Article	• Number (plural, singular, syncretic) • Gender (common, masculine, feminine, neuter, triple)
Punctuation Interjection	Preposition Unknown Affix	It cannot be specified beyond the basic grammatical category	

Table 1. Grammatical categories in *CORPES XXI*.

Some of these categories and sub-classifications are questionable (for instance, the category "unknown") and other possibilities are missing (auxiliary verbs, relational adjectives, etc.), but, the fact that the corpus allows searches that take into account the categorization of linguistic units is an important step forward, especially to obtain grammatical data, as will be illustrated below.

As pointed out in the literature (Demonte, 1999; Hummel, 2014; Bosque, 2015), many adjectives are used in certain contexts as quantifiers. One of these cases is the use of the adjective *harto* 'full' to express 'a lot of' in sentences such as *Vino harta gente* 'a lot of people came' or *Hace harto calor* 'it is very hot'. According to the classification proposed by *NGLE* (2009, § 19.2a, p. 1381), this adjective would occupy "an imprecise place" between *mucho* 'a lot' and *demasiado* 'too much' within the hierarchy of evaluative quantifies. It must also be taken into account that this quantifier can also be categorized as an adverb with the value of 'very' (*está harto contenta* 'she is very happy'), in which case it is invariable in gender and number. The *NGLE* (2009, § 19.2n, p. 1386) also points out that "this quantifier is typical of the elevated registers of accurate language in general Spanish [...] but belongs to the standard language of Chilean Spanish and that of the Andean countries, especially in Bolivia and Ecuador",[9] and also that "the use of *harto* in many American countries is not associated with the archaic connotation[10] that characterizes this quantifier in European Spanish" (*NGLE*, 2009, § 20.7a, p. 1484).

In *CORPES XXI* it is possible to get examples of this type of syntactic constructions. In fact, the search by lemma not only allows taking into account the inflected forms of the adjective (see Figure 2), but also

[9] All citations in this research are translated by the authors.

[10] This use is documented in Old Spanish until the nineteenth century, so it is not an innovation, but rather the preservation of a syntactic construction that has survived in a different way in peninsular Spanish (contexts of high formality and archaic character) and in the Spanish of some countries of America (conversational language). For an analysis of this quantifier from a historical point of view, see Espinosa (2014) and Pato (2016).

obtaining other variants of the adjective. Thus, we can find cases in which the adjective is used in the superlative form *hartísimo* 'extremely full' and also works as a quantifier meaning 'a lot' (1), a variant that perhaps the researcher had not contemplated at first:

(1) a. "Dicen que hay *hartísimo* dinero en Bolivia,
say.3PL that there.are extremely.full money in Bolivia
entonces, ¿qué es lo que está ocultando?
so what is it that is hiding
'It is said that there is a lot of money in Bolivia, so, what is he hiding?' (Escobar, Roxana: «Ni la patria ni la Iglesia evitan la pugna política por el feriado». *Eldeber.com.bo*, 2008-08-04, Bolivia)

b. … o gano cuatrocientos pesos, y tenemos
or win.1SG four.hundred pesos and have.1PL
hartísimo trabajo, ¿qué le parece?
extremely.full work what 3SG.DAT seems
'… or I win four hundred pesos, and we have a lot of work. What do you think?' (Ramírez Hita, Susana: *Calidad de atención en salud. Prácticas y*

representaciones sociales en las poblaciones quechua y aymará del altiplano boliviano, 2009, Bolivia)[11]

Likewise, the corpus makes it possible to specify the word class, which is also a great step forward to distinguish the uses of *harto* as a full adjective (*estaba harto de todo* 'he was sick of everything') from its uses as a quantifier (see Figure 2).

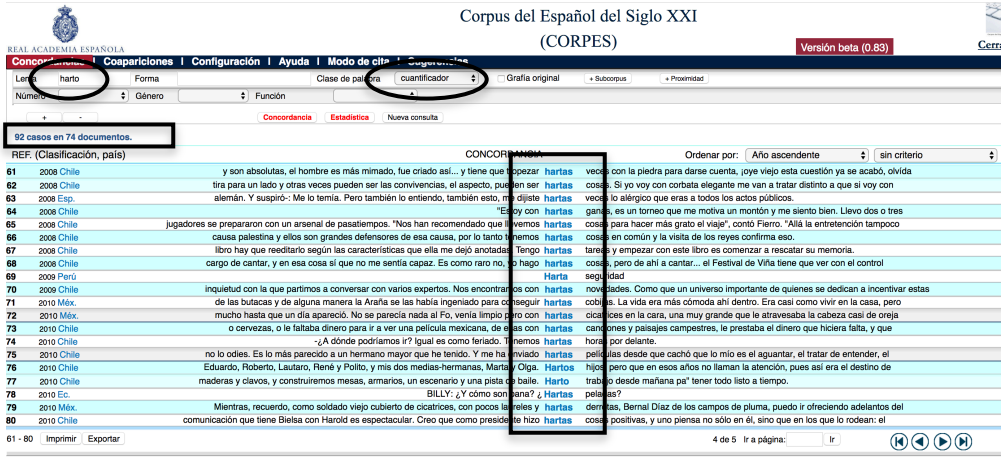


Figure 2. Search results of the quantifier *harto* in *CORPES XXI*.

The corpus indicates the number of cases per document (in this particular case, 92 cases in 74 documents) and it also offers information about the year and the geographical origin of the documents. The outputs can be further refined by using several search parameters, as we will see in the following section. If the researcher wants to know quickly the exact reference of the

[11] This example shows the use of the quantifier *harto* in the conversational language of some American countries such as Bolivia, since, as we can see, it reproduces some statements of a person from that origin.

document containing the example, it is enough to place the mouse pointer on the year of the example (see Figure 3, number 1), but to get a more detailed information, it is necessary to click on the example. Then, all the information of the document appears at the bottom of the screen and it is also possible to expand the context or, in case of oral texts, to play the audio.

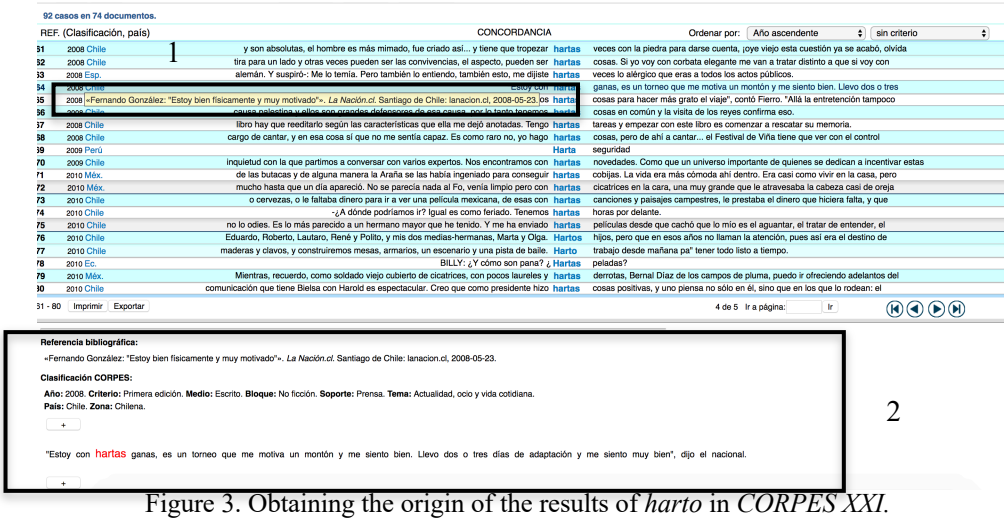


Figure 3. Obtaining the origin of the results of *harto* in *CORPES XXI*.

3.2. Complex search

CORPES XXI allows for two types of complex searches: subcorpus search and proximity search. From the standpoint of syntactic variation, both possibilities are very useful, given that the former offers the option of filtering the search by geographical origin (among other parameters) and the latter allows searching combination of words.

3.2.1. Subcorpus search

The possibilities of the extraction of information by subcorpus are closely related to the structural parameters presented in section 2. The corpus allows extracting information according to the Spanish varieties from the subcorpus search. We can discriminate between European and American Spanish, and, within the latter, delimit by linguistic areas or countries according to the distribution we indicated in section 2 (see Figure 4). Additionally, the search can be filtered with the parameters described in section 2 (oral or written texts, fiction or non-fiction, topic, textual typology, etc.):

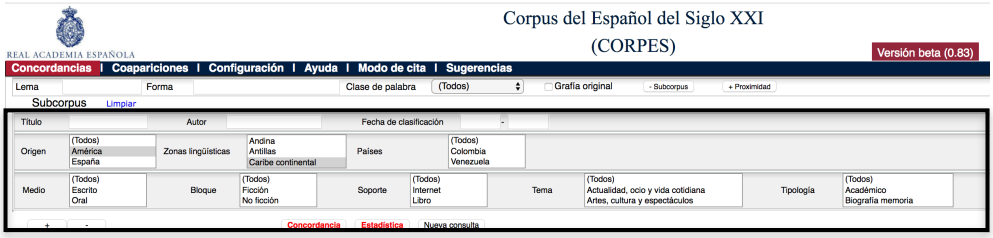


Figure 4. Subcorpus search of *CORPES XXI*.

Taking as an example the use of *harto* as a quantifier meaning ‘a lot’, we can check if this use is only documented in American Spanish or if a given construction is only found in oral texts or in fiction texts.

3.2.2. Proximity search

Another aspect that makes this corpus a valid tool for the study of syntactic variation is the possibility of extracting data on a unit combination through the proximity search. Within this type of search, the use of lemmatization and categorization greatly broadens the possibilities of obtaining data. Besides, if we indicate the distance between the linguistic units and the order in which they appear (see Figure 5), we are able to identify sequences with interpolated elements.

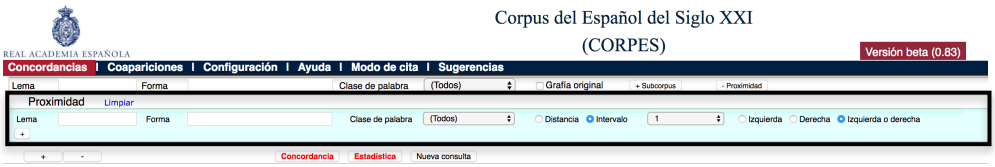


Figure 5. Proximity search of *CORPES XXI*.

Only a few years ago researchers couldn't obtain information on the use of the form *ello* 'it' as a direct object. The proximity search of this corpus allows us to discover more about this use from a diatopic point of view.

According to the *NGLE* (2009), the demonstrative *eso* 'that' and the personal pronoun *ello* 'it' differ in the fact that *ello*, contrary to *eso*, never appears as direct object except "in the colloquial Spanish of Peru and other zones of the Andean area [...] that the written language occasionally reflects" (*NGLE*, 2009, § 16.2e).

To corroborate or refute this, we start from the premise that the personal pronoun does not have inflection, so we can look for it using the form

search instead of the lemma search. Given that the direct object, in its unmarked position, usually appears immediately after the verb, we can indicate that the position of *ello* will be postverbal (extending, if needed, the range of proximity). Thus, using the word class search, it is possible to search for any verb followed by the pronoun *ello* (see Figure 6).

Figure 6. Search results for *ello* as a direct object in *CORPES XXI*.

The corpus offers examples of the use of *ello* 'it' as a direct object (*sustituían ello* 'they replaced it', *ejemplificando ello* 'exemplifying it', *ver ello* 'to see it'), but also of its use with other syntactic functions (*siendo ello* 'being it', *bastaría ello* 'it would be enough', *dicho ello* 'said it'). Since the corpus does not distinguish syntactic functions, the researcher must perform the corresponding syntactic analysis and discard those cases that do not correspond to the phenomenon sought.

One can also verify the use of *ello* outside the areas indicated by the *NGLE* (2009), such as Colombia, and also outside the colloquial language:

(2) a. Hay pruebas para demostrar *ello*, dijo.
There.is evidence.PL for prove it said
'There is evidence to prove it, he said.' («El culebrón de la reforma política». *El Tiempo*, 2001-05-20, Colombia)

b. ¿Qué han hecho ustedes para ver *ello*?
what have.2PL done you.PL to see it
'What have you done to see it?' (*Debate: funciones del Congreso y de los congresistas*, 3/4, 01/12/01, CNR, Peru)

Due to the complexity of the construction, the researcher must precisely arrange and further delimit the data obtained, but despite these limitations, the corpus proves to be a tool of great utility to collect syntactic constructions such as these.

The search through grammatical category allows attesting syntactic combinations such as the possessive plus demonstrative construction. The *NGLE* (2009, §§ 17.4z and 18.2i-j) indicates that this type of structure was common in Spanish until the 18th century and considers it as archaic

nowadays,[12] but it still survives in some dialectal varieties of the western Iberian Peninsula and in some zones of the Andean area.

In this particular case the category specification that *CORPES XXI* incorporate distinguishes clearly between the use of the elements as pronouns or as determiners. This allows us to obtain the relevant data by searching through grammatical category and proximity (see Figure 7).

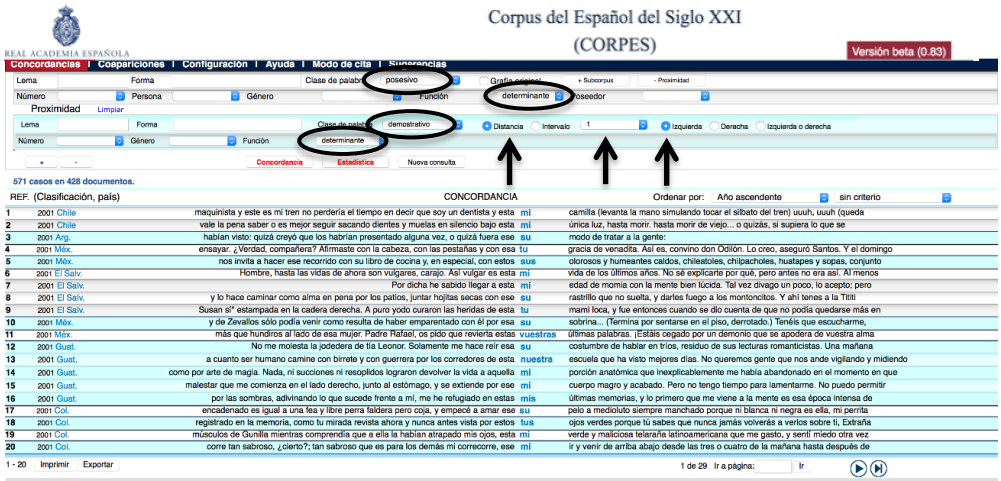


Figure 7. Search results of the demonstrative + possessive construction in *CORPES XXI*.

We realize that this construction is used in current Spanish, as the following examples show:

(3) a. *Esta mi* segunda luna de miel con madame Arnoux

this my second honey moon with madam Arnoux

[12] In fact, the *NGLE* (2009, § 17.4z) relates this structure to the Old Spanish construction with a definite article and a possessive (*la mi casa*), since "the prenominal possessive brings some relational information as a restrictive modifier […], so that the demonstrative or the preceding article constitute the true determinant".

terminó poco después de aquella cena.

Finished.3PL few after of that dinner

'My second honey moon with madam Arnoux ended shortly after that dinner.' (Vargas Llosa, Mario: *Travesuras de la niña mala*, 2006, Peru).

b. De dónde vendrán nuevas alegrías, […], a

from where come.FUT.3.PL new joys to

estas nuestras tierras …

these our.F.PL lands

'From where new joys will come to our lands …' (Mateo Vasquez, Eddy: "Presa de Monte Grande: Umbral del Sur redimido". Listindiario.com, 2005-10-06, Dominican Republic)

c. … mientras posteo en *este nuestro* blog …

while post.1SG in this our blog

'… while I post in our blog …' (Alonso Coto, Manuel: "Medios off como soporte de campañas online". *Marketing Weblog.* marketing.blogs.ie.edu, 2007-09-18, Spain)

d. … vos cantabas tan bien con *esa tu* guitarra.

you sang so well with this your guitar

'You sang so well with your guitar.' (Núñez, Agustín: *Brillo de luna*, 2005, Paraguay)

3.2.3. Combination of searches

The possibilities of the corpus increase if we combine searches. Hence, not only can we search simultaneously by subcorpus and by proximity, but also by proximity twice or more. This allows us to refine our search by specifying the form, the lemma or the grammatical category preceding or following the item.

In order to illustrate the combination of subcorpus search and proximity search, we will use the syntactic structure formed by synthetic and analytic comparatives (*Tu ordenador es más mejor que el mío* 'Lit. Your computer is more better than mine'). Kany (1963/1969, p. 71) points out that it is common practice to combine the analytical and synthetic comparative in Spanish. According to him, this construction is already attested in Latin and is found in classic authors, but nowadays it is relegated "to illiterate people in popular and rustic general use, both in Spain and America". The explanation for this phenomenon is due, on the one hand, to the fact that most adjectives (the exceptions are *mejor* 'better', *peor* 'worse', *mayor* 'bigger' and *menor* 'smaller') are built in the comparative degree through analytical processes and, on the other hand, to the fact that speakers reanalyze synthetic comparatives as positive degree forms and, for this reason, they include these comparative adjectives in analytical comparative

structures[13] (see Vigara Tauste, 2010). The *NGLE* (2009, § 13.3d) refers to this construction in the following terms: "in the rural language of many Spanish-speaking countries, lexical comparatives combined with syntactic comparatives are attested". The account of this phenomenon is not, then, very precise, but *CORPES XXI* can greatly help to refine our knowledge.

If the corpus is consulted regarding the forms of analytical comparatives of superiority (*más* 'more'), inferiority (*menos* 'less') and equality (*tan* 'as much as'), we can refine the search by incorporating proximity to any comparative adjective and limiting it to America (see Figure 8).

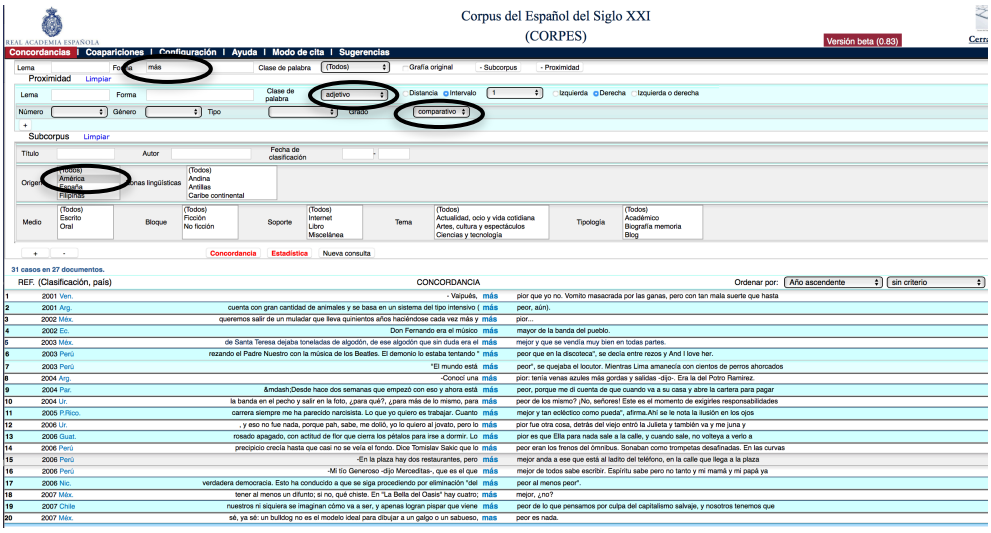


Figure 8. Search results of the combination of synthetic and analytical comparatives in *CORPES XXI*.

[13] Sánchez López (2006, p. 39) explains that "other adjectives, being etymologically comparative, currently lack this value and are exclusively positive. These are the adjectives *superior, inferior, anterior* and *posterior*".

We found examples in countries such as Mexico, Colombia, Uruguay, Argentina, Peru, Paraguay, Venezuela, Guatemala and Chile, but not in Cuba or other parts of Central America, such as Honduras, Costa Rica or El Salvador, contrary to what Kany (1963/1969, p. 71) pointed out:[14]

(4) a. — Mi tío Generoso —dijo Merceditas— que es
my uncle Generoso said.3SG Merceditas that is
el que *más mejor* de todos sabe escribir.
the that more better of all knows write
'My uncle Generoso —said Merceditas—, who is the one that better knows writing.' (Ferrini, Ernesto: *La tristeza de los burros*, 2006, Peru)

b. … es el peor filme del año. *Tan peor* como el
is the worse film of.the year so worse as the
libro, la miniserie y la novela.
book the miniseries and the novel
'… is the worse film of the year, as bad as the book, the miniseries and the novel.' (Morales, Nicolás: "Los sopores del 2010. Sopor y piropos". *Revista Arcadia.com*, 2010-12-15, Colombia)

c. Kerry no es una alternativa, pero es *menos peor*

[14] 71 cases are collected in Spain, but most of them respond to the older structure to refer to people of 'advanced age', which indicates that the greater adjective refers to age and not to size and, therefore, is an adjective in a positive degree. Only four cases were a combination of the two comparative structures.

Kerry not is a alternative but is less worse
que Bush.
than Bush
'Kerry isn't an alternative, but he isn't as bad as Bush.'
(Rascón, Marco: "El modelo Kerry de legitimación". *La Jornada*, 2005-05-10, Mexico)

d. … y se basa en un sistema del tipo intensivo
and CL base in a system of.the type intensive
(*más peor*, aún).
more worse even
'… it is based on a system of the intensive type (even worse).' (Montiel, Eduardo F.: "Medio Ambiente: Manejo de desperdicios, qué hay que hacer... lo que hay que cuidar". *Producción Agroindustrial del Noa.* produccion.com.ar, 2001-08, Argentina)

The possibility of doing several proximity searches simultaneously can be combined with the search by lemma and by grammatical category. This allows obtaining examples of syntactic constructions such as the Spanish explicative relative sentences containing the expletive use of *mismo* 'which' or the phenomenon of number discordance in the accusative pronoun.

Regarding the expletive use of *mismo* in explicative relative sentences (*Fue un intenso bombardeo, mismo que se apaciguó después* 'It was a

heavy bombing, which was later on pacified'), the *NGLE* (2009, § 13.11n) points out that this phenomenon is attested in Mexico, Central America, the Andean area and in the youth speech of certain areas of the River Plate area. The structure *mismo que* involves the omission of the definite article in front of it, so the construction is used as *el cual/la cual* 'who, which':

(5) a. … para conseguir cuatro puntos más, *mismos que*
for get four points more same.M.PL that
no fueron suficientes …
not were enough
'… in order to get four extra points, which were not enough …' (Velázquez, Ariel: "Hornets arruina la noche a Ayón". *El Universal.mx*, 2012-10-08, Mexico)

b. … en presencia de su esposa, *misma que* ipso facto
in presence of his wife same.F.SG that ipso facto
no dudó un instante.
not hesitated one instant
'… in the presence of his wife, who didn't hesitate for a single moment.' (Cuevas Molina, Rafael: *Una familia honorable*, 2008, Guatemala)

c. … lo podrían decir algunas personas que la
3SG.ACC could say some people that 3F.SG.ACC
han experimentado, *mismas que* […] cedieron

have.3PL experienced same that give.in.PST.3PL
a la dieta cítrica.
to the diet citrus
'… it could be said by some people that have experienced it, who gave in to the citrus diet'. (Garcés, Laura: "Piedras renales y dieta cítrica". *Blog Salud y Belleza Natural*. saludnatural.biomanantial.com, 2008-01-08, Spain).

In order to find this data in *CORPES XXI*, we need to combine different resources offered by the corpus. In this case, we will use the search by the lemma *mismo* (which allows us to take into consideration all the inflected forms at the same time), and we will combine two criteria in the proximity search to filter the element that appears before or after *mismo*. First of all, we specify the category *relative* to the right of the element searched, which allows us to be sure that *mismo* is expletive —since it is the use that *mismo* gets when preceded by a relative, in which case it is a redundant element with respect to the other relative; secondly, we specify the punctuation mark, which is crucial to find this type of constructions because the expletive *mismo* only appears in explicative relative sentences, that is, in

those delimited by commas.[15] Consequently, the search will have the following structure: comma+*mismo*+relative (see Figure 9).

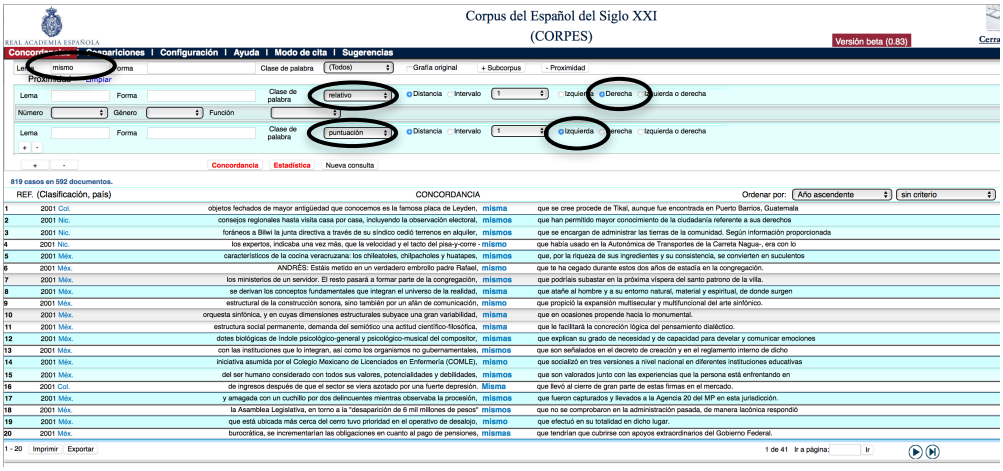


Figure 9. Search results of expletive *mismo* 'which' in the explicative relative sentences in *CORPES XXI*.

Number disagreement in accusative pronouns is another interesting case of syntactic variation. In sentences like *Aquello* [SG] *se los* [PL] *dije bien claro a tus amigos* 'I told that very clearly to your friends' the accusative clitic *los* agrees with the indirect object (*a tus amigos* 'to your friends') and not with the direct object (*aquello* 'that'). The *NGLE* (2009, § 35.2h) notes that, sometimes, when the pronoun *se* refers to the indirect object, the direct object agrees with thein number with the referent of the dative[16] because *se*

[15] This is also a big advantage to look for parenthetic discourse markers, for example.
[16] Rivarola (1985), Mello (1992) and Fernández Soriano (1999, pp. 1257-1258) consider that the two clitic pronouns (indirect object+direct object) form a morphological unit where the plural morpheme *–s* is inserted at the end. Besides, Company (1992, 1998) shows that this construction was gramaticalized by subjectivization, because speakers reanalyze the morpheme *-s*, which has a plural value and features of the dative [+ animated, + human].

is invariable. This only occurs when the indirect object has a plural referent and the direct object has a singular referent, so we are dealing with a phenomenon of hypercharacterization of number.

In addition, the RAE indicates that this construction is "common in the oral and colloquial language of many areas of America, as well as in Canary Spanish" and that it is extending "to the cultured registers (in Mexico, Continental Caribbean and part of the Central American, River Plate and Andean areas); in others (Chile, Spain and part of the Andean and Antillean areas), it is not considered to be part of these registers" (*NGLE,* 2009, § 35.2h).

In order to be able to attest the existence of this syntactic variant in *CORPES XXI*, we search the pronominal sequence *se los* and *decir*. In this case, it is difficult to find examples using the corpus possibilities, because not all the resulting examples belong to this kind of structure. Thereby, starting from the pronoun *los*, we have searched for the form *se* that immediately precedes it and the lemma *decir* (that is, all the conjugated forms of this verb) located immediately to its right, without specifying the grammatical category of any of these elements. The search result is as follows:

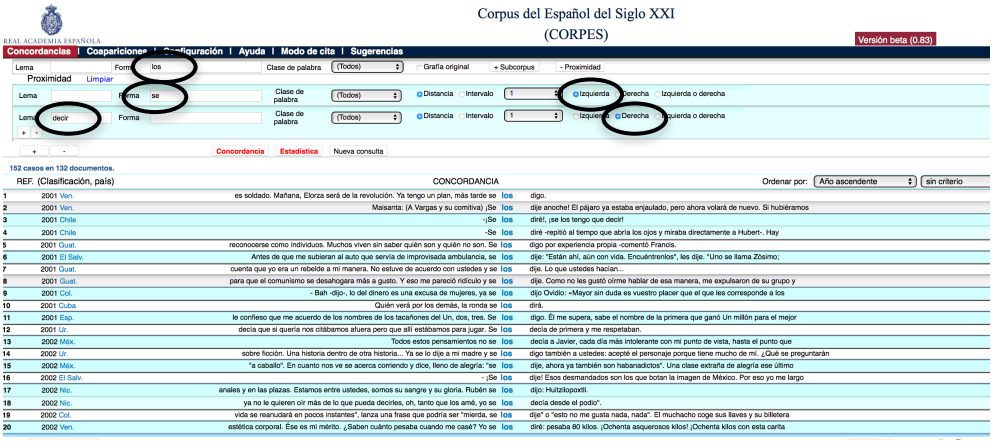


Figure 10. Search results of the structure *se los* in *CORPES XXI*.

CORPES XXI registers 152 cases of this construction in 132 documents. Some examples are reproduced below:

(6) a. … *se los dije*, ahora ya también
3.DAT 3PL.ACC told.1SG now already also
son habanadictos.
are Havana-addicts
'… I told them, now they are Havana addicts too.'
(Partida, Eugenio: «La Habana Club». *La Habana Club y otros relatos*, 2002, Mexico)

b. Pero vivir es más que saberse vivo. *Se*
but live.INF is more than know.INF.SE alive 3.DAT
los digo yo que ahora que supuestamente menos
3.M.PL.ACC. told I that now that supposedly less
vivo …

alive

'But living is more than knowing oneself alive. I told it now, when I am apparently less alive.' (Infante Guell, Manuela: *Rey planta.* www.escenachilena.uchile.cl, 2014-06-16, Chile)

In these examples, the clitic *los* is plural, but it is referring (anaphorically or cataphorically) to an statement, which has to be expressed in singular, and the clitic *se* has the plural referent *ustedes* 'you' (addressee with 3rd person morphology).

In this construction gender disagreement is more marked than number disagreement (*NGLE,* 2009, § 35.2i). In order to test this generalization, we use the same search criteria as in the previous case, but substituting the form *los* for *las*. The search result provides 36 cases in 33 documents; only two of them show gender disagreement:

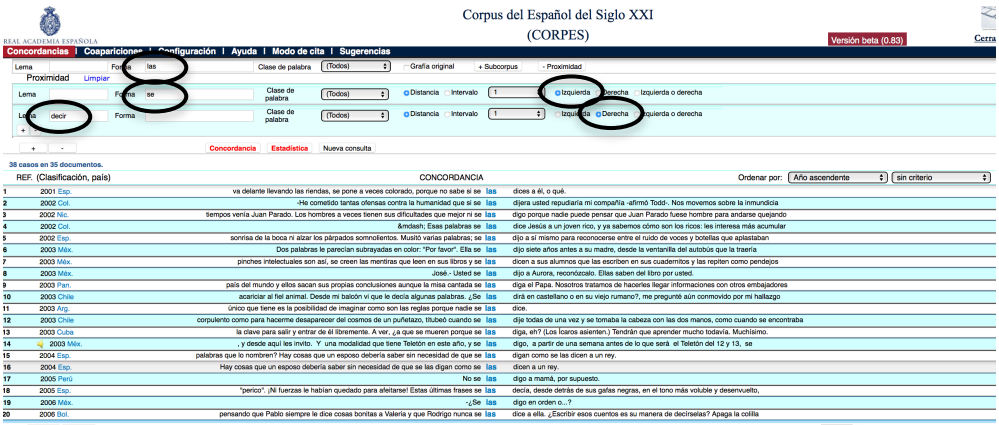


Figure 11. Search results of the structure *se las* in *CORPES XXI*.

(7) a. … Aurora sabía del *Platillo de Nudillos* mezclados,
Aurora knew of.the stew of knuckles mixed
porque usted *se la dijo* a la madre Pilar
because you 3.DAT 3F.SG.ACC told to the mother Pilar
en medio de la confesión; como tantas otras …
in middle of the confession like so.many others
'… Aurora knew about *The Mixed Knuckles Stew* because you told it to Mother Pilar in the middle of the confession; like so many others …' (Leñero Franco, Estela: *El Codex Romanoff*, 2005, Mexico)

b. Y una modalidad que tiene Teletón en este año, y
and a modality that has Teletón in this year and
se las digo, […] se podrá hacer la
3.DAT 3F.PL.ACC tell IMP can.FUT.3SG do.INF the
recaudación.

collection

'And a modality that Teletón has this year, and I tell you, the collection will be possible.' (*Primera emisión: grabación en directo, 11/11/03, Imagen Informativa.* Mexico, Oral)

In both cases, the accusative pronoun agrees in gender and number with the referent of the dative pronoun. In (7a) the accusative clitic should be *lo* (it refers to a previous statement) but it surfaces as *la* (feminine and singular), showing agreement with the indirect object *a la madre de Pilar*; in (7b) the referent of the direct object might be a nominal element (*el libro* 'the book') or the sentence *se podrá hacer la recaudación*, but the clitic *las* is agreeing with the female audience attending the television program.

3.3. Processing the results

Until now we have described the search procedures of *CORPES XXI* and we have emphasized their utility to obtain syntactic data. The corpus offers two more basic tools to carry out a first treatment of the results: different sorting options and statistics.

3.3.1. Sorting the results

Although sorting the examples may not seem relevant at first, the level of development that this function has in *CORPES XXI* makes it a very useful resource from the standpoint of syntactic variation. In this sense, the corpus allows basic orderings, such as grouping the results by year (ascending or descending order), country, linguistic area, author or title. However, what makes this corpus interesting for the study of syntactic variation is that it allows ordering the examples by the elements that appear before or after the searched form, lemma or grammatical category (see Figure 12). Besides, the corpus also offers the possibility of sorting the examples by "pivot" (as much for form as for lemma or for grammatical category), which allows grouping together the examples that share the same form, the same lemma or the same grammatical category (see Figure 12).

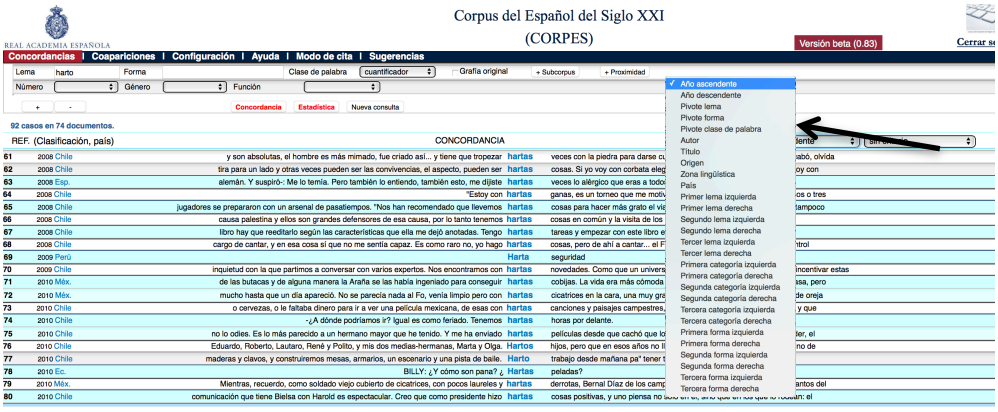


Figure 12. Sorting possibilities in *CORPES XXI*

These ordering possibilities are helpful for the analysis of the data. For example, in the case of the use of *harto* 'full' as a quantifier meaning 'a lot',

we can sort the examples by considering the grammatical category of the element to the right of the quantifier, so that we group all the combinatorial possibilities according to the grammatical category of the element with which it co-appears. This allows the researcher to provide closer analyses of, for example, the type of nouns ([±count], [–count] used as [+count] in plural or singular) combined with the quantifier (see Figure 13).

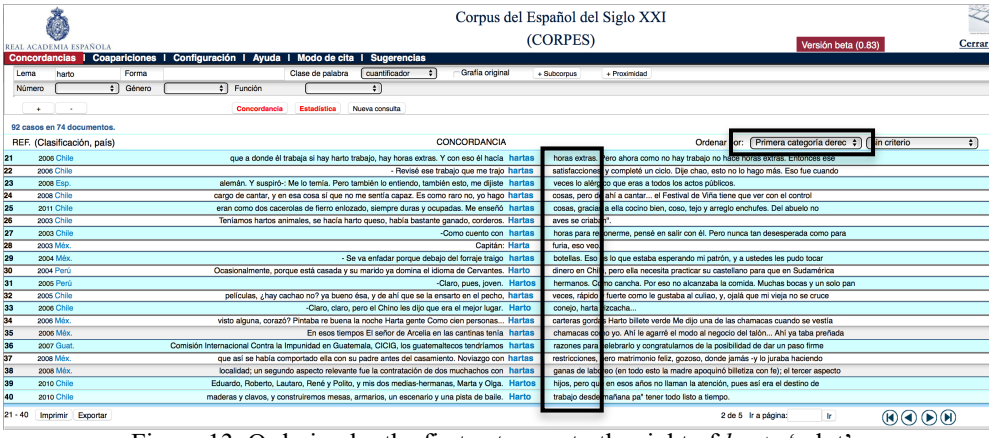


Figure 13. Ordering by the first category to the right of *harto* 'a lot'

In the same way, this classification is also helpful in the analysis of constructions in which *ello* 'it' functions as a direct object. Hence, the output of the searches can be ordered by pivot on the form and, next, the different verbs that combine with *ello* 'it' as a direct object can be grouped together, which may be helpful in the syntactic analysis (see Figure 14).

Figure 14. Ordering by pivot on the form with the examples containing *ello* as a direct object

3.3.2. Statistics

CORPES XXI allows obtaining statistics of the search results. The corpus offers statistical data of the searched phenomena with regard to five parameters: country, area, period, topic and textual typology. Furthermore, the corpus shows both the absolute frequency —that is, the total number of cases that can be found within the corpus plus the total number of documents containing this type of cases— and the standardized frequency for each million of words. By offering standardized frequencies, we have a more realistic idea of the representativeness of this phenomenon in the corpus. This fact is another strength of this corpus, since the frequencies often depend on a corpus being balanced in all its parameters (see Rojo, 2014; Rojo, 2016b).

The corpus offers these two frequencies in the total of the obtained results (see Figure 15, number 1) and it also shows these frequencies according to the five parameters pointed out (which is shown through tables and graphics) (see Figure 15). In addition, researchers can sort these frequencies from highest to lowest according to their interests.

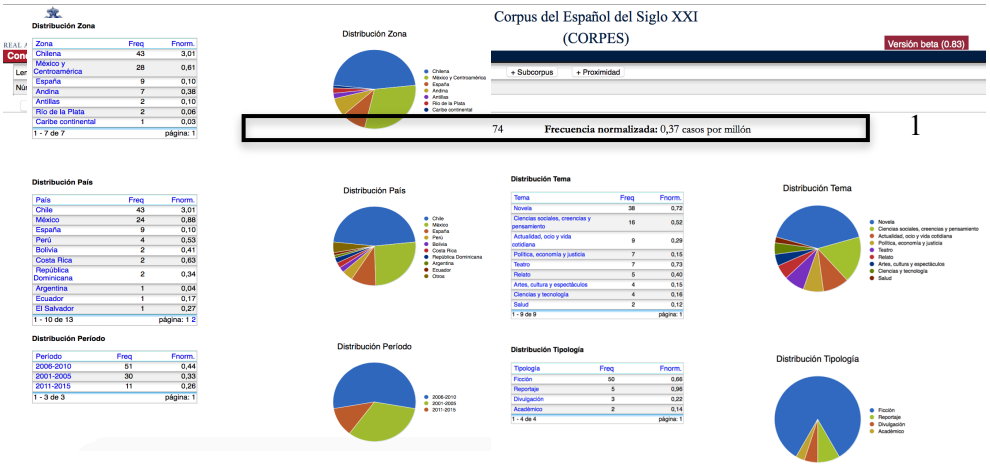


Figure 15. Statistics of the quantifier *harto* in the *CORPES XXI*.

Nevertheless, *CORPES XXI* is a corpus "under construction and, therefore, with imbalances and misalignments that will gradually disappear in later versions", as pointed out in its presentation in the RAE's website. For example, not all texts incorporate textual typology and there is still a lack of texts from some countries, topics and years to cover the forms envisaged according to these parameters. In addition, we consider necessary to note that a corpus is a tool for managing large amounts of information, but the researcher's intervention in the interpretation of the data is unavoidable,

since he/she is who must assess the relevance of the information for his/her research. Despite this, the corpus, which consists of almost 312 million forms, is a tool characterized by its flexibility, since the researcher can adapt it according to his/her interests and, therefore, create a customized corpus.

Regardless of the fact that the corpus is in elaboration process, the possibility of performing a statistical treatment of the results can greatly facilitate the job of the researcher in the interpretation of the data. For instance, the statistical tool of *CORPES XXI* can be used to know the geographical distribution of a syntactic phenomenon like the use of expletive *mismo* in explicative relative sentences. If we look at the absolute frequency of this phenomenon, we can see that the area in which it is more attested is Mexico and Central America (90,7% of the cases), followed at a great distance by the Andean and Continental Caribbean areas (both with a 2,4% of the cases). These data agree with the considerations of the *NGLE* (2009, 13.11n) regarding the extent of this phenomenon, except for the case of Continental Caribbean, since the *NGLE* did not indicate its use in this area. Given that the statistics also allows obtaining the distribution of this phenomenon by countries, one can observe its high frequency in Mexico (76,3% of the cases), and, then, it could be argued that this is a characteristic phenomenon of this country. The rest of the countries are very far from Mexico in the statistics (Honduras, with a 3,8%; Guatemala, with a 3,5% or Nicaragua, with a 2,8%).

The standardized frequency corroborates these considerations about the extent of the phenomenon: the area of Mexico and Central America, which, as already pointed out, had a higher absolute frequency, have the highest relative frequency too. However, in some instances this is not the case. This is so because standardized frequency better determines the representativeness of any phenomenon in the corpus, since it takes into account the number of forms assigned to each parameter (area, year, country, topic). That is, it is possible that a phenomenon has a low presence in absolute terms, but if the number of forms assigned to that parameter is scarce, then the representativeness of the phenomenon increases. In the case of the expletive use of *mismo* in explicative relative sentences, in USA there are only 8 cases in absolute data (1%); however, the representativeness of these cases is important because this country is assigned very few forms in the whole corpus. Hence, its standardized frequency is higher (2,51) than in areas like the Andean or the Continental Caribbean ones, which are in the second and the third place with regard to their absolute frequency. Taking into account this issue in relation to frequencies, it is important to note that this tool allows us to verify that this construction is frequent in Mexico and in Central America (as pointed out by the *NGLE*), especially in México, and that, by linguistic contact, it is spreading to the Spanish of the USA. On the other hand, the results show its scarce use in other areas such as the Andean area or the Continental Caribbean zone (see Figure 16).

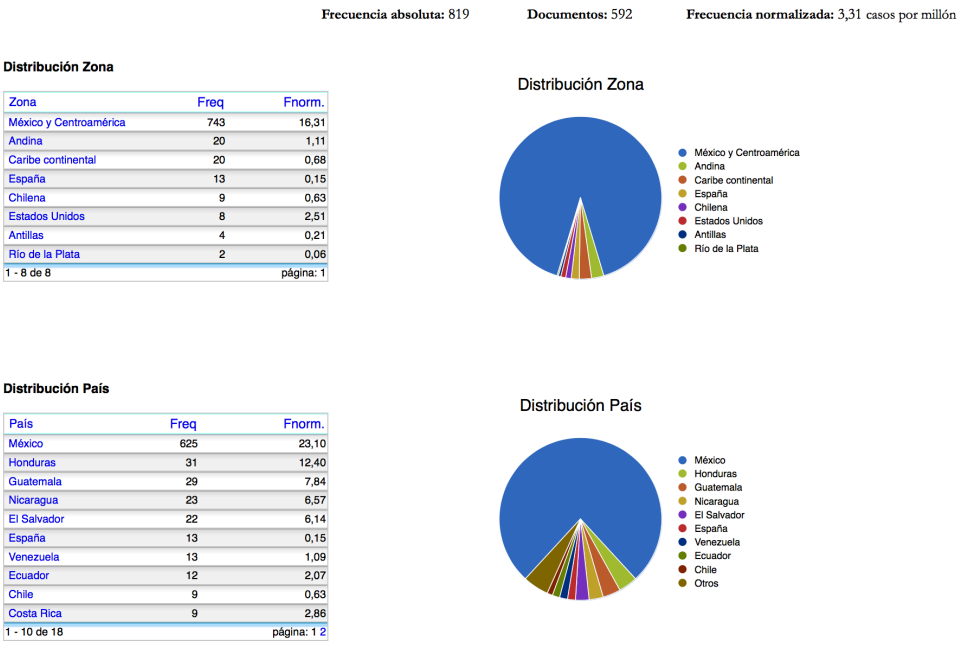


Figure 16. Statistics of expletive *mismo* in the explicative relative sentences in *CORPES XXI*.

It is also interesting to compare the statistics of two related constructions in order to observe the distribution among variants. For example, if we compare number disagreement with gender disagreement in accusative pronouns, we observe that both the absolute and the standardized frequency of number disagreement (absolute frequency: 152 cases in 132 documents; standardized frequency: 0,61) is much higher than that of gender disagreement (absolute frequency: 38 cases in 35 documents; standardized frequency: 0,15). These data lead to the conclusion that gender disagreement in accusative pronouns is scarce in current Spanish. Besides, when comparing the two variants on the basis of the statistical data offered by the corpus, one can also infer some conclusions from a diatopic

perspective. For instance, number disagreement has a low frequency in the Spanish spoken in Spain (4 cases of 132, with a standardized frequency of 0,04), which contrasts with the usual attestation of this phenomenon in American Spanish, especially in Mexico and Central America (67 cases of 132, with a standardized frequency of 1,47), Continental Caribbean (29 cases of 132, with a standardized frequency of 1,00) and River Plate area (21 cases of 132, with a standardized frequency of 0,64). By contrast, gender disagreement —despite being scarcely attested in general terms— is attested with the same degree of vitality both in Spain and in Mexico and Central America (10 cases of 38, with a standardized frequency of 0,21), even though in Chile or River Plate it shows more standardized frequency or the same (0,28 and 0,21, respectively) (see Figure 17).

se los

Distribución Zona

Zona	Freq	Fnorm.
Chilena	4	0,28
México y Centroamérica	10	0,21
Río de la Plata	7	0,21
Antillas	3	0,16
España	10	0,11
Andina	2	0,11
Caribe continental	2	0,06
1 - 7 de 7		página: 1

se las

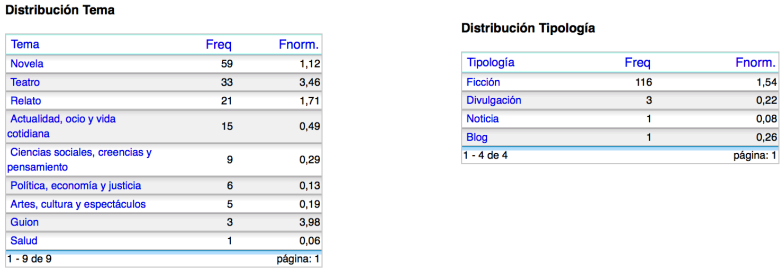
Distribución Zona

Zona	Freq	Fnorm.
México y Centroamérica	67	1,47
Caribe continental	29	1,00
Río de la Plata	21	0,64
Chilena	14	0,98
Antillas	8	0,42
Andina	7	0,38
España	4	0,04
Estados Unidos	2	0,62
1 - 8 de 8		página: 1

Figure 17. Statistics of the use of *se los/se las* in several linguistic areas in *CORPES XXI*.

As we have previously pointed out, the statistics offered by *CORPES XXI* also account for the distribution of the searched phenomenon according to the topic and the typology of the texts in which it appears. In this sense, one can check if a given phenomenon is related to a particular topic or a

particular type of text. For example, in the case of number disagreement in the direct object, it is possible to notice that the textual typology plays a role in the use of these structures, since fiction presents 1.54 cases by million. Regarding this, theatre, a clear example of orality, presents 3.46 cases by million of this phenomenon, followed by short story (1.71) and novel (1.12) (see Figure 18).



Distribución Tema

Tema	Freq	Fnorm.
Novela	59	1,12
Teatro	33	3,46
Relato	21	1,71
Actualidad, ocio y vida cotidiana	15	0,49
Ciencias sociales, creencias y pensamiento	9	0,29
Política, economía y justicia	6	0,13
Artes, cultura y espectáculos	5	0,19
Guion	3	3,98
Salud	1	0,06
1 - 9 de 9		página: 1

Distribución Tipología

Tipología	Freq	Fnorm.
Ficción	116	1,54
Divulgación	3	0,22
Noticia	1	0,08
Blog	1	0,26
1 - 4 de 4		página: 1

Figure 18. Statistics of the use of the construction *se los* according to the typology and the topic of the texts in *CORPES XXI*.

Taking this fact into account, it is possible to relate these typologies to a particular degree of formality. Hence, fiction texts (novels, short stories, theatre) could correspond to a high degree of formality. However, these correspondences could lead to wrong conclusions because fiction texts can contain fragments with a low degree of formality. For instance, the combination of synthetic and analytic comparatives is more attested in fiction texts, with a standardized frequency of 0,26, especially in short stories (0,73), theatre (0,41) and novels (0,13). Nevertheless, it is necessary to point out that most of the examples found, despite their typology, contain

fragments close to orality, such as speeches of characters in novels and narratives, or the expression of personal opinions in the case of journalistic texts (see examples in 3 above).

Therefore, in the absence of a classification of the texts according to their formality, the statistics by typology and by topic must be complemented with the analysis of the context in which the examples appear, which will lead the researcher to determine whether formality is a relevant factor in the use of each particular phenomenon.

4. Conclusions

In this chapter we showed the suitability of corpora to do research on syntactic variation. Based on the results obtained in this study, we have shown the great possibilities of *CORPES XXI* in the analysis of linguistic variation in current Spanish. First, the lemmatization and grammatical categorization of words makes it possible to search for morphological and syntactic phenomena and their analysis from a diatopic point of view. The searches allowed by the corpus are appropriate to find data about syntactic phenomena, especially the search for proximity. Besides, the possibility of combining searches increases the advantages of this corpus from the standpoint of syntactic variation. As we have illustrated, one of the weaknesses of the corpus is that it does not offer a classification of the

documents from a diastratic and a diaphasic perspective, and hence it is the researcher who must interpret whether these variables are significant or not for the analysis of the text itself. In the same way, searches or classifications by syntactic functions cannot be performed either, which would be desirable in order to improve the exploitation of the corpus for syntactic purposes.

Therefore, *CORPES XXI* is an ideal tool to obtain syntactic variation data in Spanish if one knows all the possibilities we described. *CORPES XXI* is a reliable complement (or even a clear alternative) to the use of linguistic atlases in the study of syntactic variation in Spanish, given that linguistic atlases usually lack syntactic information and do not offer the great versatility of this corpus.

References

Bosque, I. (2015). Las categorías gramaticales (2ª ed.). Madrid: Síntesis.

Company Company, C. (1992). Un cambio en proceso: "el libro ¿quién *se los* prestó?". In E. Traill (Ed.), Scripta philologica in honorem Juan M. Lope Blanch (pp. 349-363). México: UNAM.

Company Company, C. (1998). The interplay between form and meaning in language change. Grammaticalization of cannibalistic datives in Spanish. *Studies in Language*, *22* (3), 529-565.

Demonte, V. (1999). El adjetivo: clases y usos. La posición del adjetivo en el sintagma nominal. In I. Bosque & V. Demonte (Eds.), Gramática descriptiva de la lengua española (pp. 129-216). Madrid: Espasa Calpe.

Espinosa, R. Mª. (2014). Adverbios de cantidad, foco, polaridad y modalidad. In C. Company (Dir.), Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales (Vol. 1, pp. 939-1117). México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Fernández Soriano, O. (1999). El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos. In I. Bosque & V. Demonte (Eds.), Gramática descriptiva de la lengua española (pp. 1209-1273). Madrid: Espasa Calpe.

Hummel, M. (2014). Adjetivos adverbiales. In C. Company (Dir.), Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales (Vol. 1, pp. 615-733). México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Kany, Ch. (1963/1969). Sintaxis Hispanoamericana. Madrid: Gredos.

Koch, P., & Oesterreicher, W. (2007). *Lengua hablada en la Romania: español, francés, italiano*. Madrid: Gredos.

Mello, G. de (1992). *Se los* for *se lo* in the spoken cultured Spanish of eleven cities. *Hispanic Journal*, *13* (1), 165-179.

Pato, E. (2016). Cuestiones de gramaticalización: *harto, cierto,* adverbios en *-mente* y adverbio *y* en documentos colombianos del siglo XVI. *Cuadernos de la ALFAL*, *8*, 202-218.

PRESEEA (2017, July). Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Retrieved from http://preseea.linguas.net.

RAE. (2017, July). Corpus del Español del Siglo XXI (CORPES XXI). Retrieved from http://web.frl.es/CORPES/view/inicioExterno.view. Beta version 0.83 of June 2016.

RAE-ASALE. (2009). Nueva gramática de la lengua española. Madrid: Espasa.

Rivarola, J. L. (1985). *Se los* por *se lo*. *Lexis*, *9*, 239-242.

Rojo, G. (2014). Hispanic Corpus Linguistics. In M. Lacorte (Ed.), The Routledge Handbook of Hispanic Applied Linguistics (pp. 371-387). New York: Routledge.

Rojo, G. (2016a). Los corpus textuales del español. In J. Gutiérrez-Rexach (Ed.), Enciclopedia lingüística hispánica (pp. 285-296). New York: Routledge.

Rojo, G. (2016b). *Citius, maius, melius*: del CREA al CORPES XXI. In J. Kabatek (Ed.), Lingüística de corpus y lingüística histórica iberorrománica (pp. 197-212). Berlin-Boston: Walter de Gruyter.

Sánchez López, C. (2006). El grado de adjetivos y adverbios. Madrid: Arco/Libros.

Vigara Tauste, A. M.ª (2010). Gramática, "excepción", norma y uso: a propósito de la construcción *más mayor*. Aspectos sincrónicos y diacrónicos (I). *Revista de la Sociedad Española de Lingüística*, *40*(2), 123-140.