MENDIVE
REVISTA DE EDUCACIÓN

Review article

Theoretical and methodological design of the Corpora of Chinese Learners of Spanish

Diseño teórico y metodológico del Corpus de Aprendientes Chinos de Español

Desenho teórico e metodológico do Corpus de Chineses Aprendizes de Espanhol

Jialing Gou [1]
https://orcid.org/0000-0002-0303-974X
Diusbel Rodríguez Roque [2]
https://orcid.org/0009-0003-4224-8928
Lidia Ester Cuba Vega [1]
https://orcid.org/0000-0002-0405-5194

[1] University of Havana. Cuba.
wawaana0601@gmail.com , lidiacuba@fenhi.uh.cu
[2] Sichuan International Studies University. China
diusbelrr2010@gmail.com

Received: January 27, 2023
Accepted: May 25, 2023.

ABSTRACT
This paper deals with corpora of learners and is part of a doctoral research entitled Attenuation in the written expression and interaction of Chinese learners of Spanish, whose objective (to characterize the use of attenuation in the written production of these learners) and the lack of data to achieve it, leads to the design of a corpora of learners, a source of reliable empirical data for the analysis of the characteristics of the interlanguage of these learners. Starting from the idea that a corpus always responds to a previous design, this work pursues the theoretical and methodological design of a computerized corpora, representative of the oral and written variants of Spanish as a foreign and second language of Chinese learners, in order to have a database that allows us to carry out rigorous research on their interlanguage and identify the main difficulties in their learning, in order to design appropriate didactic strategies to solve them. The Dialectical-Materialist Method is assumed as the general method of sciences and philosophical support, and theoretical and empirical methods are used, namely, historical-logical, analysis-synthesis, systemic-structural, documentary analysis and modeling. The main results show the theoretical systematization on linguistics of corpora and corpora of learners and, from it, the theoretical and methodological design of a specialized, sample and representative, multilingual, mixed, open and transversal-oriented corpus; in addition, the conception of three stages for the creation of Corpora of Chinese Learners of Spanish: design, data collection and construction.

Keywords: linguistics of corpora; corpus of learners; Chinese learners; Spanish as a foreign/second language.

RESUMEN

Este trabajo versa sobre los corpus de aprendientes y forma parte de una investigación doctoral titulada La atenuación en la expresión e interacción escrita de estudiantes chinos de español, cuyo objetivo (caracterizar el empleo de la atenuación en la producción escrita de estos estudiantes) y la carencia de datos que permita cumplirlo, conduce al diseño un corpus de aprendientes, fuente de datos empíricos fiable para el análisis de las características de la interlengua de estos estudiantes. Partiendo de la idea de que un corpus siempre responde a un diseño previo, este trabajo persiguió como objetivo diseñar teórica y metodológicamente un corpus informatizado, representativo de las variantes oral y escrita del español como lengua extranjera y segunda de los estudiantes chinos, para contar con una base de datos que permita efectuar investigaciones rigurosas sobre su interlengua e identificar las principales dificultades en su aprendizaje, a fin de diseñar estrategias didácticas adecuadas para solucionarlas. Se asumió el Método Dialéctico-Materialista, como método general de las ciencias y sustento filosófico, y se emplearon métodos teóricos y empíricos, a saber, el histórico-lógico, el análisis-síntesis, el sistémico-estructural, el análisis documental y la modelación. Los resultados principales exhiben la sistematización teórica sobre lingüística del corpus y corpus de aprendientes y, a partir de ella, el diseño teórico y metodológico de un corpus especializado, muestral y representativo, multilingüe, mixto, abierto y de orientación transversal; además, la concepción de tres etapas para la creación de Corpus de Aprendientes de Chinos de Español: diseño, recolección de datos y construcción.

Palabras clave: lingüística de corpus; corpus de aprendientes; estudiantes chinos; español como lengua extranjera/segunda lengua.

RESUMO

Este trabalho trata do corpus de aprendizes e faz parte de uma pesquisa de doutorado intitulada Atenuação na expressão e interação escrita de estudantes chineses de espanhol, cujo objetivo (caracterizar o uso da atenuação na produção escrita desses estudantes) e a falta de de dados que permite a conformidade, leva ao desenho de um corpus de alunos, uma fonte de dados empíricos confiável para a análise das características interlínguas desses alunos. Partindo da ideia de que um corpus responde sempre a um desenho prévio, este trabalho teve como objectivo conceber teórica e metodologicamente um corpus informatizado, representativo das variantes orais e escritas do espanhol como língua estrangeira e segunda de estudantes chineses, para ter uma base de dados que permite pesquisar rigorosamente a sua interlíngua e identificar as principais dificuldades na sua aprendizagem, a fim de desenhar estratégias didáticas adequadas para resolvê-las. Assumiu-se o Método Dialético-Materialista, como método geral de ciência e sustentação filosófica, e foram utilizados métodos teóricos e empíricos, nomeadamente, histórico-lógico, análise-síntese, sistêmico-estrutural, análise documental e modelagem. Os principais resultados expõem a sistematização teórica sobre linguística do corpus e corpus de aprendizes e, a partir dela, o desenho teórico e metodológico de um corpus especializado, amostral e representativo, multilíngue, misto, aberto e de orientação cruzada; Além disso, a concepção de três etapas para a criação de Corpus de Aprendizes de Chinês de Espanhol: concepção, coleta de dados e construção.

Palavras-chave: linguística de corpus; corpus de aprendizes; estudantes chineses; Espanhol como língua estrangeira/segunda língua.

INTRODUCTION

In the area of general linguistics and applied linguistics in particular, efforts are made today to work with real and significant data that enable the most faithful possible analysis of the object of study. This entails the collection, organization and adequate coding of extensive samples of the linguistic reality that is intended to be investigated. In this sense, Information and Communications Technologies (ICTs) have considerably facilitated the construction of corpora to the extent that they have made possible the collection and organization of texts in electronic format, which also allows their analysis. , coding and annotation through applications developed for these purposes and, consequently, favors the recovery of statistical information and language samples through a computer search tool. As Rojo (2022) points out, "the spread of computers, as well as the generalization of the use of the Internet, have meant an authentic instrumental revolution in linguistics and corpus linguistics is probably its clearest exponent" (p. 73) .

Corpora are characterized by being composed of a voluminous amount of real data representative of a language, linguistic variety, dialect, etc. or the interlanguage of students of a specific foreign language (FL) or second language (L2), which are stored in electronic format and are coded so that they can be analyzed scientifically, so that empirical results can be reached that allow phenomena to be studied in depth. and linguistic elements in order to better understand language and languages, and establish relationships between linguistic theories and realities. But, to build a corpus that facilitates the recovery of homogeneous and quantifiable data that favors the development of empirical theories about a specific object of study, an adequate design is essential.

This article starts from Rojo's (2021) statement that "all the aspects necessary to understand what a corpus is [...] are synthesized in the idea that they always respond to a prior design" (p. 81). And, precisely, in these pages a theoretical systematization is carried out on what is understood today by corpus linguistics, linguistic corpus and learner corpus, as well as a review of the methodological principles that govern the constitution of an unjust corpus to substantiate from the theoretical and methodological point of view the construction of the Corpus of Chinese Learners of Spanish (CACHE). But why is the creation of this corpus necessary?

In recent years, a considerable number of corpora of ELE/L2 learners have been built. However, there are still few that include samples of Spanish interlanguage from Chinese informants. This can be seen in the corpus repositories of Spanish learners at the Catholic University of Leuven and the Complutense University of Madrid (UCM). A review of these repositories allows us to confirm that the corpora that include samples of the interlanguage of Chinese learners of Spanish are the Corpus Oral de Español como Lengua Extranjera (CORELE), FonoELE , DIAZ Corpus, Corpus de Interlengua Española de Aprendices Sinohablantes (SINEAS), the Written Corpus of Spanish as L2 (CEDEL2), the 台灣西語學習者語料庫(Corpus of Taiwanese Learners of Spanish, CATE) and the Corpus Learners of Spanish (CAES)

Of them, the first three (CORELE, FonoELE and DÍAZ Corpus) constitute monolingual oral corpora that collect samples of the interlanguage produced by Spanish students with different mother tongues. It should be noted that the design criteria of each one respond to the objectives with which they were created and, therefore, are different. For example, CORELE was collected for the study of error analysis of oral production and FonoELE for the development of research on the acquisition and learning of the phonic component in ELE. They all collect samples of the interlanguage of speakers of different mother tongues, but the samples of Chinese informants are a minority compared to those of speakers of European languages such as Portuguese, Italian and German.

The latter would not affect the conduct of research with samples of Chinese speakers if there were a representation of all levels of competence and if a homogeneous sample could be established in this sense, but CORELE is only composed of speeches from speakers accredited with the levels of A2 and B1 competition, FonoELE includes from level A2 to C1 and Diaz Corpus does not declare this metadata. Furthermore, as the design criteria and objectives of these corpora are different and only FonoELE documents and describes elements of the corpus design, offering relevant information about its construction, it is impossible to carry out contrastive studies with these computerized empirical databases.

Of the remaining corpora listed, SINEAS and CAES constitute written corpora. The SINEAS includes only samples of the Spanish interlanguage of Chinese informants, while the CAES incorporates that of students of various first languages. For its part, 台灣西語學習者語料庫 (CATE) only includes the language produced by Spanish students from Taiwan and CEDEL2 includes students of various foreign languages, including Mandarin Chinese. It should be noted that these last two corpora are mixed, that is, they contain oral and written texts; Furthermore, they are multilingual since they contain native corpora of the first languages of the students learning Spanish, a characteristic of great value for the development of contrastive studies.

The Spanish Interlanguage Corpus of Sino-speaking Learners (SINEAS) collects texts written by Chinese students studying Hispanic Philology at four Chinese universities and by students of Chinese nationality on mobility (in linguistic immersion) at three Spanish universities since the 2016 academic year. -2017 until the 2019-2020 academic year. The samples are from students with different levels of linguistic proficiency, from A1 to C1. The texts collected, in different textual types (description, narration, exposition and argumentation), were handwritten based on writing tasks as stimuli integrated into 43 tests, then they were digitized faithfully respecting the handwritten version, and have been annotated based on of a set of social and linguistic metadata that allows filtering searches. Statistically, it can be noted that CINEAS (version 1.0, December 2019) contains samples produced by Chinese students with a total of 4378 written texts 4339 (its current volume is 435,000 words). Only part of the CINEAS corpus can be accessed, access to the entirety is restricted. Only researchers from the University of Lleida project will be able to access all the data available in CINEAS. The general public can only make simple queries, without error tagging and other metadata, of less than 50% of the corpus.

The Corpus Apprentices of Spanish (CAES), in its current version (2.1, March 2022), includes 1,045,097 linguistic elements, produced by 2,544 students of different language levels (A1-C1) and different L1s (German, Arabic, Chinese Mandarin, French, Greek, English, Italian, Japanese, Polish, Portuguese and Russian), who wrote 2 to 3 texts each, comprising some 6561 writing tasks integrated into 2544 tests. This large corpus, with open access, has been built thanks to the collection of samples in different centers of the Cervantes Institute and universities in a large number of countries, in a period ranging from October 2011 when this project promoted and financed by the Cervantes Institute until December 2020.

Although, without a doubt, the CAES constitutes a large database of the Spanish interlanguage produced by Chinese students, since it currently includes 108,942 linguistic units distributed by different levels (A1, 19,751; A2, 23,292; B1 48 766; B2, 16,892; and C1, 241) produced by this ELE/El2 student body and is made up of a wide variety of texts (biographies, letters, emails, argumentative essays, narratives, notes/notices, postcards, etc.), we must point out at least three aspects that are considered negative when carrying out studies with this corpus.

Firstly, the writing tasks from which the interlanguage samples of this corpus were

collected imitate, to a large extent, the DELE tests, which, in the opinion of this author, given her experience in preparing students to take the DELE exams, and in the review of syllabi of the DELE Examiner Accreditation Courses carried out, limit the creativity of the students and only provide an interlanguage conditioned by pre-established and mandatory guidelines. On the other hand, as samples are collected in various countries, there would be ELE and EL2 students, as well as students of Spanish as a heritage language, etc. Thus, the failure to indicate the metadata referring to the immersion or non-immersion context where these students learn, or the time of contact with the Spanish language such as years of study or stays abroad, limits the development of contrastive and sociolinguistic studies based on of these variables. On the other hand, the level of instruction of the students and the programs in which they learn Spanish are unknown; it would be inappropriate to compare the Spanish interlanguage of a Chinese student studying Hispanic Philology and that of a student taking a 6-year Spanish course. months, even if both are accredited with an A1 level.

The 台灣西語學習者語料庫 (Corpus of Taiwanese Learners of Spanish, CATE), as shown by its two constituent subcorpuses, 台灣西語學習者書面語語料庫 (Written Corpus of Taiwanese Learners of Spanish, CEATE) and the 台灣西語學習者口語語料庫 (Oral Corpus of Taiwanese Learners of Spanish, COATE), is an oral and written corpus that collects the Spanish language produced by students from Taiwan. It constitutes a valuable database with some 34,000 words produced by ELE students in Taiwan, however, its access is restricted. Furthermore, two negative aspects can be pointed out regarding the research. Firstly, this corpus does not indicate the metadata referring to the language of origin and, although it can be assumed that the majority of the texts come from informants whose L1 is Mandarin Chinese, if the sociolinguistic situation of Taiwan is considered, the L1 could also be other such as Taiwanese (*min nan*) and Hakka (*kejia*). Furthermore, it can be noted that the representativeness of the interlanguage of these students is affected, since only essay texts are included and, as a consequence, the diversity of text types required by a corpus of learners does not continue.

The Written Corpus of Spanish as a L2 (CEDEL2) is a large linguistic corpus that contains texts produced by students of Spanish as a second/foreign language. Its new version (CEDEL2, version 2, 2021) currently has a total of 4,399 participants and 1,105,936 words (744,950 from ELE/L2 students and 360,986 from native speakers), making it one of the largest corpus of its guy. This corpus of learners contains data from Spanish learners with different L1 (students' mother tongue): English, German, Dutch, French, Portuguese, Italian, Greek, Russian, Japanese, Arabic and Chinese. Also, for comparative purposes, CEDEL2 contains subcorpuses of texts produced by native speakers of Spanish (Spanish and Latin American varieties), English, Japanese, Portuguese, Greek and Arabic. It does not yet contain samples from native Mandarin Chinese informants.

Although CEDEL2, unlike CINEAS and CATE, and following the *open data science philosophy ,* is available in its entirety and can be consulted online and downloaded freely for research and teaching purposes, it still does not collect an extensive and representative sample of texts produced by Spanish students whose mother tongue is Mandarian Chinese. When observing the statistics offered on its official website, it is found that, of the 4,399 informants, there are only 22 Chinese, 20 of them provided written texts and 2 participated in the collection of oral texts. These figures are insignificant compared to the data collected with informants from other L1s, for example, with English as L1, 1931 participated, with Japanese 243, with Greek 216, with Portuguese 164 and with Russian 101. Furthermore, the native corpus with linguistic data from Mandarin Chinese.

Furthermore, the interlanguage data of Chinese students collected in CEDEL2 is insufficient for rigorous studies. On the one

hand, it is not a representative sample of the interlanguage of these students, since there are no informants of all levels of competence and of the levels represented the number of informants is tiny. On the other hand, it should also be noted that of the 20 written texts collected, 17 belong to female informants and only 3 to men. Therefore, it is impossible to form a homogeneous sample that allows a study to be carried out taking into account social and linguistic factors that are fundamental for the purposes of this research. It can also be indicated that the 20 texts from Chinese informants were written based on the same task, which required the writing of a description-narrative, so there is no representativeness of textual genres either.

Until now, there is no corpus available, among the existing ELE learner corpora, that serves to deeply and rigorously investigate the Spanish interlanguage of Chinese students, taking into account certain linguistic and extralinguistic variables. CACHE comes to fill that gap.

The strength that the learning of Spanish as a foreign language is gaining in China, the lack of depth in some research on the interlanguage developed by Chinese students - largely due to the fact that the empirical data used for the analysis are limited and the collection instruments of data are inadequate - and, above all, the lack of a corpus for the study of the interlanguage of Spanish students that offers an extensive sample of oral and written texts produced by students of Chinese origin who speak Mandarin and with metadata labels that indicate precise information on the social and linguistic factors of the informants that allow truly generalizable conclusions to be drawn in a rigorous manner, justify the design and construction of the CACHE, which, in addition, must be open access.

General objective: To theoretically and methodologically design a computerized corpus, representative of the oral and written variants of Spanish as a foreign and second language of Chinese learners, to have a database that allows rigorous research to be carried out on their interlanguage and to identify the main difficulties in their learning, in order to design appropriate linguistic strategies for their solution.

DEVELOPMENT

Corpus linguistics, a new theory, a linguistic (sub)discipline or a new methodology?

«Linguistics is an empirical cultural science [...] whose object of study is language and languages» (Rojo, 2021, p. 35). Although this consideration is recent, today it is widely accepted. However, the theoretical deepening of corpus linguistics (CL), a young reality (six decades of history), reveals terminological issues on which conceptual ambiguities, unresolved terminological vaguenesses coexist explicitly or implicitly and, consequently, , the convergence of multiple opposing perspectives.

Precisely, the first terminological problem lies in defining what LC is. Are we facing a new theory? Does it constitute a (sub)discipline or is it, rather, a new methodology? There has been much debate to answer these questions; However, there is still not much consensus in the literature. An example of this is seen in the divergence of criteria between the authors who reflect on the theoretical issues of LC. The most accurate answer seems to be the one offered by Rojo (2021), who states that:

> Indeed, LC is not a theory: the data from a corpus can be analyzed from very different theoretical frameworks […]. Nor does it seem that it can be considered as a linguistic (sub)discipline like morphology, syntax, sociolinguistics, etc.: the use of corpora occurs in grammatical, historical,

> sociolinguistic, lexicographical and many other specialized fields. Finally, it is not easy to view it as a methodology in the strictest sense of the word. (Red, 2021, p.47)

Rojo (2021) himself, referring to the difficulty of considering LC as a methodology, points out that « Leech (1992, p. 106) considers that LC is "a new research enterprise , and in fact a new philosophy approaches to the subject "» (p. 47) and later emphasizes that «more recently, Leech (2011, p. 158) has insisted on considering that the LC " is not a pure and simple methodology , but is more like a methodology than a scientific domain "» (p.47). We can therefore observe the complexity that authors such as Leech have noted when defining what CL is and the ambiguity of their proposals.

According to Rojo (2021), a clear and operational idea of the character of LC is to consider that it is an approach to the study of linguistic facts that is empirically oriented and based on the detailed analysis of a large amount of data (the corpora). which makes clear his opposition to both rationalist and traditional descriptive linguistics […]. In effect, LC constitutes a form of approach to the study of linguistic phenomena and elements based on certain assumptions about which aspects of the analysis are really relevant. (p. 48)

Rojo (2021), taking as reference the proposals of Leech (1992), Biber , Conrad and Reppen (1998), Tognini -Bonelli (2001), Gries (2006, 2009) and Bolaños (2015), considers that LC is characterized mainly due to the following features:

- Be empirical, focus on the analysis of real usage patterns in naturally produced texts. Therefore, LC is more interested in performance than in competence and in describing what is found in languages rather than in linguistic universals.
- Use large textual corpora as the basis of the analysis. To the extent that these corpora are well constructed and representative, what is found in the sample can be projected to the population, that is, to the language. The objective, therefore, is not only to describe and explain what is found in the corpus, but also everything that can be seen in the language or linguistic variety from which it has been extracted.
- The data come from texts produced naturally, which allows the data obtained to be related to the variations due to the different registers and types of text, a relationship that is not possible in the data obtained through experimental designs . Naturally, the variations can be located on the diachronic, diatopic and diastratic lines.
- Make heavy use of computers to carry out at least part of that analysis. The use of computational resources in the construction and exploitation of corpora is a necessity derived from their volume, since only in this way is it possible to analyze data sets of a size that would be impossible without these resources.
- Use quantitative and qualitative analytical techniques. It is important to note that textual corpora are the most convenient and appropriate resource to study everything related to the frequency of linguistic phenomena and elements.
- Carry out (or, at least, pretend to carry out) systematic and exhaustive analyzes of all the relevant cases located in the corpus of what you intend to study. That is, the corpus is not treated simply as a kind of database from which a few cases are extracted and others rejected, but rather the entire corpus is taken into consideration.

Therefore, summarizing Rojo's (2021) position, it can be concluded that LC is an empirical approach that analyzes real data - « what people actually say and write » (

Aarts , 2002, p. 4) -, representative of the linguistic acts that occurred in a linguistic community, with the idea of understanding the system that has made them possible. And, the way to achieve this is by examining the texts or fragments of texts, oral or written, contained in representative corpora, whose exhaustiveness requires their computerized construction, to work with computers for the selective extraction and recovery of information and, finally, perform statistical processing of these large masses of information.

Likewise, Rojo (2021) considers that the LC methodology also meets the characteristics of the "empirical cycle" of Krug, Schlüter and Rosenbach (2013):

> Firstly, objectivity, which means that the data used must be completely independent of the people carrying out the research and the tools used to obtain it. Secondly, reliability and replicability, which guarantee that the data obtained will be the same in extractions carried out at different times. Finally, the relevance of the data used for the phenomenon analyzed. (2021, p. 50)

So, it must be considered that LC is an empirical orientation that deals with the analysis of real and objective data gathered in linguistic corpora, using new technologies and computer programs. It thus constitutes an empirical orientation in which "the quantitative growth of knowledge about the behavior of languages and speakers has given rise to significant qualitative growth" (Rojo, 2021, p. 50).

What is a linguistic corpus?

In general, assuming an operational definition as a theoretical framework for research is a complex task. And, of course, the heterogeneity of approaches when defining what is understood as a corpus within the LC is an example of this complexity. An in-depth review of the specialized bibliography makes this critical systematization possible with the intention of assuming that which exposes the most relevant characteristics of a corpus within the current LC or to be able to construct a rigorous definition for the present study.

After consulting the definitions of linguistic corpus offered by Francis, Kuèera and Mackie (1982), Sinclair (1991, 2005), Leech (1992), Biber (1993), Torruella and Llisterri (1999), McEnery , Xiao and Tono (2006) , Parodi (2008), Villayandre (2008), Hincapié (2018), Hincapié and Bernal (2018), Hincapié and Rubio (2018), Lemnitzer and Zinsmeister (2008) and Rojo (2021), the definition offered by this The last author includes, in summary, the most characteristic features of a linguistic corpus in current CL. According to Rojo (2021), a corpus is:

> … a set of (fragments of) texts, oral or written, produced under natural conditions, jointly representative of a language or a linguistic variety, in its entirety or in some of its components, which are stored in electronic format and They code with the intention that they can be analyzed scientifically. (p.1)

Rojo (2021) himself reflects on the aspects included in this definition. Their analysis is discussed below.

According to this linguist, "the texts that make up the corpus must have been produced in natural situations" (2021, p. 1). He then explains that, before including the texts or fragments of texts in the corpus, they must have been created as a literary work, a journalistic text, an epistolary text (in the case of written texts) or as a conversation, a conference , a speech (if they are oral texts). In this way, Rojo emphasizes that these are texts constructed with real communicative intention and not conceived to illustrate a certain linguistic phenomenon.

The other aspect that Rojo (2021) analyzes is that of representativeness. According to him, the texts or fragments of texts contained in the corpus must be "jointly representative of a language or a linguistic variety at a certain moment in its history or throughout a certain period" (p. 1). But, aware of the complexity contained in the idea of representativeness, he explains that at a minimum "the set of texts integrated into a corpus must give an adequate vision of what it aims to represent" (p. 1).

It also analyzes the computational character or nature of the corpus. Rojo (2021) suggests that due to the size of the corpora, in order to recover the information necessary for its study, it is unavoidable to transfer the texts to electronic format. And he states: "although conceptually it may be thought that the electronic format is not a constitutive feature of the definition of corpora, the reality is that they can only be handled if they have this character" (p. 2).

On the other hand, coding is important. In this regard, Rojo (2021) points out that "the texts that form a corpus must be coded so that it is possible to achieve selective retrieval of information" (p. 2). It cannot be lost sight of that selectivity is one of the methodological procedures of the LC. For this reason, texts are encoded: metadata must be added, that is, paralinguistic information associated with the text. Coding, that is, the inclusion of metadata in each of the texts that make up a corpus, "makes possible its scientific study and, more specifically, the selective recovery of the information it contains" (p. 2) to the extent in which this selective data extraction can be carried out using a query application designed for such purposes.

Coding also involves linguistic annotation. Rojo (2021) states that the texts that make up a corpus are also usually subjected to linguistic annotation processes, that is, "a series of information referring to their lexical and grammatical characteristics" is usually added (p. 2), because "The scientific analysis of a corpus considerably increases its possibilities if the texts that make up it have also been subjected to a process of linguistic annotation" (p. 2).

After this theoretical systematization of corpus definitions, there is one last consideration that is worth adding. A corpus, as Rojo states "is made up of texts, but it is much more than a simple aggregate of texts." In the same terms as this author, "the key word is design" (2021, p. 3). Consequently, "the construction of a corpus involves the systematic integration of texts according to a specific design" (Rojo, 2021, pp. 23-24), which means that "each corpus has the general configuration that corresponds to the objectives." with which it is built» (Rojo, 2021, p. 3). This is one of the reasons why there are different types of corpora and each one of them suitable for its specific purpose.

Classification of linguistic corpora

As concluded in the previous subsection, the construction of a corpus responds to a specific design according to the purpose for which it is built. Hence, there is a variety of corpora that could be classified according to different perspectives. Among the most notable corpus classification typology proposals in the specialized bibliography, those of Torruella and Llisterri (1999), Procházková (2006), Villayandre (2008), Cruz (2017), Hincapié and Bernal (2018), Hincapié and Rubio (2018) and Rojo (2021).

After reflecting on the typologization of these authors, their classification criteria and the types of corpora that they include within these categories, it is observed that some classification criteria are very general and, consequently, include a great variety of types of corpora that do not They are neither comparable nor opposite; Furthermore, some of the corpora mentioned constitute a subtype or variant of others listed in the same list.

On the other hand, other classification criteria are imprecise because, for example, they seem to indicate issues related to size, but in reality they present types of corpora that differ rather by the

size of the selected sample and by their purpose, and it is the latter that is conditions the extension and not vice versa. It must also be added that some of the classification labels could be misleading and by themselves cannot present the reality they try to describe.

Based on what has been analyzed, it is considered pertinent to classify the corpora according to the different objectives for which their design is conceived, and depending on this, taking into account another series of properties such as the types of texts they include and their characteristics, the number of languages, the degree of added information, etc. Precisely Rojo (2021) proposes a corpus typology taking these assumptions into account. According to this author, corpora can be:

- *Reference or specialized corpus.* These classification labels refer to the purpose for which the corpus has been created, the main classification criterion. Reference corpora, also called general corpora, are those designed with the intention of offering a resource where linguistic phenomena and elements that occur in a given language can be analyzed; therefore, they attempt to cover an entire linguistic domain and can be composed of different subcorpuses . For their part, specialized corpora are those composed of the selection of texts with certain common characteristics and belonging to very varied fields. Learner corpora are also specialized , consisting of oral or written samples of students with different degrees of mastery of a foreign language.
- *Oral and/or written corpus.* Corpora may contain oral and/or written texts. These labels are obvious: written corpora are those that contain written productions, while oral corpora are made up of oral interventions (and their transcriptions). There are also so-called mixed corpora that contain texts of both kinds and more recently there is talk of multimodal ones.
- *Total or sample corpuses.* A total corpus, for example, is one that contains all the work of an author or literary movement, all the speeches given by a president, etc. But, as Rojo (2021) states, these corpora have restricted purposes, which is why sample corpora are more used within the LC, which are conceived as a supposedly representative sample of a certain language, variety, etc. For example, a corpus of contemporary Spanish, a variety of Spanish, Spanish from the press of a country and in a specific period, etc.
- *Closed or open corpuses.* A closed corpus is one that is planned with a certain size, and when the predetermined size has been reached, it is considered finished and its composition is no longer altered, unless some type of linguistic annotation is added or modified. . On the other hand, an open corpus does not start with an already established size, but is conceived to grow as the availability of texts makes it possible and exploitation applications allow it.
- *Monolingual or multilingual corpus.* This classification responds to the selection criterion of the languages of the texts that make up a corpus. Monolingual corpora contain texts that belong only to one language, while multilingual ones are made up of texts from more than one. It should be noted that the latter can be presented in two forms: parallel multilingual corpora and comparable multilingual corpora. The former contain "the same text" in two or more languages, that is, they are made up of aligned translations. On the contrary, comparable corpora are made up of texts also in two or more languages, but without being translations of each other.
- *Synchronous or diachronic corpuses.* If a corpus focuses on the general characteristics of a

language at a given time, it is synchronous in orientation; On the other hand, if it tries to reveal the existing variation in any of the axes, it is diachronic in orientation.

- *Encoded or unencrypted corpus.* Taking into account the information added to the texts that allow data to be recovered selectively, the corpora may or may not be coded. The coded ones, the most used corpora in the current LC, add extratextual information (place, date, type of text, etc.) consistent with the organization and design of the corpus. That is, a header with so-called metadata is added to the texts of these corpora. On the contrary, non-coded ones do not add them.
- *Annotated and unannotated corpuses.* Also according to the information added to the texts that favors data retrieval selectively, corpora can be classified as linguistically annotated and unannotated corpora. The first, the most used corpora in current LC, are analyzed at different degrees and levels: phonological, morphological, lexical, syntactic, semantic, pragmatic, etc. That is, the texts are subjected to a linguistic annotation process that considerably favors scientific analysis.

Characteristics of the corpora

This subsection aims to identify the most important characteristics of a linguistic corpus. However, it must be emphasized that any characterization established could be insufficient given the enormous number of purposes for which a corpus can be conceived, designed and built. In this sense, it should also be noted that the list of characteristics should not have a hierarchical nuance, nor should it be presented as a closed enumeration.

A large list of important characteristics of linguistic corpora can be inventoried; in fact, there are some very extensive ones, namely those offered by Torruella and Llisterri (1999), Bowker and Pearson (2002), Villayandre (2008), Parodi (2008).) and Rojo (2021). But, as Parodi (2008) states, "the description of a corpus lies importantly in the search for a specification of its prototypical characteristics" (p. 108), which are considered contained in the definition of linguistic corpus offered by Rojo (2021) that has already been cited and commented on in the body of this text and from which prototypical features can be broken down such as:

- be a set of texts or fragments of texts;
- be produced in natural conditions;
- be representative of a language or a linguistic variety, in its entirety or in some of its components;
- be stored in electronic format and
- be coded so that they can be analyzed scientifically.

Thus, although each linguistic corpus must meet those specific characteristics consistent with its specific objective and pre-established design, all must respect at least these prototypical features that define what is understood today as a corpus within the field of LC.

Corpus levels

Depending on its design, we can find different levels in a linguistic corpus. A review of some with open access allows us to identify at least three: corpus, subcorpus and components.

Precisely, Torruella and Llisterri (1999) present this classification of hierarchical levels. At a first level they place the **corpus** , as a set of language samples of any type that are taken as a model of a predetermined state or level of language, which once analyzed should allow for improving knowledge of the linguistic system of the language it represents. At a second level they place the **subcorpus**, understood as a static selection of texts, derived from a normally more general and complex corpus, which is divided into groups of more specific textual samples; but it can also be a dynamic selection of

texts from a growing corpus: a certain number of texts intended to increase some section of a general corpus. Finally, they locate the **components**, understood as a collection of samples of a corpus or a subcorpus, which respond to a very specific specific linguistic criterion and, consequently, reflect a specific type of language. Compared to the corpora and subcorpus that are heterogeneous, the components are very homogeneous.

Corpus of learners: towards a definition and characterization

Once LC was consolidated, in the 90s, learner corpora emerged (Granger, 2002), a type of specialized corpus according to the classification offered by Rojo (2021). A bibliographic review allows us to confirm that there is no standardization when it comes to actually defining what these corpora consist of, and the definitions that are found (Sinclair, 1991; Granger, 2002; Hincapié, 2018; Calero, Serrano and Gómez, 2020; Lozano, 2022) tend to omit features that are considered prototypical of corpora in general and even others that characterize those of learners in particular.

Rojo (2021) explains the purpose of learner corpora within his corpus typology. This author proposes that these corpora are made up of texts produced by students of a specific second (L2) or foreign language (FL, with different languages of origin and different degrees of mastery of the L2/FL. They are corpora that are built precisely to study the characteristics of the corresponding interlanguage and that, logically, cannot be considered representative of the L1 in question. (p. 25)

In other of his recently published works, Guillermo Rojo together with Ignacio Palacios dedicates a chapter to the corpora of learners of Spanish as L2. In this reference, Rojo and Palacios (2022) maintain that a learner corpus consists of the collection of texts formulated by learners of one (or several) L2 and coming from one (or several) L1, with the possibility of also indicating, among the traits incorporated in the coding, the level of knowledge of the L2, the type of general training, the years dedicated to the study of the L2, the country of origin, etc. (p.75)

Next, these authors state that, like all corpora, those of learners "incorporate the texts as they have been originally produced, with the minimum coding necessary for the adequate recovery of data" (Rojo and Palacios, 2022, p.75) . Finally, they add that from this minimum of coding, additional features of morphosyntactic annotation, concordances, transcription and sound alignment can be added if it is oral samples, error coding, etc.

However, although these definitions (Rojo, 2021; Rojo and Palacios, 2022) offer insights into what is distinctive about learner corpora, they do not present all of the prototypical features that a linguistic corpus should manifest as stated in the previous section. Therefore, based on the general definition of linguistic corpus formulated by Rojo (2021) commented in this same text and, taking into account the distinctive features of learner corpora that authors such as Granger (2002), Hincapié (2018), Calero, Serrano and Gómez (2020), Rojo (2021) and Rojo and Palacios (2022), the following definition is proposed:

> A learner corpus is a specialized corpus, made up of a set of texts, oral or written, that are real, which implies that they come from linguistic acts actually performed and produced under natural conditions by students of a specific second language (L2) or foreign language (FL), with one or different mother tongues (L1) and with different degrees of mastery of the L2/FL, which is stored in electronic format and encoded with metadata and linguistic annotations in order to be able to study scientifically and with a empirical orientation the

characteristics of the corresponding interlanguage that are manifested throughout the learning process.

Corpus of Chinese Learners of Spanish: design, collection and construction

Objective: To build a computerized corpus, representative of the oral and written variants of Spanish as a foreign and second language of Chinese learners, to have a database that allows rigorous research to be carried out on its interlanguage and to identify the main difficulties in their learning, in order to design appropriate linguistic and educational strategies for their solution.

Stages in building CACHE

For the construction of CACHE, the three stages have been established, summarized in Scheme 1 (Figure 1), and specific actions have been planned for each of them. They are described below.

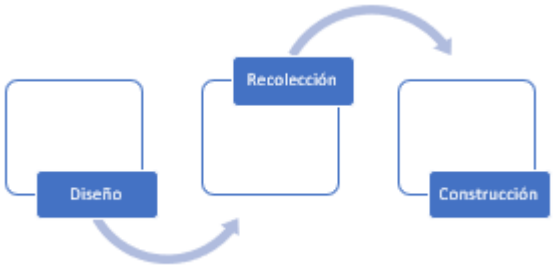

Fig. 1- Planned stages for the construction of CACHE

Design stage: This first stage focuses on design and, therefore, corresponds to the period prior to data collection. In it, the following actions have been established:

1) Critically review the specialized bibliography with a view to systematizing the theoretical and methodological references that underpin the design of a corpus.

2) Build a macro design of the learner corpus taking into account the systematized theory.

3) Assume the theoretical and methodological principles on which the design of the corpus to be built is based.

4) Determine linguistic and extralinguistic metadata that will be used in the coding and linguistic annotation of the language samples that favor the construction of a corpus representative of the Spanish interlanguage of Chinese learners.

5) Create and validate the oral and written expression and interaction tasks that will be used as stimuli for data collection, taking into account the general objective of the corpus and the descriptors established for these skills in the Common European Framework of Reference and the Plan Curriculum of the Cervantes Institute.

6) Design and validate an appropriate computer instrument for collecting linguistic and extralinguistic metadata and obtaining oral and written language samples, which will be included in the corpus.

7) Determine the computer linguistic tool that will be used for linguistic and extralinguistic coding and linguistic annotation.

8) Determine and design the computer system with which the data will be managed and which will constitute the means by which the corpus can be freely accessed online.

Data collection stage: This stage corresponds to the period in which the data is collected. In it, the following action has basically been contemplated:

1) Collect, based on the tasks and instruments created and validated in the design stage, the linguistic samples of Spanish produced by Chinese learners, taking into account the agreed objectives, selection criteria and linguistic and extralinguistic metadata.

Construction stage: This stage corresponds to the period in which all the language samples obtained are systematized and the construction of the corpus is completed. It contemplates the following actions:

1) Select for inclusion in the corpus those language samples that meet the agreed selection criteria and discard the remaining ones.

2) Carry out, automatically and manually, the coding and annotation process of the samples approved for inclusion in the corpus, based on the metadata and agreed labels.

3) Build the computer tool for searching the corpus, a Web, that allows the recovery of statistical and linguistic information; and publish it with open access for use by the scientific community.

4) Publish the corpus, according to the determined route, with open access for use by the academic and scientific community.

Execution of the design stage

After completing the first task of the design stage (the critical review of the theory and methodology related to the design of the corpora), the execution of the remaining actions planned for this period continued. Next, the most significant aspects of the CACHE design, resulting from the fulfillment of these tasks, are presented.

Macro Design Overview

The corpus of learners that is intended to be built will be called the Corpus of Chinese Learners of Spanish (CACHE). Once built, it will be a specialized, sample and representative corpus, multilingual - it will include a native control corpus in Spanish and Mandarin Chinese -, mixed - it will include oral and written samples -, and open; In addition, it will be transversally oriented.

As seen in Scheme 2 (Figure 2), CACHE will be made up of four subcorpuses : two interlanguage subcorpuses of Spanish learners and the two control subcorpuses in Spanish and Mandarin Chinese.

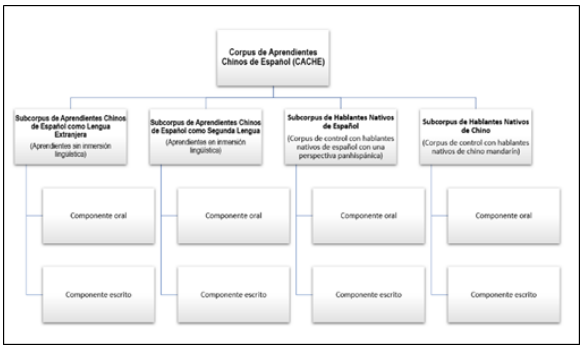

Fig. 2- CACHE macro design

The core of the corpus is made up of the two subcorpus of learners: the Subcorpus of Chinese Learners of Spanish as a Foreign Language, in which learners without linguistic immersion will participate, and the Subcorpus of Chinese Learners of Spanish as a Second Language, in which learners in immersion contexts. Both subcorpuses will include two levels: an oral component and a written component. It will also be made up of two control subcorpuses , the language spoken and written by native Spanish speakers and Mandarin Chinese. A Subcorpus of Native Spanish Speakers will be included , in which language samples from native Spanish speakers (from any of the linguistic communities of the Hispanic world) will be collected; and a Native Chinese Speaker Subcorpus, in which language samples from Mandarin Chinese speakers will be collected.

Theoretical and methodological principles for the foundation of the design

The design of any corpus must be based on specific theoretical and methodological criteria. It is considered that the principles of creating a corpus formulated by Sinclair (2005) are valid as the main foundation for the design of learner corpora. The creation of CACHE is governed by the ten principles of corpus creation by said author described below:

Principle for the selection of content : «The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise» (Sinclair, 2005, p. 1). The texts that will make up CACHE will be produced in natural conditions, so they will be created according to external criteria, that is, depending on their communicative function. Internal criteria referring to the language of the text will not be considered, that is, with the aim of collecting a certain linguistic structure.

Principle of representativeness: «Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen» (Sinclair, 2005, p. 2). CACHE cannot be a representative corpus of Spanish, because it is a corpus of learners. This corpus will be representative of the interlanguage of Chinese learners of Spanish. Now, when we talk about representativeness here, we do not intend to represent the "totality" of the Spanish language produced by Chinese learners, since a corpus that aims to do so would have to collect data, for example, in all the educational centers where it is learned. Spanish in China and where Chinese people learn Spanish abroad. Thus, said representativeness is achieved with the inclusion of interlanguage samples at different stages of development since learners from levels A1-C2 will participate in their construction. In fact, CACHE will constitute a corpus based on a transversally oriented sampling that will consider the different levels of competence. On the other hand, in order to obtain a variety of linguistic structures, different communicative tasks will be offered as stimuli for textual productions and the diversity of types of texts will be taken into account, other aspects that representativeness provides since Diversity of topics and types of texts favors the elicitation of various linguistic forms and structures.

Contrast principle: "Only those components of corpora which have been designed to be independently contrastive should be contrasted" (Sinclair, 2005, p. 3). What is described in the second principle favors compliance with the third, because representativeness with the inclusion of interlanguage samples at different stages of development with the participation of learners from levels A1-C2, allows contrasting different stages of the interlanguage. But, in addition, the design of CACHE includes the formation of two interlanguage subcorpuses: a Subcorpus of Chinese Learners of Spanish as a Foreign Language where learners in non-linguistic immersion contexts will participate and a Subcorpus of Chinese Learners of Spanish as a Second Language where learners will participate. in linguistic

immersion contexts. Each of these subcorpus will have an oral and written component. The design, methodology, and tasks planned for data collection will be the same for both subcorpuses, so that the possibility of carrying out interlanguage contrasts between learners in non-immersion and linguistic immersion contexts is guaranteed. Furthermore, the collection of data from the Subcorpus of Native Spanish Speakers and the Subcorpus of Native Chinese Speakers is conceived in the same way, which allows reliable contrasts to be made. As Cruz (2017) states, «the empirical analysis of a learner corpus is generally carried out by comparing it with a "control corpus" that is made up of texts with similar characteristics (topic, type of text, etc.) produced by native speakers». (p. 138). Being able to contrast the interlanguage of Chinese students with the language produced by native Spanish speakers makes it possible to achieve greater objectivity in the analysis.

Principle for determining the criteria structural: "Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination" (Sinclair, 2005, p. 5). Since CACHE tries to be a large corpus, this criterion is very important. However, this criterion cannot be applied in the way it would be applied to extensive reference corpora. CACHE is a corpus of learners and its structural framework design, described in the third principle, is already governed by exact criteria: depending on the context of study of the learners (with or without linguistic immersion) it is structured into two subcorpuses that, In turn, and taking into account the means of production of the texts (oral or written), they are divided into two components. In addition, the micro design will contemplate a division by language levels A1-C2.

Principle for establishment of labeling : «Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications» (Sinclair, 2005, p. 5). To comply with this principle, when the process of encoding CACHE texts, including metadata and linguistic annotation begins, a linguistic tool will be selected that, after the encoding process, produces a new XML file with the text. with metadata and annotations and keep the original file with clean text.

Principle for determining the sample: «Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events or should get as close to this target as possible. This means that samples will differ substantially in size»

(Sinclair, 2005, p. 7). In CACHE, only complete texts will be included without considering their length in number of words, because, as Sinclair (2005) emphasizes, selecting samples of the same size has no value from a linguistic point of view, but the integrity and representativeness of the texts complete if it acquires high value.

Principle for the selection of documentation: «The design and composition of a corpus should be fully documented with information about the contents and arguments in justification of the decisions taken» (Sinclair, 2005, p. 8). Given that CACHE is a corpus that aims to be a free access data source with which numerous linguists and Spanish teachers will be able to develop various studies on the interlanguage of Chinese students who study Spanish, the Web that will host it will contain all the documents that They theoretically and methodologically support the design and data collection, information related to the metadata included in the corpus and the linguistic annotation process, etc.

Principle of balance: «The corpus builder should retain, as target notions, representativeness, and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components" (Sinclair, 2005, p. 9). In this sense, the interest in representativeness and balance is explicit in the design of CACHE itself, which contains a written component and an oral component in each of its subcorpus. In this way, the results obtained from the use of CACHE can be extrapolated to both written and oral interlanguage.

Principle for determining the subject matter: "Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria" (Sinclair, 2005, p.10). In relation to this, compliance with the first principle referring to the content of the corpus and the second that requires representativeness, requires being governed by the ninth. The controls that will be carried out will be related to external criteria such as the level of interlanguage, but criteria will never be applied that condition the language samples in search of specific linguistic structures.

Principle of homogeneity: **«**A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided» (Sinclair, 2005, p. 14). From its design, CACHE considers respect for the homogeneity of its components and establishes clear criteria to avoid the introduction of atypical texts in the corpus. For example, texts that do not constitute a complete communication unit and others that show signs of translation use will be excluded.

Linguistic and extralinguistic metadata and annotation

In these pages, we have repeatedly talked about the objectivity achieved with corpus-based linguistic research; It has been declared that corpus linguistics is, precisely, an empirical approach. It has always been considered that this objectivity is achieved with the construction of a corpus representative of the linguistic reality in question; However, this degree of objectivity does not only respond to exhaustiveness, it also depends, to a large extent, on the use of appropriate linguistic and extralinguistic metadata, also guarantors of representativeness and the possibility of contrasts. Thus, CACHE will have a set of well-defined metadata that will favor the exploitation of the corpus, since all the metadata used in coding and annotation will allow information to be filtered in automatic searches.

As previously stated, since this research only takes into account the written component of the Subcorpus of Chinese Learners of Spanish as a Foreign Language, that of Native Spanish Speakers and that of the Subcorpus of Native Spanish Speakers, only the linguistic (textual) and extralinguistic (sociolinguistic) metadata that will be used in the coding of the texts of these subcorpuses.

Linguistic metadata

As can be seen in Table 1, the linguistic metadata that is considered for the coding of the texts reflects important information related to the genre and type of text, the type of task, the number of words and paragraphs and the use of dictionaries, in addition , indicate whether the text was written in a controlled context (in the classroom, under the supervision of teachers, etc.) or not and the identity of those in charge of examining the texts to determine their inclusion or exclusion from the corpus.

Table 1- Linguistic metadata for coding

Metadata	Description
sample id	SUBACHELE_CE_001 SUBHNE_CE_001(Subcorpus name + written component + sample number)
Predominant textual genre	Expository / descriptive / narrative / argumentative
Kind of text	Letter / email / story / biography / opinion article / critical review / message (blog / forum), etc.
Task type	Written expression / written expression and interaction
Number of words	-
Number of paragraphs	-
Dictionary use	But
Number of searches carried out	-
Checked	But
First reviewer	Name (would always be the author)
Second reviewer	Name (would always be a native Spanish specialist)

Extralinguistic metadata

For its part, Table 2 lists the extralinguistic metadata that are taken into account in the coding of the texts. Among them, the usual social variables in sociolinguistics (sex, age, level of education, mother tongue) stand out, considered ascribed variables. But other variables called acquired are also included, such as foreign languages, the level acquired in those foreign languages, stays abroad, studies and study centers, and motivation to study Spanish.

Table 2- Extralinguistic metadata for coding according to information from informants

Metadata	Description
Learner ID	SUBACHELE_CE_001_H1A1 (Sample ID + sex (M/F) + educational level (1/2) + level in Spanish (A1-C2)) SUBHNE_CE_001_M2 (Sample ID + sex (M/F) + educational level (1/2))
Sex	Man (H) Woman (M)
Age	
About the level of instruction	
Finished studies	Pre-university (P) / Bachelor's (G) / Master's (M) / Doctorate (D)
Studies in progress	Degree (G) / Master (M) / Doctorate (D)
Undergraduate / Postgraduate	Title
Study Center	Name
Current year	
About the language	
Mother tongue (L1)	Mandarin Chinese / Spanish /Other Which one?
Second mother tongue	Other Chinese languages, geolects or dialects learned within the family
10 foreign language/second language	English / Japanese / Korean /others Which one?
Self-assessment of the estimated level in that language (Subjective)	-
Level in that language (Objective-Certifications)	International certifications
20 foreign language/second language	English / Japanese / Korean /others Which one?
About the study of Spanish (only for Chinese informants)	
Self-assessment of the estimated level in Spanish (Subjective)	A1/ A2 / B1/ B2/ C1/ C2
Spanish level (Objective-Certifications)	EEE4 /8; CATTI 3 / 2 / 1; SIELE A1/ A2 / B1/ B2/ C1; DELE A1/ A2 / B1/ B2/ C1/ C2
Spanish level (Objective-Level test)	Level test result
Time learning Spanish	months or years
Study or exchange stay in a Hispanic country	But
Hispanic country	Country name
Time spent in that country	months or years
Work stay in a Hispanic country	But
Hispanic country	Country name
Time spent in that country	months or years
Motivation / Attitude towards the study of Spanish (Self-assessment)	1-10

Some of this metadata will be used to encode the texts on the corpus website and will be used to filter information in automatic searches, for example, age, language level, etc. However, others will be useful to disambiguate any doubts that arise when the data is examined. Thus, for example, when faced with a text from an informant who is in the third year of his degree in Hispanic Philology and has been studying Spanish for about two years and does not have any certification that accredits his language level, the language level with which he You have self-assessed yourself with the level indicated by the placement test, and the year of your degree will be taken into account. In this way, you will be able to better position yourself at the language level that corresponds to you. In short, some of this data will be used in the design of the corpus, while others will only allow for a better evaluation of the collected texts.

CONCLUSIONS

The critical review of the theory available in the specialized bibliography on corpus linguistics, linguistic corpora and, in particular, learner corpora, allowed us to systematize throughout these pages the fundamental theoretical and methodological notions and propose a precise definition. corpus of learners. Thus, the first contribution of this research is situated in the theoretical order with the theoretical systematization made available for future studies in this area.

Another theoretical and also practical contribution is manifested in the theoretical

and methodological foundation of the design of the Corpus of Chinese Learners of Spanish and in the determination of the stages for its creation: design, data collection and construction. Likewise, the results achieved in the execution of the design stage are exhibited: the conception of the corpus based on its objectives, its generalities, its subcorpus and components, the theoretical and methodological principles on which its creation is based and the linguistic metadata. and extralinguistics for their codification.

The Corpus of Chinese Learners of Spanish (CACHE), once built, will be a specialized, sample and representative, multilingual, mixed, open and transversally oriented corpus. It will be accessible to the scientific community in order to promote the development of research that contributes to the knowledge of the characteristics of the interlanguage produced by Chinese students of Spanish as a foreign/second language. Work continues on the data collection stages and its construction process.

BIBLIOGRAPHIC REFERENCES

Aarts, J. (2002). Does Corpus Linguistics Exist? Some Old and New Issues. Egil, B. L. y Hasselgren, A. (eds). *From the COLT's Mouth ... And Others' Language Corpora Studies. In Honour of Anna-Brita Stenström - Language and Computers* (pp. 117). Editions Rodopi B.V. https://brill.com/display/title/30088

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243-257. https://doi.org/10.1093/llc/8.4.243

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511804489

Bolaños, S. (2015). La lingüística de corpus: perspectivas para la investigación lingüística contemporánea. *Forma y Función*, 28, 1, 31-54. https://doi.org/10.15446/fyf.v28n1.51970

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language A Practical Guide to Using Corpora* (1era ed.). Routledge. https://www.routledge.com/Working-with-Specialized-Language-A-Practical-Guide-to-Using-Corpora/Bowker-Pearson/p/book/9780415236997

Calero F., Mª A.; Serrano Z., Mª I.; Gómez-Devís, Mª B. (2020). Codificación y etiquetado en los corpus de aprendices y su aplicación didáctica: la propuesta del corpus de interlengua española de aprendices sinohablantes (CINEAS), *E-AESLA*, 6, 206-222. https://cvc.cervantes.es/lengua/eaesla/eaesla_06.htm

Cruz, M. (2012). *Lingüística de corpus y enseñanza del español como 2/L.* Arco Libros. https://doi.org/10.5565/rev/doble.35

Cruz Piñol, M. (2017). *Lingüística de corpus y enseñanza del español como 2/L* (2da ed.). Editorial La Muralla. https://www.arcomuralla.com/detalle_libro.php?id=872&ideditorial_get=1

Francis, W. N., Kucera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Houghton Mifflin.

https://lib.ugent.be/en/catalog/rug01:000049253

Granger, S. (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. En Lllt.6. John Benjamins Publishing Company. https://benjamins.com/catalog/lllt.6

Gries, S. T., & Stefanowitsch, A. (2006). Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis. Mouton de Gruyter. https://books.google.com.pe/books?id=D_nVI5mFtkAC&lr=&num=20

Gries, S. (2009). What Is Corpus Linguistics. *Language and Linguistic Compass* 3, 12251241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Hincapié, D. (2018). Corpus de aprendientes de español como lengua extranjera y segunda lengua (CAELE/2): el componente escrito. *Forma y Función*, 31(2), 129-144. https://doi.org/10.15446/fyf.v31n2.74659

Hincapié M., D. y Bernal C., J. (2018). *Lingüística de corpus.* Instituto Caro y Cuervo. https://bibliotecadigital.caroycuervo.gov.co/1703/1/Linguistica-de-corpus-2018.pdf

Hincapié, D. y Rubio, R. (2018). Diseño y construcción del CAELE2: Base para una planificación curricular. *Hechos y Proyecciones del Lenguaje*, 23 (1), 42 52. https://revistas.udenar.edu.co/index.php/rheprol/article/view/3842

Krug, M., Schlüter, J., & Rosenbach, A. (2013). Introduction: Investigating language variation and change. In M. Krug & J. Schlüter (Eds.), *Research Methods in Language Variation and Change* (pp. 1-14). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511792519

Leech, G. (1992). Corpora theories of linguistic performance. En Svartvik, J. (Ed.). *Directions in Corpus Linguistics*. (pp. 105-122). Mouton de Gruyter. https://doi.org/10.1515/9783110867275

Leech, G (2011). Principles and Applications of Corpus Linguistics. En Viana, V., Zyngier, S. y Barnbrook, G. *Perspectives on Corpus Linguistics (Studies in Corpus Linguistics, 48)* (pp. 155170). John Benjamins Publishing Company. https://doi.org/10.1075/scl.48

Lemnitzer, L., & Zinsmeister, H. (2008). Korpuslinguistik. Eine Einführung. 2. Auflage. Tübingen: NarrFrancke Attempto Verlag. https://doi.org/10.1515/infodaf-2008-2-362

Lozano, C. (2022). CEDEL2: Diseño, compilación e interfaz web de un corpus online para la investigación de adquisiciones de L2 en España. *Investigación en un segundo idioma*, *38*(4), 965-983. https://doi.org/10.1177/02676583211050522

McEnery, A. M., Xiao, R. Z. and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book.* Routledge Applied Linguistics Series. Routledge. https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm

Parodi, G. (2008). Lingüística de corpus: Una introducción al ámbito. *RLA. Revista de lingüística teórica y aplicada*, 46(1), 93-119.

https://doi.org/10.4067/S0718-48832008000100006

Procházková, P. (2006). Fundamentos de la lingüística de corpus. Concepción de los corpus y métodos de investigación con corpus. https://docplayer.es/78249763-Fundamentos-de-la-linguistica-de-corpus.html

Rojo, G. (2022). *Introducción a la lingüística de corpus en español. LinRed*, (XIX). https://erevistas.publicaciones.uah.es/ojs/index.php/linred/article/view/1974

Rojo, G. y Palacios, I. (2022). Los corpus de aprendientes de español como L2. En: Parodi, G., Cantos-Gómez, p., Howe, C. (Eds). *Lingüística de corpus en español, (pp. 73-88). The Routledge Handbook of Spanish Corpus Linguistics.* https://www.routledge.com/Linguistica-de-corpus-en-espanol—The-Routledge-Handbook-of-Spanish/Parodi-Cantos-Gomez-Howe/p/book/9780367350123

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press. https://repositorio.uchile.cl/bitstream/handle/2250/139995/Corpus-concordance-collocation.pdf?sequence=4

Sinclair, J., & Wynne, M. (2005). How to build a corpus. En *Developing Linguistic Corpora: A Guide to Good Practice* (p. 108). AHDS. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf

Torruella, J. y Llisterri, J. (1999) Diseño de corpus textuales y orales. En Blecua, J.M., Clavería, G., Sánchez, C., Torruella, J. (Eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos.* Universidad Autónoma de Barcelona. (pp. 45-77). Milenio. https://dialnet.unirioja.es/servlet/articulo?codigo=595883

Tognini-Bonelli (2001). *Corpus Linguistics at WorK.* John Benjamins Publishing Company. https://doi.org/10.1075/scl.6

Villayandre Llamazares, M. (2008). Lingüística con corpus (I). *Estudios humanísticos. Filología*, *30*, 329-349. https://doi.org/10.18002/ehf.v0i30.2847

Conflict of interests:

The authors declare not to have any interest conflicts.

Authors' contribution:

The authors participated in the design, analysis of the documents and writing of the work.

Cite as
Gou , J., Rodríguez Roque, D., Cuba Vega, LE (2023). Theoretical and methodological design of the Corpus of Chinese Learners of Spanish. *Mendive. Journal of Education, 21* (4), e3347. https://mendive.upr.edu.cu/index.php/MendiveUPR/article/view/3347