e-ISSN 2385-3042

Annali di Ca' Foscari. Serie orientale
Vol. 58 – Giugno 2022

Learner Corpus Research Meets Chinese as a Second Language Acquisition: Achievements and Challenges

Alessia Iurato
Università Ca' Foscari Venezia, Italia; Universität Bremen, Deutschland

Abstract The article sheds light on Chinese as a Second Language Learner Corpus Research, emphasising advances and lacks in this field. First, the paper describes the potential of learner corpora in the investigation of learner language. Second, it provides an overview of Chinese learner corpus-based research and reviews existing L2 Chinese learner corpora. The paper highlights the lack of L2 Chinese learner corpora collecting data from Italian learners. Therefore, it emphasises the importance of compiling corpora to conduct studies on the acquisition of L2 Chinese by learners whose L1s are other than English or Asian languages.

Keywords Learner corpus research. Chinese as a second language acquisition. Corpus linguistics. L2 Chinese learner corpora. Learner corpus construction.

Edizioni
Ca'Foscari

Peer review

Submitted 2022-02-21
Accepted 2022-04-07
Published 2022-07-05

Open access

Citation Iurato, A. (2022). "Learner Corpus Research Meets Chinese as a Second Language Acquisition: Achievements and Challenges". *Annali di Ca' Foscari. Serie orientale*, 58, [1-34].

DOI 10.30687/AnnOr/2385-3042/2022/01/024

1 Introduction

The area of linguistic enquiry known as ‘Learner Corpus Research’ (LCR) originated in the late 1980s and it has mushroomed in recent years, since interest in computer learner corpora is growing fast, as well as the recognition of their theoretical and practical value (Granger 2002; 2012).

LCR has created an important link between the two previously distinct fields of corpus linguistics and foreign/second language research. LCR started as “an offshoot of corpus linguistics” (Granger, Gilquin, Meunier 2015b, 1), a field of study that had revealed enormous potential in investigating a wide range of native languages, but that had never explored non-native varieties.

LCR applies the paradigm of corpus linguistics to a wide range of (research) purposes in foreign/second language acquisition (Granger 2002; 2012). It is for this reason that LCR and corpus linguistics share a set of common features (Meunier 2021).

Over the last few decades, the field of linguistics has experienced an important paradigm shift from the study of language as an abstract mental representation to the study of language in its actual use (Zhang, Tao 2018). Corpus linguistics analysis has facilitated and supported this significant transition, as it makes it possible to systematically study patterns of language use through the investigation of large electronically stored and automatically processed collections of language samples (Zhang, Tao 2018). Established as an officially independent discipline in 1960, corpus linguistics began to expand in the 1990s, developing itself concurrently with the advancement of computational technology (Brezina, McEnery 2021). Compared to other linguistic approaches, corpus linguistics has attracted widespread interest among scholars because it displays several strengths relating to the support of computational tools. First, it bases linguistic analysis on naturally occurring data rather than mental abstract intuition; second, corpus data are empirical, constituting an important resource for interpreting patterns of language use in natural contexts; third, it utilises a large collection of texts, making it easy to provide information about the frequency of occurrence of linguistic features; finally, it analyses large collection of texts which allow to compare different varieties of a language or languages (Brezina, McEnery 2021). As summarised by McEnery et al. (2019), there are many advantages of using corpora, such as the shareability of data sets, which promotes re-use of data to develop further research, and the greater scale of analysis, which allows to draw broader conclusions.

The above-mentioned features lead corpus linguistics to establish itself as an increasingly reliable research approach, which became widely adopted in linguistic studies. It goes without saying that all those strengths also apply to LCR.

The present article will not give an overview of corpus linguistics research, as this is beyond its scope; sources for a more in-depth study of the subject can ba found in the literature reviews by Biber, Conrad and Reppen (1998), Brezina and McEnery (2021), McCarthy and O'Keeffe (2010b), McEnery and Hardie (2011), McEnery and Wilson (1996), Sinclair (1991), Stubbs and Halbe (2012), Tognini Bonelli (2010).

Following the definition provided by Brezina and McEnery (2021, 11) of corpus linguistics as "an approach to the study of language that uses computers to analyse large amounts of language data, both written and spoken, [called] corpora", learner corpus linguistics can be defined as a linguistic methodology which is founded on the use of electronic collections of learners' data (written and spoken), which we call 'learner corpora'.

The rationale of this paper is to identify achievements in LCR gained over the last thirty years, focusing on advances, challenges, pedagogical implications, and future directions of Chinese as a Second language Learner Corpus Research (CSL LCR). In detail, the first section of the paper will explore the development of LCR as an independent field; LCR core issues, potentials, and limits will be investigated as well. Furthermore, the study will answer the question whether LCR and Second Language Acquisition (SLA), despite both partaking to the broader field of L2 studies, have finally met. The second section of the paper will highlight how, in spite of the fact that in the last few years in Italy there has been a sudden and significant increase in the teaching of Chinese at all levels, from school to university, due to a widespread interest from learners in this language, research on Chinese acquisition in the Italian context is not very developed, partly because of little general scientific interest in this field until a few years ago. Moreover, the study will show that in the Italian scientific scenario, although there has been an indisputable growth in studies on the acquisition of Chinese, there is a lack of research applying the rigorous LCR methodology. The paper will therefore discuss the necessities of compiling L2 Chinese corpora which collect data from L1 Italian learners. This issue will be addressed by presenting recent attempts in the scientific community which incorporate the LCR methodology with research on L2 Chinese acquisition by Italian-speaking learners.

2 The Definition of 'Learner Corpus' and the Specificity of Learner Corpus Data

As a specific type of corpora, computer learner corpora are defined by Granger (2002, VII) as "electronic collections of spoken or written texts produced by foreign or second language learners in a variety of language settings". Analogously, Barlow (2005, 335) states that learner corpora are "digital representation of the performance or output […] of language learners". In Stewart, Bernardini and Aston's words (2004, 2), "learner corpora consist of writing or speech produced by language learners, or of materials written for language learners".

McEnery, Xiao and Tono (2006, 5) defined a corpus in corpus linguistics as a

> collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety.

To follow up on McEnery, Xiao and Tono's (2006) definition, Meunier (2021, 23) claimed that

> a learner corpus is thus a specific type of corpus which […] can broadly be defined as a collection of machine-readable texts consisting in representative samples of the language written and/or spoken by learners of an additional language.

As Gilquin (2015) summarises, the main characteristic that distinguishes a learner corpus from any other corpus is that it represents language as produced by foreign or second language learners. On the other hand, what distinguishes it from the data used in the previous SLA studies is that is representative of learner language use. These two distinctive features of learner corpora have led researchers to provide a more detailed definition of learner corpora as "systematic collections of texts produced by language learners" (Nesselhauf 2004, 125), where 'systematic' in Nesselhauf's words (2004, 127) means that

> the texts included in the corpus were selected on the basis of a number of – mostly external – criteria (e.g. learner level(s), learners' L1(s) [mother tongue(s)]) and that the selection is representative and balanced.

Agreeing with this definition, Callies and Götz add that learner corpora can be defined as "systematic collections of authentic, continuous, and contextualized language use (spoken or written) by L2 learners stored in electronic format", by stressing that "language samples should be representative of learners' contextualized use" (Callies, Götz 2015, 3). In this respect, Granger emphasises the two essential criteria for learner corpus data, i.e. the length of language samples and the context in which the language is produced:

> the notion of 'continuous text' lies at the heart of corpushood. A series of decontextualized words or sentences produced by learners, while being bona fide learner production data, will never qualify

> as learner corpus data. In addition, it is best to restrict the term ‘learner corpus’ to the most open-ended types of tasks, viz. those tasks that allow learners to choose their own word ordering rather than being requested to produce a particular word structure. (Granger 2008, 261)

In other words, learner corpus data are intended to be produced in open-ended tasks that allow learners to choose their own wording in spoken or written composition (Callies 2015a).

When defining learner corpora, a thorny issue is stating their degree of authenticity and naturalness, as

> [l]earner corpora represent a (more or less) naturalistic kind of data, collected with no (or very little) control over what learners say or write. (Gilquin 2021, 133)

This statement reveals that, as generally reported in the literature of LCR (Gilquin 2015; 2021; Granger 2002; 2012; Meunier 2021, among others), it is inaccurate to refer to learner corpora as collections of fully natural data. The concept of authenticity is indeed tricky in the case of learner language (Granger 2002). Learner language occurring in learner corpora is meant to be as authentic as possible (Meunier 2021); however, Granger’s (2008, 260) description of learner corpora as “electronic collections of (near-)natural foreign or second language learner texts assembled according to explicit design criteria” suggests that corpora may include texts that are not naturally occurring texts. So, the definition of learner corpora as “authentic” collection of learners’ data (Callies, Götz 2015, 3) stresses that the language produced by learners is meant to be considered merely situationally and interactionally authentic in the context of the SLA classroom (Callies, Götz 2015).

In this respect, Granger (2008, 259) points out that

> the term near-natural is used to highlight the “need for data that reflects as closely as possible ‘natural’ language use”.

So, it can be inferred that “[t]he content of a learner corpus will, more often than not, be exactly those activities which are natural in the context of a second language classroom” (Gilquin, Gries 2009, 7), such as role plays, speaking and reading activities, writing, etc. As stated by Granger (2002, 8),

> [i]n relation to learner corpora, the term ‘authentic’ […] covers different degrees of authenticity, ranging from ‘gathered from the genuine communication of people going about their normal business’ to ‘resulting from authentic classroom activity’.

It is in fact acknowledged that some degree of artificiality is always involved in the language teaching context and that learner data are therefore rarely fully natural.

From this it follows that studies collecting purely spontaneous oral or written learners' production in LCR are fairly rare, as learners' data cannot easily be spontaneously collected. This is because,

> for learners (especially foreign language learners), the target language fulfils only a limited number of functions, most of which are restricted to the classroom context. (Gilquin 2015, 10)

In fact, when learners are engaged in one of the required activities, such as writing a composition or role-playing with their classmates, they focus on practicing what they have learned to improve their language proficiency, rather than simply conveying a not premeditated message. Consequently, data collected under these circumstances cannot be considered the authentic linguistic output of "people going about their normal business" (Gilquin 2015, 10), as it is usually the case with fully natural corpus data. So, although according to Ellis (1994) learner corpus data fall within the more open-ended types of SLA data as natural language use data, and although they are supposed to be 'authentic', because they contain data gathered from the genuine communications of people doing their regular business,

> fully natural learner data is difficult to collect, especially in foreign language settings which give learners few opportunities to use the L2 in authentic everyday situations. (Granger 2012, 8)

Following this train of thought, Granger (2012) points out that it is counterproductive to analyse naturalistic data because of their drawbacks, such as: a) impossibility of exploring some specific language features because of the scarcity of the data; b) lack of control of certain factors external or internal to the learner that may affect learners' production; c) difficulty in the interpretation of the data.

In light of all the above, the traditional definition of what counts as a learner corpus needs to be expanded and renewed (Tracy-Ventura, Myles 2015). What is defined as a learner corpus has always been a hot topic in the SLA and LCR community, and in the past literature there is no general agreement on its definition. However, currently SLA and LCR communities unanimously agree that corpora can no longer be intended as fully authentic collections of learners' data just because they contain natural language use data produced by learners who use the L2 for authentic communication purposes.

Nowadays, both communities consider a learner corpus as a collection of computerised continuous, spontaneous, contextualised, representative (near-)natural (written or spoken) data produced by foreign or L2 learners, and gathered through those activities which are ordinarily carried out in the teaching and learning of second/foreign languages.

3 Potentials and Benefits of Learner Corpora

Similarly to other categories of corpora, learner corpora also allow to search through and analyse millions or even billions of words, a task that would be almost unmanageable without the existence of adequate computational technology. As highlighted by Leech (2011, 7),

> [i]f asked what is the one benefit that corpora can provide and that cannot be provided by other means, I would reply 'information about frequency'.

Leech highlights the great contribution of using corpora: they allow us to obtain information about the frequencies of occurrence of linguistic features and the contexts of their use. In the specific case of learner corpora, we can observe frequency and distribution of linguistic features in learner language use in language sampled in the corpora. This perspective of analysis is undoubtedly unique to (learner) corpus analysis (Brezina, McEnery 2021). Learner corpus linguistics, similarly to corpus linguistics, is able to answer research questions as the following:

- Is the linguistic feature of interest underrepresented or frequently distributed in learner language use?
- Which is the frequency rate of the linguistic feature of interest in learner language use?
- What are the typical collocations of the word of interest?
- What are the typical contexts of use in which the word of interest generally occurs?

Learner corpora allow us to answer the above mentioned questions not only in terms of quantitative analysis, but also in terms of qualitative analysis: this combination is the real strength of learner corpus linguistics (Brezina, McEnery 2021). Corpora provide information on how typical or unusual a linguistic feature is in learner language as sampled by the corpus (quantitative approach), and they simultaneously inform us about the context in which it occurs and, most of the time, the language background and the metadata of the learner (qualitative approach). In learner corpus linguistics, therefore, these two approaches of analysis are complementary and strongly connected: qualitative information is often a starting point for further quantitative analysis, whereas, to have a deeper and more correct interpretation of quantitative analysis, it is often necessary to re-analyse the text from which the data were extracted and engage in qualitative analysis (Brezina, McEnery 2021).

Although LCR has developed as a branch of corpus linguistics, we can state that nowadays it is an independent field of study (Granger 2021b). In the past thirty years, a considerable body of literature on LCR has been published, analysing a wide range of linguistic issues, from lexical to grammatical topics (Zhang, Tao 2018). The inauguration of the biannual *International Conference of Learner Corpus Research* in 2011, the foundation of the *Learner Corpus Association* in 2013, and the publication of the *Handbook of Learner Corpus Research* in 2015 all attest to the increasing relevance of LCR as an autonomous field of study and the worldwide growth of the LCR research community.

4 Learner Corpus Research and Second Language Acquisition: Have They Ever Met?

Although LCR and SLA studies both fall into the wider field of L2 studies, "it must be acknowledged that they are still essentially two different worlds" (Granger 2021a, 243).

The analysis of learner data is not new in linguistics. Written and spoken data have always been collected and investigated in SLA studies (Granger, Gilquin, Meunier 2015b). However, for a long time, in the field of SLA data were rather artificial, as they were collected by means of highly controlled tasks. Therefore, data in SLA were not regarded as a realistic reflection of learners' oral communication skills. Moreover, since data were always collected in small quantities, they were lacking in representativeness and statistical reliability. The need to overcome these theoretical and methodological gaps, as well as the need to create more "learner-aware/learner-focus pedagogical tools" (Granger, Gilquin, Meunier 2015b, 1), encouraged the emergence of learner corpora.

Differently from previous data collections analysed in SLA research, working on electronic collections of L2 data has brought two main advantages (Granger, Gilquin, Meunier 2015b). First, as these data collections are usually very big and collect data from a large number of participants, they are arguably more representative than smaller data collections gathered from a smaller number of students. Second, as the data are computerised, the analysis procedures are faster, and the data can be analysed for different research purposes. Part-of-speech (POS) taggers, for instance, assign each word in the learner corpus a tag labelling its grammatical category, facilitating the study of learners' use of specific grammatical categories, such as adverbs or prepositions. As for the analysis of errors, it is possible to add error annotation by means of specific software tools which allow to identify errors labelled with the same tag in a very large corpus in a short time.

The original project behind the development of major learner corpora was

> to enrich existing corpus collections with learner varieties and pass on the advances made in computerized corpus linguistics to applied linguistics. (Le Bruyn, Paquot 2021b, 1)

This placed the roots of LCR outside the domain of theory-driven SLA research.

However, the expectation that learner corpora could have become a relevant resource for theory-driven SLA research is not borne out (Le Bruyn, Paquot 2021b). Early learner corpus research was met with scepticism by the SLA community. Bell, Collins and Marsden (2021, 235) attribute LCR's scarce popularity within SLA research to its "preoccupation for coding errors, L1 transfer errors, and deviations from a target-like norm", at a stage when SLA research had already moved beyond descriptive generalisations.

Despite LCR has been the target of fierce criticism by the field of SLA since the 1990s, over the time LCR has improved consistently, by refining its theories and techniques, and SLA has recognised its developments (Granger 2012; 2021a; Meunier 2021). In the last few years, LCR and SLA have started to interact, nonetheless Myles (2015; 2021) notes that there is not a real systematic collaboration.

Granger (2012; 2021a) has often highlighted the usefulness of LCR methodological approach and software tools for SLA research, as well as the importance of SLA theory for the analysis of learner corpora:

> [i]t is now time that corpus linguists and SLA specialists work more closely, since the few studies that have used LCR [methodology] to test an SLA hypothesis demonstrate the potential of a more SLA-informed approach. (Granger 2012, 8)

As stressed by Tracy-Ventura and Myles (2015), corpora and corpus linguistics techniques should be an integral part of the toolkit of every SLA specialist devoted to the research and analysis of second/foreign language development. There are at least two main reasons why SLA experts should convert to and appreciate the use of corpus linguistics techniques. First, most of the current SLA hypotheses were based on research involving a small number of students. Therefore, findings cannot be considered generalizable and statistically reliable. Secondly, the use of electronic corpora could streamline and speed up the research process (Myles 2015). Moreover, corpus tools could allow SLA specialists to consult a vast amount of data for different research purposes. For example, as reported by Tracy-Ventura and

Paquot (2021b), the use of 'regex'[1] would allow for the extraction of linguistics patterns. Corpus tools can also be adopted to retrieve lexical bundles and collocations, which are the starting point for the study of formulaic language in learner language. The use of a concordance would allow to retrieve at once all instances of linguistic items of interest from an annotated learner corpus, thus reducing the duration of the research process. Furthermore, spoken learner corpora could be a valuable resource for the study of oral language development: transcriptions can be searched so that linguistic items of interest can be rapidly located.

LCR studies have been regularly criticised for being merely descriptive (Granger 2021a; Myles 2021). It is a common tendency to believe that studies in the field of learner corpus linguistics mainly provide statistical analyses, without dealing with adequate interpretation of the data. It must be emphasised that LCR, actually, recognises and attaches significative importance to the interpretation of findings. This theme was one of the most debated topics during the last *Graduate Student Conference in Learner Corpus Research 2021*, which took place in October 2021 at the Inland Norway University of Applied Sciences. On that occasion, the President of the Learner Corpus Association repeatedly stressed the need to restore the right balance between statistical analysis and data interpretation (Granger 2021b). In fact, only by placing the methodological paradigm of the learner corpus linguistics at the service of acquisitional studies will we be able to truly enhance the potential of this methodology.

Now, the main question is: in the reality of L2 studies, have LCR and SLA really met? As stated by Granger (2021a), the convergence has not yet been achieved. Myles (2015, 309) agrees and states that

> second language researchers have been rather slow in taking advantage of learner corpora and their associated computerized methodologies […], and LCR is not always fully informed by SLA research.

The reason why the limitation of LCR has been, and arguably still is, the lack of theoretical interpretations in the analysis of data is because it is mainly corpus linguists that started research activity in the field of LCR (Granger 2012). This can be considered positive, as they were able to adapt corpus linguistics techniques to the analysis of learners' data, by designing new corpora according to strict criteria which have been revised to meet the needs of LCR. This process required experience and corpus expertise. However, the down-

[1] 'Regex' or 'regexp', short for 'regular expression', is a sequence of symbols and characters expressing a string or pattern to be searched for within a longer piece of text.

side is that, as corpus specialists, and not as SLA specialists, they neglected the interpretations of data based on SLA theories; therefore, their research was mainly descriptive, relatively limited to the illustration of the corpus data, and lacking in theoretical frameworks.

Although there is still no real synergy between the two fields, they clearly have a lot to learn from each other:

> L2 studies would benefit greatly if SLA researchers were more familiar with the research carried out in LCR, and vice versa, resulting in more cross-referencing of each other's work in their respective publications. (Granger 2021a, 254)

SLA would provide the theoretical foundation, which is usually lacking in LCR; LCR, on the other hand, would offer descriptions of learner language use from a wide variety of L1 backgrounds at different proficiency levels. It is important that future works move towards this direction, to the mutual benefit of the two fields.

5 Development and Achievements of Learner Corpus Research

LCR emerged approximately 30 years ago thanks to the noteworthy research work conducted by Sylviane Granger and her research team at the Catholic University of Louvain, in Belgium (Zhang, Tao 2018; Granger 2021b). In the late 1980s, she conceived the idea of the Centre for English Corpus Linguistics (CECL), which gave rise to the creation of learner and multinational corpora especially for pedagogical purposes. Since its foundation, the Centre has produced fourteen corpora, some of which are amongst the largest of their type and collected data from numerous countries (Gráf 2017).

Granger's efforts also led to the creation of two new learner corpus methodologies: Contrastive Interlanguage Analysis (CIA) (Granger 2002) and Computer-aided Error Analysis (CEA) (Granger 2002). CIA is based on the combination of two types of comparison: non-native speakers/non-native speakers (NNS/NNS) comparison and native speakers/non-native speakers (NS/NNS) comparison. The first consists in comparing learners' data from different learner populations and different L1 backgrounds; the latter involves the comparison of learners' data with native speakers' data (Granger 2002). CEA, on the other hand, consists in "devising a standardized system for error tags and tagging all the errors in a learner corpus" (Granger 2002, 14). CEA is different from previous Error Analysis[2] studies because it is computer-aided, involves a higher degree of standardisation, and because learner "errors are presented in the full context of the text, alongside non-erroneous forms" (Granger 2002, 13).

Unsurprisingly, the earliest data sets collected data from learners of L2 English with different L1 language backgrounds (Granger 2012; Zhang, Tao 2018; Gráf 2017). The first to be compiled, and one of the most notable examples of 'commercial learner corpora', is the *Longman Learner Corpus* (Gilquin 2015; Gillard, Gadsby 1998). Its compilation was followed by the launch of the most renowned learner corpus: the *International Corpus of Learners* (*ICLE*) (Granger 2003). The *ICLE* was created in the 1990s by Sylviane Granger and her associates at the Catholic University of Louvain and collects written productions by intermediate and advanced L2 English learners. As a result of international cooperation work, the *ICLE* contains 3.7 million words produced by over 3,000 learners of L2 English with sixteen L1 different backgrounds. Moreover, in the *ICLE* more than 20 tasks and learner variables are documented (Granger et al. 2009). The corpus has been designed and compiled with the aim of developing analysis of high-frequency linguistic phenomena at the morphological, grammatical, lexical, and discourse levels (Granger 2003).

The *Cambridge Learner Corpus* (*CLC*) (Nicholls 2003) was compiled to support English language teaching publishers to produce a wide range of learning tools, such as dictionaries and course books.

The *NUS Corpus of Learner English* (*NUCLE*) (Dahlmeier, Ng, Wu 2013) was established for the annotation and evaluation of grammatical error correction systems.

Following this lead, a consistent number of learner corpora have gradually developed for the analysis of other European languages. Originally, studies were limited exclusively to the analysis of data of English learners, given the role of English as the major lingua franca of the world (Granger 2002; 2012). However, in recent decades, an increasing number of L2s has been the subject of studies in the field of LCR, which has therefore experienced an exponential growth. The *Learner Corpora Around the World* database,[3] managed by the

[2] Error Analysis is a type of linguistic analysis that focuses on the errors appearing in learner language. It consists of a comparison between the errors made in the target language and that target language itself. Error analysis emphasises the significance of learners' errors in second language. It determines whether those errors are systematic, and (if possible) explain what caused them. For a detailed discussion of the topic, see Corder 1975; Ellis 1985; 1987; 1994; Lüdeling, Hirschmann 2015; Richards 1980; Wallace Robinett, Schachter 1983.

[3] The *Learner Corpora Around the World* database is searchable at: `https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html`.

Centre for English Corpus Linguistics of the University of Louvain, currently collects 190 learner corpora, 100 (52.6%) representing L2 English, the rest focusing on other target languages (Italian, Spanish, French, German, Korean, Finnish, Arabic, Portuguese, Russian, etc.). Unfortunately, this database currently counts only two available L2 Chinese learner corpora: the *Jinan Learner Corpus* (*JCLC*)[4] (Wang, Malmasi, Huang 2015) and the *Spoken Chinese Corpus of Informal Interaction*,[5] which is not available for public use. In addition to the two above-mentioned, a review of existing L2 Chinese learner corpora that are not currently included in the database of *Learner Corpora Around the World* will be provided later in this paper.

6 Chinese as a Second Language Learner Corpus Research

Studies on corpus linguistics have developed considerably in China in the last decades. This paper, however, will merely focus on the development of Chinese as a Second Language Learner Corpus Research (CSL LCR).[6]

In Chinese, 'learner corpus' is identified as *xuéxízhě yǔliàokù* 学习者语料库 (learner corpus) or *zhōngjièyǔ yǔliàokù* 中介语语料库 (interlanguage corpus). Works in CSL LCR started in the late 1990s and have flourished over the past fifteen years (Zhang, Tao 2018). The unstoppable increase in the construction of corpora of L2 Chinese has led to a parallel exponential growth in the acquisitional studies of L2 Chinese over the last decade. Scholars in this field have produced a large body of studies exploring the acquisition of L2 Chinese at different levels from different perspectives. Another important indicator of the progress of this field is the establishment of the biennial CSL corpus research conference series *The International Symposium of Chinese Interlanguage Corpora: Construction and Application* (Zhang, Tao 2018; Xu 2019). The symposium first convened in Beijing in 2012, and the related conference proceedings, published by the *Journal of Chinese Language Teachers Association*, report findings on CSL learner corpora construction and their application. This research tradition is further strengthened by the first *International*

[4] The *JCLC* is searchable in the list of the *Learner Corpora Around the World* database at: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[5] The *Spoken Chinese Corpus of Informal Interaction* is searchable in the list of the *Learner Corpora Around the World* database at: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[6] For overviews of Chinese corpus linguistics, see the comprehensive overviews on the topic in Basciano, Gatti, Morbiato 2020; Feng 2006; McEnery, Xiao 2016; Xu 2015; Zhan et al. 2006.

Conference on Corpora of Chinese Spoken Interlanguage, which first was convened in 2015 (Zhang, Tao 2018). On both occasions, scholars' discussions focused on Chinese interlanguage corpora and related issues, such as research on the construction of Chinese interlanguage corpora, corpus-based research on the acquisition of Chinese sentence patterns and syntax, and corpus-based research on the acquisition of Chinese characters and words.

7 L2 Chinese Learner Corpora

The past twenty years have seen an exponential boom in Chinese learner corpus-based studies, due to China's growing global influence and the resulting increase of Teaching Chinese as a Second Language courses (Xu 2019). From this it follows that

> the construction of Chinese interlanguage corpora has become very popular in the wake of the augmented enrolment of international learners of Chinese. (44)

The central subject of CSL LCR has been the description of learner language, with a particular focus on learner errors (Zhang, Tao 2018). In fact,

> [e]arly learner corpora, such as the L2 Chinese Interlanguage Corpus and the HSK Dynamic Composition Corpus […], were only tagged for learner errors rather than language in its totality. Therefore, early CSL LCR used error analysis as its primary analytical framework. (50)

The earliest research was exclusively focused on the description of the taxonomy of errors; scholars limited their analysis to identifying the canonical taxonomies of underuse, overuse, and misuse of a target linguistic feature. Thanks to the use of large-scale learner corpora, they were also able to extract information on the frequency rate of learner errors, and then they tried to provide interpretations of findings based on quantitative analysis. However, as stressed by Tono (2003), it was necessary to move from pure descriptive taxonomies to exploring motivations for interlanguage errors. In order to provide an adequate picture of learner language use, later research began studying both errors and correct usages of linguistic features. Nowadays, as Zhang and Tao (2018, 50) claim, "[t]he norm of current CSL LCR is to look at language in its totality". Therefore, current research (Zhang 2010; Zhang 2014, among others), for example, examine the acquisition of specific linguistic features not only looking at learner errors, but also comparing the frequency of use of these linguistic features by learners with different L1s. Findings from investigations of patterns of acquisition and development, as well as of individual variation in learner language use, over the last few years have helped researchers and teachers, among other things, to understand learners' abilities, challenges, and developmental trajectories (Lu, Chen 2019).

7.1 Review of L2 Chinese Learner Corpora

As this is a growing field, it is increasingly difficult to be kept up-to-date and fully informed of the vastness of learner corpus projects around the world. A review of existing L2 Chinese learner corpora will be provided, with the aim to be as comprehensive as possible, including the major and the small-scale corpora found during the research carried out so far.

The first Chinese learner corpus project was born in a totally autonomous way from the corpus linguistics research carried out in Europe and USA. It is the *L2 Chinese Interlanguage Corpus* (*Hànyǔ zhōngjièyǔ yǔliàokù xìtǒng* 汉语中介语语料库系统), constructed from 1993 to 1995 by the research team led by Chu Chengzhi and Chen Xiaohe (Chu, Chen 1993; Chu et al. 1995) at the Beijing Language Institute, now Beijing Language and Culture University (BLCU). The corpus, as the first interlanguage Chinese data set, includes 5,774 essays written by 1,365 CSL learners from 96 different countries studying L2 Chinese at nine universities in China. As for the corpus size, it consists of 3,528,988 Chinese characters. The written data are POS tagged, parsed, and error-annotated. The data are also supplemented by rich ethnographic learners' metadata, documenting learners' sociolinguistic variables. Moreover, the corpus allows users to search by character, single word, sentence, discourse levels, or by learner metadata (Zhang, Tao 2018). Unfortunately, this corpus is not available for public use.

In the late 1990s and early 2000s, the aforementioned *L2 Chinese Interlanguage Corpus* was followed up by "the hitherto most frequently cited L2 Chinese learner corpus" (Xu 2019, 45), i.e. the *HSK*[7] *Dynamic Composition Corpus* (HSK *dòngtài zuòwén yǔliàokù* HSK 动态作文语料库) (Zhang 2003). The Corpus Version 1.0, launched in 2006, was compiled by the International Research and Development Center for Chinese Education at BLCU. This first version contained 10,740 compositions, including 4 million characters. Over the years, the corpus has been expanded and renewed, and in 2008 the new Corpus Version

[7] The HSK test (*Hànyǔ Shuǐpíng Kǎoshì* 汉语水平考试) is the Chinese language proficiency test of Mainland China for non-native speakers such as foreign students and overseas Chinese.

1.1 was launched. Xun Endong from the School of Information Science at BLCU provided the data of version 1.1. The corpus included 11,569 essays with 4.24 million characters. Since the technology of the Corpus Version 1.1 had become obsolete, the team has recently worked on the construction of version 2.0, which is now available.[8] The Corpus Version 2.0 retained all the data of Version 1.1. The overall corpus design and the development of the last software system are currently in charge of the team led by Zhang Baolin at BLCU, in collaboration with Beijing Weishu Technology Co., Ltd. Among the new features of the Corpus Version 2.0, there is the possibility of developing graphs for statistical analyses. Moreover, users can also add and edit error annotations to the corpus. As for learners' data, the corpus collects essays written by L2 Chinese learners who took the HSK Chinese language proficiency test in the period between 1992 and 2005. The genre of the essays is mainly narrative or argumentative. As reported by Zhang and Tao (2018), 88.81% of contributors to the corpus is from an Asian region or country, and 64% of the data is collected from Korean and Japanese learners. Learners' metadata are added to each composition; they include information on the learner's profile (such as age, nationality, L1 background) and the results of the reading test, the listening test, the written test, the spoken test, accompanied by the HSK total score and the certificate awarded. The error annotation is also added to the corpus at the levels of punctuation, character, lexicon, grammar, and discourse (Zhang, Tao 2018). The *HSK Dynamic Composition Corpus (Version 2.0)* is online, freely available for public use; once registered, users can start their research and check the scanned copies of learners' original compositions.[9]

In addition to the team working at BLCU, which can be considered the leader in L2 Chinese learner corpus research (Xu 2019), other research teams in the field have constructed their own distinctive L2 Chinese learner corpus projects. For example, National Taiwan Normal University (NTNU) has developed three important learner corpora: the *Chinese Character Errors Corpus*, the *Chinese as a Second Language Spoken Corpus*, and the *TOFCL Learner Corpus*.

The *Chinese Character Errors Corpus* (*CCEC*) (Teng et al. 2007), is the first and "arguably the earliest" (Xu 2019, 45) learner corpus collecting data to merely analyse learner errors in the writing of traditional characters. In fact, only error annotation is added to the corpus. The corpus collects data from 124 students at beginner, intermediate, and advanced levels from 22 different countries and from 15

[8] The *HSK Dynamic Composition Corpus (Version 2.0)* is searchable online at: `http://yuyanziyuan.blcu.edu.cn/en/info/1043/1501.htm`.

[9] Information about the *HSK Dynamic Composition Corpus (Version 2.0)* is available at: `http://yuyanziyuan.blcu.edu.cn/en/info/1043/1501.htm`.

different L1 backgrounds. Since the main contributors to other existing L2 Chinese learner corpora were Japanese and Korean learners, the *CCEC* decided to exclude learners whose L1s were Japanese and Korean (Zhang 2013). The error annotation consists of tagging misspellings, which are categorised into nine codes. The scanned version of the erroneous characters is stored and attached to the corpus. As reported by Xu (2019, 46),

> [t]he National Taiwan Normal University corpus […] might be disqualified as a corpus because its size is too small, and only individual characters, rather than running texts, were recorded in the database.

This corpus is not available for public use.

A similar corpus collecting misspelled Chinese characters, but in their simplified form, is the *Continuity Corpus of Chinese Interlanguage of Character-Error System* (*Hanzi Pianwu Biaozhu de Hanyu Lianxuxing Zhongjieyu Yuliaoku* 汉字偏误标注的汉语连续性中介语语料库) (Zhang 2017), which was developed at Sun Yat-sen University in Guangzhou.[10] It includes written texts, which were tokenised and POS tagged. Misspelled characters are tagged and, similarly to the *CCEC*, the image files of the original hand-written texts are stored along-side each entry in the corpus.

The *Chinese as a Second Language Spoken Corpus* consists of a collection of 450 learners' data gathered from the standard Mandarin language proficiency test, namely Test of Chinese as a Foreign Language (TOCFL), that has been adopted in Taiwan. Learners are grouped into basic and advanced proficiency levels, and their L1 backgrounds are English, Japanese and Korean. In total, the corpus contains 450 tests with 773,000 characters (Zhang, Tao 2018).[11]

The *TOCFL Learner Corpus* (Chang 2013) is the first learner corpus of traditional Chinese annotating grammatical errors (Lee, Tseng, Chang 2018). It contains written essays that students completed for the TOCFL test collected since 2016. The learners' data are accompanied by rich metadata including information on their L1, their CEFR level,[12] and information relating to the text genre, text function, text length, and score. Contributors to the learner corpus

[10] The *Hanzi Pianwu Biaozhu de Hanyu Lianxuxing Zhongjieyu Yuliaoku* corpus is available at: `https://languageresources.github.io/2018/06/24/朱述承_汉字偏误标注的汉语连续性中介语语料库/`.

[11] The *Chinese as a Second Language Corpus* is available online at: `http://140.122.83.243/mp3c`.

[12] The Common European Framework of Reference for Languages (CEFR) is an international standard for describing language ability. It describes language ability on a six-point scale (A1, A2, B1, B2, C1, C2), from A1 for beginners, up to C2 for those who

are from 42 different L1 backgrounds (mainly Japanese, followed by English, Vietnamese, Korean and Indonesian) (Zhang, Tao 2018). The corpus consists of 5,092 essays, with 1,740,000 characters and 1,140,000 words. A total of 33,835 grammatical errors in 2,837 essays and their corresponding corrections have been manually annotated by the team to analyse inappropriate linguistic usages (Lee, Tseng, Chang 2018). Two different error sets of error classifications were used simultaneously to tag grammatical errors: the first set

> denotes the coarse-grained surface differences, while the […] [second set] denote[s] the fine-grained linguistic category. The course-grained error types originate from comparing erroneous sentences with the correct usages. […] The fine-grained error types focus on representing linguistic concepts. (Lee, Tseng, Chang 2018, 2299)

The annotation was developed by Chinese native-speaking annotators specifically trained to follow the team's annotation guidelines for the error-tagging task. As for the coarse-grained level, four error types are identified (missing, redundant, incorrect selection, and incorrect word ordering); whereas at the fine-grained linguistic category level, 36 error types were catalogued (Lee, Tseng, Chang 2018). Once annotated, the team formatted the data in four sections: essay, learner, text, and mistake (Lee, Tseng, Chang 2018). The TOCFL is publicly available online to facilitate further research.[13]

Recently, some Chinese interlanguage projects have assumed the goal of compiling more balanced learner corpora covering both spoken and written interlanguage. The *Mandarin Interlanguage Corpus* (*MIC*) (Tsang, Yeung 2012) and the *Guangwai-Lancaster Chinese Learner Corpus* (*GWLCLC*)[14] are two cases in point.

The *MIC* (Tsang, Yeung 2012) is a small-scale learner corpus compiled at the University of Hong Kong. The corpus collects written and spoken data from pre-intermediate to intermediate Mandarin learners from different L1s. Both written and spoken production were collected in the form of coursework and examinations, amounting a total of approximately 50,000 characters and 60 hours of oral output. The *MIC* tags the errors at the character level to avoid "the situations where errors receive different treatments by the research team

have mastered a language. The *TOCFL Learner Corpus* groups learners into four different proficiency levels (A2, B1, B2, C1) following the CEFR standards (Cheng 2013).

[13] The *TOCFL Learner Corpus* is available online at: `http://nlp.ee.ncu.edu.tw/resource/tocfl.html`.

[14] Information about the *GWLCLC* can be found online at: `https://www.sketchengine.eu/guangwai-lancaster-chinese-learner-corpus/`.

and the user, and as a result do not turn up in the research" (Tsang, Yeung 2012, 190). The aims of the *MIC* project are: identifying and tracking both written and spoken language patterns from Mandarin learners of different L1s; facilitating research comparing features among learners of different L1s and possibly different proficiency levels; enhancing the development of teaching and assessment materials of L2 Chinese Teaching (Tsang, Yeung 2012).[15]

The *GWLCLC*[16] is a written and spoken L2 Chinese corpus collecting data produced by 886 learners from 80 different countries studying at Guangdong University of Foreign Studies (GDUFS) in China. Learners are classified into beginner, intermediate, and advanced proficiency levels, according to the HSK Chinese Proficiency Test score standards. The *GWLCLC* was built by Hai Xu and his team at GDUFS, in collaboration with Vaclav Brezina at Lancaster University. Originally, the project was initiated by Richard Xiao, whose vision was to bring corpus linguistics to the analysis of L2 spoken and written Chinese. The funding for the corpus was obtained by Xiao to whom the corpus is also dedicated. The corpus currently consists of 1,2 million words. It has both a spoken (621,900 tokens, 48%) and a written (672,328 tokens, 52%) part, and it is fully error tagged. Metadata are also incorporated in the corpus. The data (spoken and written) were collected from exams and tutorial sessions carried out at GDUFS. The written texts consist of short pieces and essays on a given topic. Spoken data comprise interactions typically between native and non-native speakers of Chinese and involve one, two, or multiple speakers. The corpus is a balanced sample that can be used to explore various theoretical and practical issues pertaining to the acquisition of Chinese as a foreign language. The corpus is searchable online on Sketch Engine.[17]

Two recently compiled large-scale written corpora are also worth mentioning, i.e. the *Jinan Chinese Learner Corpus* (*JCLC*) (Wang, Malmasi, Huang 2015) and the *Yet Another Chinese Learner Corpus* (*YACLC*) (Wang el al. 2021). The *JCLC* (Wang, Malmasi, Huang 2015) is a large-scale corpus of L2 Chinese written texts produced by learners at beginner, intermediate and advanced levels from 59 different L1 backgrounds learning Chinese at different universities in China. Some data are also collected from universities outside China. The *JCLC* project, started in 2006, aims to create a learner corpus similar to the *ICLE*. The *JCLC* is an ongoing project since new data continue to be collected and added to the corpus. The corpus currently

[15] The *MIC* is not available online; for the relevant publication, see Tsang, Yeung 2012.

[16] The *GWLCLC* is available online at: `https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fguangwai`.

[17] Information about the *GWLCLC* is available online at: `https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fguangwai`.

contains 5.91 million Chinese characters across 8,739 texts, which are completed with a rich set of metadata. The corpus has not been annotated yet, but, as declared by Wang, Malmasi and Huang (2012, 122), "[t]he inclusion of error annotations and manual corrections is another potential avenue for future work". The corpus is freely available to the research community upon contact with the research team.[18]

The *YACLC* (Wang et al. 2021) is a large-scale written multidimensional-annotated learner corpus. It contains 32,124 sentences from 2,421 essays, provided by around 50,000 CFL learners. Its main characteristic is the multidimensional annotation: for each sentence, an annotator was allowed to provide a variety of revisions. According to Wang et al. (2021), revisions include grammatical or fluency correction. Grammatical correction consists in making the sentence conform to grammar; whereas fluency correction consists in making the sentence fluent and native-sounding. 183 annotators were recruited and instructed to annotate the corpus, and each sentence was analysed and annotated by ten annotators which worked on a crowdsourcing platform specifically designed for the annotation process. This corpus is available online to "further enhance the studies on Chinese International Education and Chinese automatic grammatical error correction" (Wang 2021, 1).[19]

Let us now turn to an overview of the oral corpora produced to date. The *Spontaneous Chinese Learner Speech Corpus* (Wu, Shih 2014) is a large-scale spoken learner corpus developed at University of Illinois at Urbana-Champaign. It consists of 185 audio and video recordings, which were collected during Chinese speech training classes on a weekly basis from 2004 to 2009. The speakers in this corpus include 11 Chinese language teacher, 11 Korean-speaking learners, 23 English-speaking learners and 86 Chinese heritage learners. Participants were asked to complete two different oral open-ended tasks, and each of the tasks was designed to fit in a 50-minute class. The data were transcribed through a transcription website, where each speaker turn was presented individually with a link to the audio-video files. The data "has been used for perceptual ratings and acoustic analyses on oral fluency and foreign accent" (Wu, Shih 2014, 124). According to Wu and Shih (2014), this spoken corpus is a prolific resource with speech samples for various research topics.

Other smaller-scale corpora are also available online and allow users to analyse the L2 acquisition from different perspectives. The *Spoken Chinese Corpus of Informal Interaction* is a small-scale corpus created by Lin Lu at Massey University, in New Zealand. The

[18] The *Jinan Chinese Learner Corpus* is accessible online, upon contact with the research team, at: `https://hwy.jnu.edu.cn/jclc`.

[19] The *YACLC* is available online at: `http://cuge.baai.ac.cn/#/`.

corpus collects spoken data from English-speaking intermediate and advanced learners from New Zealand and Australia. The data are collected from informal conversation and interaction between 14 learners of L2 Chinese and Chinese native speakers. It is included in the list of *Learner Corpora Around the World* and is available online.[20]

The *COPA Corpus* (Zhang 2009) collects speech recordings from 120 college students learning Mandarin in Hong Kong. The data are gathered from conversation with Chinese native speakers. The corpus is part of the *SLABank* collection,[21] which is a component of *TalkBank*[22] dedicated to providing corpora for the study of second language acquisition and learning. The corpus is available for online browsing and download via *TalkBank*.[23]

Likewise, the *HKPU Corpus* is also part of the *SLABank* collection and is available via *TalkBank*.[24] It is a small-scale corpus and contains speech recordings of 20 college students learning Mandarin in Hong Kong. The tasks involve oral interviews.

The *Chinese Subcorpus of LINDSEI*[25] is a spoken learner corpus developed at South China Normal University by the team directed by He Anping. It is included in the database online of *LINDSEI Partners*, but, unfortunately, no further information is available.[26]

[20] The *Spoken Chinese Corpus of Informal Interaction* is available online at: https://github.com/blculyn.

[21] The *SLABank* is a component of *TalkBank* dedicated to providing corpora for the study of second language acquisition. It is available at: https://slabank.talkbank.org/.

[22] *TalkBank* is a project organised by Brian MacWhinney at Carnegie Mellon University. Its goal is to foster fundamental research in the study of human communication with an emphasis on spoken communication. Currently, it provides repositories in 14 research areas. Data in *TalkBank* have been contributed by hundreds of researchers working in over 34 languages internationally who are committed to principles of open data-sharing. Further information is searchable at: https://talkbank.org/.

[23] Information about the *COPA Corpus* and the link to it are available at: https://www.clarin.eu/resource-families/L2-corpora.

[24] The *HKPU Corpus* and related information are available at: https://slabank.talkbank.org/access/Mandarin/HKPU.html.

[25] *LINDSEI* is the *Louvain International Database of Spoken English Interlanguage*. This project was launched in 1995, five years after the start of the *ICLE* by the research team working at the Catholic University of Louvain. The aim of this project was to provide a spoken counterpart to *ICLE*, containing oral data produced by advanced learners of English from several mother tongue backgrounds. To date, eleven components have been completed and made available online. *LINDSEI* and *ICLE* have been built according to similar principles and share as many as ten mother tongue backgrounds. This means that they can be used in combination with each other to compare spoken and written interlanguage. Nowadays, *LINDSEI* contains also several subcorpora including a wide variety of L2s. Further information on *LINDSEI* and *LINDSEI partners* is searchable at: https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html.

[26] The *Chinese Subcorpus of LINDSEI* is searchable online at: https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei-partners.html.

In addition to the corpus projects mentioned so far, a few other L2 Chinese interlanguage corpora cited in the literature include those developed at Ludong University (Hu, Xu 2010), collecting data from Korean learners, and Nanjing Normal University (Xiao, Zhou 2014), which also provide a taxonomy for error annotation.

The *UCLA Heritage Language Learner Corpus* (Ming, Tao 2008)[27] is worth a separate discussion. It collects data from Chinese heritage language (HL) learners with Chinese family background. They represent "a specific group of CSL learners because they have usually acquired some degree of the Chinese language at a young age and have the advantage in listening to and speaking Chinese" (Zhang, Tao 2018, 53). The corpus was developed by the team working at the University of California, Los Angeles, and collects written data by Chinese HL learners at the intermediate level attending elementary heritage Chinese classes in 2006 and 2007. As for the corpus size, it contains approximately 1,000 samples of written essays and compositions students completed as homework assignments, with 200,000 characters. The genres of the texts are argumentative, narrative, and descriptive. The corpus is POS and error tagged, following a coding system which was specifically developed for HL learner error annotation. According to Zhang and Tao (2018), it is the first Chinese learner corpus built in North America.

7.2 L2 Chinese Learners' Input Corpora

Apart from the corpus-based Chinese interlanguage research, there is another specific type of learner corpora: 'L2 Chinese learners' input corpora' (Xu 2019). According to Xu (2019, 43),

> [t]he term 'input corpus' is used by some learner corpus linguists meaning the collection of learners' language exposures such as teachers' talk in class as well as the written texts that the learners are confronted with in learning.

The language input refers to written texts and Chinese language teaching textbooks that learners are likely to read in real life. The corpus compilers gather these specific input resources and then turn them into teaching and learning resources (Xu 2019). For example, the corpus of Chinese textbooks for international students developed by the research team at Xiamen University (Su 2010) is freely available online, and it has become a useful resource for researchers and

[27] Information about the *UCLA Heritage Language Learner Corpora* is searchable at: `https://nhlrc.ucla.edu/nhlrc/home`.

CSL teachers.[28] It collects data from 11 different CSL textbooks published from 1996 to 2007 which have been digitalised and converted into a corpus format. The corpus contains 771,350 Chinese characters (Xu 2019).

A similar project has been developed at Sun Yat-sen University in Guangzhou. The team has compiled a CSL textbook corpus which includes data from 1,752 textbooks published from 2006. Specifically, 1,802 out of 3,212 textbooks were published outside China (Zhou et al. 2017). This corpus is not available for public use.

As reported by Xu (2019), the teams working in the compilation of the above-mentioned corpora have conducted research on the coverage of vocabulary and grammar points across different CSL textbooks in order to provide useful learning and teaching sources.

8 Ongoing Research in CSL Learner Corpus Construction

The compilation of CSL LCR, as with learner corpora of many other languages, remains a challenging research topic for the research community. Nowadays, many research groups are engaged in the development of learner corpora that can be adopted in Chinese language teaching and learning in the future. For instance, current efforts in L2 Chinese learner corpus work at BLCU are devoted to the construction of a new large-scale corpus project: the *International Corpus of Learner Chinese*. The projected corpus will collect approximately 50 million characters, including 45 million written interlanguage Chinese and five million spoken interlanguage Chinese characters (Cui, Zhang 2011; Zhang, Cui 2013). It will collect data of learners from a wide variety of L1 backgrounds and learning contexts. Moreover, it will

> comprise five sub-corpora: raw corpus, annotated corpus, statistical information corpus, metadata corpus, and Chinese native speaker primary and middle school student corpus. (Zhang, Tao 2018, 54)

The corpus will be made available online to interested researchers and educators (Zhang, Tao 2018).

The team led by Liang Yuan at The Education University of Hong Kong is working on the construction of a CSL learner corpus for character-writing error that will be made searchable as an online

[28] The corpus of Chinese textbooks for international students is freely available at: `https://ncl.xmu.edu.cn/?query404path=https%3A%2F%2Fncl.xmu.edu.cn%2Fshj%2FDefault.aspx`.

resource for researchers and educators in the field of CSL. This project will collect handwriting data by students who learn L2 Chinese and will develop a tag-set for the error annotation.[29]

The present review on L2 learner corpora and the above-mentioned description of current work in progress in the field of CSL LCR demonstrate that most learner corpus projects mainly collect data from Asian or English-speaking learners and there is a shortage of corpora which collect data from learners whose L1 are European languages. Zhang and Tao (2018) are also in alignment with the results of the present survey. In fact, they state that since 2010 there has been a significant surge of interest in CSL LCR, but

> currently available corpora are unbalanced, with data mainly from Asian learners (specifically Korean and Japanese); there are far less data from European-speaking regions. (Zhang, Tao 2018, 54)

In this respect, Istvanova (2021), for instance, points out that the lack of L2 Chinese corpora collecting data from Slovakian learners hinders scientific production in the field of Chinese acquisitional studies and negatively affects the teaching of Chinese to Slovakian students. In order to fill this gap, Istvanova (2021) compiled a small-scale corpus of L2 Chinese with data from Slovakian learners with the purpose of gaining a better understanding of the gradual development of the learner's interlanguage and error's variation throughout the evolving language proficiency. Istvanova also hopes that the creation of the first Chinese learner corpus of Slovak students will contribute to increasing the current limited availability of teaching materials in the language combination Chinese - Slovak.

An analogous situation also arises in the context of the acquisition of Chinese by Italian-speaking learners. As stated by Romagnoli and Conti (2021), in Italy, the general interest in Chinese language has been echoed by a growing and conspicuous production of teaching materials and tools of various kinds and levels, with an ever-increasing attention to the different types of users and learning contexts. However, they emphasise that although Chinese teaching in terms of number of learners and teaching publications is in good health, research on teaching and acquisition is less developed, partially due to a general lack of scientific interest in this area until recent time. Along the same line, Iurato (2021a), emphasises the need for corpora which collect data from Italian learners of L2 Chinese that could be interrogated to investigate the acquisition of Chinese by Italian learners. However, it must be pointed out that in the last

[29] Information about the construction of the CSL learner corpus for character-writing error is searchable at: `https://www.eduhk.hk/chl/knowledge-transfer`.

few years in Italy there has been a remarkable increase in studies in the field of L2 Chinese acquisition produced by a constantly growing research community.[30] For an overview of the studies on the acquisition of L2 Chinese produced by the Italian research community, see Morbiato (2021) and Romagnoli and Conti (2021).

Nonetheless, despite the undeniable recent proliferation of acquisitional studies conducted in Italy, during the last *Graduate Student Conference on Learner Corpus Research 2021*, Iurato (2021a) pointed out that in the Italian context there is a lack of research applying the rigorous methodological framework of the LCR in the compilation of learner corpora to study the acquisition of L2 Chinese by Italian learners. She also highlighted that in LCR Chinese is understudied, and it is for this reason that more research should be developed in this field. In this regard, Iurato (2021b) is currently working on the compilation of a written and spoken L2 Chinese corpus to study the acquisition of the Chinese 是 *shì*… 的 *de* cleft construction by L1 Italian learners. She adopts a multi-method triangulated approach consisting in the combination of corpus data and experimental data to provide different insights into the phenomenon under study (Callies 2013; 2015b, Gilquin 2021). The corpus collects written and spoken data of 103 L1 Italian university learners studying at Ca’ Foscari University of Venice at elementary, intermediate, and advanced proficiency levels through open-ended tasks. The written corpus includes 53,248 Chinese characters and 2,337 occurrences of the *shì...de* construction. The spoken data (24 hours of speech) were manually transcribed and contain 19,073 Chinese characters and 1,305 occurrences of the *shì...de* construction. Moreover, Iurato collected data from 30 L1 Chinese speakers to include an equivalent native-speaker control group. All data are complemented by a rich set of metadata with learners’ and native speakers’ sociolinguistic variables. She also developed a target-oriented error taxonomy to manually annotate the grammatical errors; a pragmatic annotation was also added to detect the inappropriate use of the pragmatic functions (highlighting information and contrastive focus) of the *shì…de* cleft construction. Following Granger (2012) and Díez-Bedmar (2015), the identification of errors was carried out simultaneously by a bilingual team composed by two Chinese native-speaking experts and the researcher whose L1 is the same as the learners. Furthermore, to counterbalance potential construct underrepresentation (Tracy-Ventura, Myles 2015), she collected experimental data through additional experimental tasks. This research, to the best of our knowledge, is the first study grounded in the LCR framework which explores the acquisition of a specific syntactic linguistic feature by L1 Italian learners of Chinese.

[30] See, for example, Conti 2021; Conti, Lepedat 2021; Eletti, Casentini, Fontanarosa 2021; Gabbianelli 2020; Morbiato 2020; Romagnoli 2018; 2021; Tucci 2021.

Thanks to the funding for the Research Project of National Interest 2020 (PRIN 2020) allocated by MUR (Italian Ministry of University and Research), and co-funded by Università Ca' Foscari Venezia and Università degli Studi Roma Tre, the research group led by Bianca Basciano will work on the compilation of a new learner corpus to analyse the acquisition of Chinese resultative verbal complexes by L1 Italian learners. This project is also contextualised in the LCR framework and it is based on the combination and analysis of learner corpora and experimental data.

The two above-mentioned corpora will be useful to support the SLA and LCR community for the development of pedagogical tools. Moreover, they will be expanded and made available to other scholars and educators for further research.

9 Concluding Remarks

LCR has continuously grown over past the few decades, and it is a widely recognised branch of the broader field of corpus linguistics. As Tracy-Ventura and Paquot (2021b, 32) highlight,

> LCR has constantly questioned its role, methods, and goals, and has, as a result, evolved remarkably over the last thirty years.

The application of corpora has enabled researchers to

> explore areas as diverse as second language acquisition, psycholinguistics and natural language processing and to utilize their research findings within L2 pedagogy. (Gráf 2017, 22)

The field has seen its worldwide expansion on two main fronts: first, the increasing assortment of L2s for which learner corpora are being produced; second, the growing number of learner corpora used to analyse language learner use (Gráf 2017). As stated by Gráf (2017, 22), "[s]omewhat sadly – if perhaps not surprisingly – the target language for most learner corpora is English".

Nonetheless, this paper revealed how, since 2010, there has been a surge of interest in CSL LCR, reflecting what Zhang and Tao (2018, 54) believe is a "shift of attention to corpora in the field of CSL". In the last decade, a plethora of studies have addressed methodological issues to improve corpus construction, including collection of data, compilation, annotation, user interface, and application in teaching and learning. Moreover, the construction of L2 Chinese learner corpora has become the empirical basis for many doctoral dissertations, monographs, and research papers. Since the launch of the first L2 Chinese learner corpus in 1995, CSL LCR and related research have achieved a great deal. CSL LCR is a significant resource both for CSL research and corpus linguistics research. In fact, with the growing advancement in the construction of corpus tools, CSL LCR is intended to play an even more important role in understanding the acquisition of L2 Chinese.

As highlighted by Stewart, Bernardini and Aston (2004), there are different types of interaction between language corpora and learners. Learners may be the authors or contributors of corpus data, they may be the ultimate beneficiaries of a corpus insights, e.g. through the intermediation of the teacher, or they themselves may interrogate a corpus to better understand the syntactic, semantic, and pragmatic properties of a linguistic feature. This potential of learner corpora should be taken into account and applied in the context of Chinese learning and teaching to improve learners' explicit knowledge of the Chinese language. With regard to the pedagogical implications, it is also important to stress that LCR can be useful in two ways: the direct use and the indirect use (Zhang, Tao 2018). The direct use refers to the use of learner corpora to guide the writing of textbooks and dictionaries. CSL textbook writers and pedagogical materials developers should take advantage from

> [t]he rich understandings gained from LCR, including the acquisition orders and developmental sequences of different linguistic features, the typical errors learners tend to commit at different levels, and the desirable and less desirable L1 effects. (Zhang, Tao 2018, 56)

CSL LCR is also useful in dictionary compilation. For example, the *Cambridge Advanced Learners' Dictionary* (2003) used a learner corpus to include information about learner errors. As claimed by Zhang and Tao (2018), also L2 Chinese learner corpora could be used to compile dictionaries which includes information on frequent learner errors. As for the direct use of L2 Chinese learner corpora, teachers and students can use them for pedagogical purposes in L2 Chinese classrooms. Moreover, CSL learner corpora can support the process of CSL assessment. Similarly to other learner corpora, L2 Chinese learner corpora

> can serve as critical resources by providing quantitative, empirical information that can guide the development of assessment measures, such as placement, texts, exit tests, and other types of proficiency assessment. (Zhang, Tao 2018, 57)

The present review of existing L2 Chinese learner corpora suggests, in addition, a lack of learner corpora exclusively collecting data from Italian learners. Specifically, there is a lack in the compilation of L2

Chinese learner corpora applying the LCR methodological framework. The research carried out by Iurato (2021a; 2021b) and the PRIN research project directed by Basciano are a first contribution which aims at filling this gap; nonetheless, further studies need to be developed towards this direction in the future.

Finally, it is important to specify that, due to the unique characteristics of Chinese, corpus tools developed for English or other European languages cannot be easily applied to analyse Chinese; therefore, the compilation of learner corpora remains a challenging topic for the CSL LCR community, especially for those interested in analysing the acquisition of L2 Chinese by learners whose L1s are not English or Asian languages.

Bibliography

Basciano, B.; Gatti, F.; Morbiato, A. (2020). "Introduction". Basciano, B.; Gatti, F.; Morbiato, A. (eds), *Corpus-Based Research on Chinese Language and Linguistics*. Venezia: Edizioni Ca' Foscari, 7-15. Sinica veneziana 6. https://edizionicafoscari.unive.it/it/edizioni4/libri/978-88-6969-407-3/introduction/.

Barlow, M. (2005). "Computer-Based Analysis of Learner Language". Ellis, R.; Barkhuizen, G. (eds), *Analysing Learner Language*. Oxford; New York: Oxford University Press, 335-57.

Bell, P.; Collins, L.; Marsden, E. (2021). "Building an Oral and Written Learner Corpus of a School Programme: Methodological Issues". Le Bruyn, Paquot 2021a, 214-42. https://doi.org/10.1017/9781108674577.011.

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

Brezina, V.; McEnery, T. (2021). "Introduction to Corpus Linguistics". Tracy-Ventura, Paquot 2021a, 11-22.

Callies, M. (2013). "Triangulation". Schierholz, S.J.; Wiegand, H.E. (Hrsgg), *Wörterbücher zur Sprach - und Kommunikationswissenschaft [WSK] Online*. Berlin: De Gruyter Mouton. https://www.wsk.fau.de/.

Callies, M. (2015a). "Learner Corpus Methodology". Granger, Gilquin, Meunier 2015a, 35-55. https://doi.org/10.1017/CBO9781139649414.003.

Callies, M. (2015b). "Using Corpora in Language Testing and Assessment: Current Practice and Future Challenges". Castello, Ackerley, Coccetta 2015, 21-35.

Callies, M.; Götz, S. (2015). "Learner Corpora in Language Testing and Assessment. Prospect and Challenges". Callies, M.; Götz, S. (eds), *Learner Corpora in Language Testing and Assessment*. Amsterdam; Philadelphia: John Benjamins, 1-9.

Cambridge Advanced Learner's Dictionary (2003). Cambridge: Cambridge University Press.

Castello, E.; Ackerley, K.; Coccetta, F. (eds) (2015). *Studies in Learner Corpus Linguistics*. Bern: Peter Lang. https://doi.org/10.3726/978-3-0351-0736-4.

Chang Liping 张莉萍 (2013). “TOCFL Zuowen yuliaoku de jianzhi yu yingyong” TOCFL 作文语料库的建置与应用 (Compilation and applications of the TOCFL Composition Corpus). Cui Xiliang 崔希亮 Zhang Baolin 张宝林 (eds), *Di’er jie Hanyu zhongjieyu yuliaoku jianshe yu yingyong xueshu taolunhui lunwen xuanji* 第二届汉语中介语语料库建设与应用国际学术讨论会论文选集 (Selected papers from the 2nd International Conference on the Construction and Applications of Chinese Learner Corpora). Beijing: Beijing Language and Culture University Press, 141-52.

Chu, C.; Chen, X. (1993). “Constructing a Chinese Interlanguage Corpus”. *Shijie Hanyu Jiaoxue*, 7(3), 199-205.

Chu Chengzhi 储诚志 et al. (1995). “Hanyu zhongjieyu yuliaoku xitong yanzhi baogao” 汉语中介语语料库系统研制报告 (Research Report of The Corpus of Chinese Interlanguage [CCI 1.0]). Beijing: Beijing Language and Culture University Press.

Conti, S. (2021). “Italian Learners’ Use of Chinese Sentence-Final Particles: Marking Interrogatives in a Tandem-Learning Context. *Instructed Second Language Acquisition*, 5(2), 202-31. https://doi.org/10.1558/isla.18813.

Conti, S.; Lepedat, C. (2021). “*Situation-based Utterances* in italiano e in cinese: un confronto tra parlanti nativi e apprendenti italofoni”. Romagnoli, Conti 2021, 39-69.

Corder, S. (1975). “Error Analysis, Interlanguage and Second Language Acquisition”. *Language Teaching & Linguistics: Abstracts*, 8(4), 201-18. https://doi.org/10.1017/s0261444800002822.

Cui Xiliang 崔希亮; Zhang Baolin 张宝林 (2011). “Quanqiu hanyu xuexizhe yuliaoku jianshe fangan” 全球汉语学习者语料库建设方案 (A Proposal for the Building of the International Learner Corpus of Chinese). *Yuyan Wenzi Yingyong*, 19(2), 100-8.

Dahlmeier, D.; Ng, H.T.; Wu, S.M. (2013). “Building a Large Annotated Corpus of Learner English: the NUS Corpus of Learner English”. *Proceedings of the 8th Workshop on the Innovative Use of NLP for Building Educational Applications (BEA’13)*, 22-31.

Díez-Bedmar, M.B. (2015). “Dealing with Errors in Learner Corpora to Describe, Teach and Assess EFL Writing: Focus on Article Use”. Castello, Ackerley, Coccetta 2015, 37-69.

Eletti, V.; Casentini, M.; Fontanarosa, L. (2021). “Lo sviluppo della sensibilità sublessicale negli apprendenti italiani di cinese lingua straniera”. Romagnoli, Conti 2021, 159-80.

Ellis, R. (1985). *Understanding Second Language Acquisition*. Oxford: Oxford University Press.

Ellis, R. (1987). *Second Language Acquisition in Context*. New York: Prentice Hall.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Feng, Z. (2006). “Evolution and Present Situation of Corpus Research in China”. *International Journal of Corpus Linguistics*, 11(2), 173-207. https://doi.org/10.1075/ijcl.11.2.03fen.

Gabbianelli, G. (2020). “Video-Based Instruction and Students’ Perception of Cultural Understanding and Motivation in the Chinese Foreign Language Classroom”. *International Journal of Chinese Language Education*, 8, 37-70

Gillard, P.; Gadsby, A. (1998). “Using a Learners’ Corpus in Compiling ELT Dictionaries”. Granger, S.; Leech, G. (eds), *Learner English on Com-*

puter. London: Addison Wisley Longman, 159-71. https://doi.org/10.4324/9781315841342-12.

Gilquin, G. (2015). "From Design to Collection of Learner Corpora". Granger, Gilqin, Meunier 2015a, 9-34.

Gilquin, G. (2021). "Combining Learner Corpora and Experimental Methods". Tracy-Ventura, Paquot 2021a, 133-44.

Gilquin, G.; Gries, S.T. (2009). "Corpora and Experimental Methods: A State-of-the-Art Review". *Corpus Linguistics and Linguistic Theory*, 5(1), 1-26. https://doi.org/10.1515/CLLT.2009.001.

Gráf, T. (2017). "The Story of the Learner Corpus LINDSEI_CZ". *Studie z Aplikovane Lingvistiky*, 8(2), 22-35.

Granger, S. (2002). "A Bird's-Eye View of Learner Corpus Research". Granger, S.; Hung, J.; Petch-Tyson, S. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam; Philadephia: John Benjamins, 3-33. https://doi.org/10.1075/lllt.6.04gra.

Granger, S. (2003). "The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research". *TESOL Quarterly*, 37(3), 538-46. https://doi.org/10.2307/3588404.

Granger, S. (2008). "Learner Corpora". Lüdeling, A.; Kytö, M. (eds), *Corpus Linguistics. An International Handbook*, vol. 1. Berlin; New York: Walter de Gruyter, 259-75. https://doi.org/10.1080/00393270903392342.

Granger, S. (2012). "How to Use Foreign and Second Language Learner Corpora". Mackey, A.; Gass, S.M. (eds), *Research Methods in Second Language Acquisition. A Practical Guide*. Hoboken (NJ); Oxford: Wiley-Blackwell, 7-29. https://doi.org/10.1002/9781444347340.ch2.

Granger, S. (2021a). "Commentary: Have Learner Corpus Research and Second Language Acquisition Finally Met?". Le Bruyn, Paquot 2021a, 258-73.

Granger, S. (2021b). "Once Upon a Time… A Tale of Learner Corpus Research". Paper presented at *The Graduate Student Conference in Learner Corpus Research 2021* (Elverum, Inland Norway University of Applied Sciences, 12 October 2021).

Granger, S. et al. (2009). *International Corpus of Learner English (Version 2)*. Louvain-a-la-Neuve: Presses universitaires de Louvain.

Granger, S.; Gilquin, G.; Meunier, F. (eds) (2015a). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.

Granger, S.; Gilquin, G.; Meunier, F. (2015b). "Introduction: Learner Corpus Research – Past, Present and Future". Granger, Gilquin, Meunier 2015a, 1-5. https://doi.org/10.1017/CBO9781139649414.001.

Hu Xiaoqing 胡晓清; Xu Xiaoxing 许小星 (2010). "Mianxiang zhongwen dianhua jiaoxue de hanguo liuxuesheng hanyu zhongjieyu yuliaoku de kaifa yu jianshe" 面向中文电话教学的 "韩国留学生汉语中介语语料库"的开发与建设 (The Development of a Computer-Assisted Chinese Language Teaching Oriented Korean Students' Interlanguage Chinese Corpus). *Shuzihua Duiwai Hanyu Jiaoxue Shijian yu Fansi*, 19, 403-10.

Istvanova, M. (2021). "Chinese Learner Corpora and Creation of Slovak Learner Corpus of Chinese." *The Silk Road. Language and Culture*, 48-55.

Iurato, A. (2021a). "Compiling a Corpus of Written and Spoken L2 Chinese: Combining Pragmatic -and-Error- Annotation to Study the Chinese 是 *shì*…的 *de* Cleft Construction". Paper presented at *The Graduate Student Conference* *in Learner Corpus Research 2021* (Elverum, Inland Norway University of Applied Sciences, 12 October 2021).

Iurato, A. (2021b). "The Acquisition of the Chinese 是 *shì*… 的 *de* Construction by L1 Italian Learners: A Preliminary Analysis Based on a Learner Corpus and Experimental Data". Paper presented at the *6th International Conference on Chinese as a Second Language Research (CASLAR 6-2021)* (Washington, George Washington University, 31st July 2021).

Le Bruyn, B.; Paquot, M. (eds) (2021a). *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108674577.

Le Bruyn, B.; Paquot, M. (2021b). "Learner Corpus Research and Second Language Acquisition: An Attempt at Bridging the Gap". Le Bruyn, Paquot 2021a, 1-9. https://doi.org/10.1017/9781108674577.002.

Lee, L.; Tseng, Y.; Chang, L. (2018). "Building a TOFCL Learner Corpus for Chinese Grammatical Error Diagnosis". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resource Association, 2298-304.

Leech, G. (2011). "Frequency, Corpora and Language Learning". Meunier, F.; Cock, S.; Gilquin, G. (eds), *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: John Benjamins, 7-31.

Lu, X.; Chen, B. (2019). "Computational and Corpus Approaches to Chinese Language Learning: An Introduction". Lu, X.; Chen, B. (eds), *Computational and Corpus Approaches to Chinese Language Learning*. Singapore: Springer, 3-11.

Lüdeling, A.; Hirschmann, H. (2015). "Error Annotation Systems". Granger, Gilquin, Meunier 2015a, 135-57. https://doi.org/10.1017/CBO9781139649414.007.

McCarthy, M.; O'Keeffe, A. (eds) (2010a). *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

McCarthy, M.; O'Keeffe, A. (2010b). "Introduction". McCarthy, O'Keeffe 2010a, 1-28.

McEnery, T. et al. (2019). "Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use". *Annual Review of Applied Linguistics*, 39, 159-75.

McEnery, T.; Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T.; Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T.; Xiao, R.; Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

McEnery, T.; Xiao, R. (2016). "Corpus-Based Study of Chinese". Chan, S. (ed.), *The Routledge Encyclopedia of the Chinese language*. London: Routldge, 438-51.

Meunier, F. (2021). "Introduction to Learner Corpus Research". Tracy-Ventura, Paquot 2021a, 23-36.

Ming, T.; Tao, H. (2008). "Developing a Chinese Heritage Language Corpus: Issues and a Preliminary Report". He, A.W.; Xiao, Y. (eds), *Chinese as a Heritage Language: Fostering Rooted World Citizenry*. Honolulu: National Foreign Language Resource Center, University of Hawai'i, 167-78.

Morbiato, A. (2020). "Acquisition of Double-Nominative Constructions by Italian L1 Learners of Chinese. A Cross-Sectional Corpus Study". *Annali di*

Ca' Foscari. Serie Orientale, 56, 377-408. http://doi.org/10.30687/AnnOr/2385-3042/2020/56/015.

Morbiato, A. (2021). "Una Panoramica degli Studi sull'Acquisizione di Aspetti Sintattici e Strutture Grammaticali del Cinese da Parte di Italofoni". *La Lingua Cinese in Italia. Studi su Didattica e Acquisizione*. Roma: RomaTrePress, 87-114.

Myles, F. (2015). "Second Language Acquisition Theory and Learner Corpus Research". Granger, Gilquin, Meunier 2015a, 309-31. https://doi.org/10.1017/CBO9781139649414.014.

Myles, F. (2021). "Commentary: As SLA Perspective on Learner Corpus Research". Le Bruyn, Paquot 2021a, 258-73.

Nesselhauf, N. (2004). "Learner Corpora and their Potential in Language Teaching". Sinclair, J. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 125-52. https://doi.org/10.1075/scl.12.11nes.

Nicholls, D. (2003). "The Cambridge Learner Corpus – Error Coding and Analysis for Lexicography and ELT". *Proceedings of the Corpus Linguistics 2003 Conference (CL'03)*, 572-81.

Richards, J. (1980). "Second Language Acquisition: Error Analysis". *Annual Review of Applied Linguistics*, 1, 91-107.

Romagnoli, C. (2018). "The Acquisition of Mandarin Sentence Final Particles by Italian Learners". *International Review of Applied Linguistics in Language Teaching*, 58(4), 475-94.

Romagnoli, C. (2021). "Dire quasi la stessa cosa: l'apprendimento dei sinonimi in cinese Come lingua straniera". Romagnoli, Conti 2021, 71-86.

Romagnoli, C.; Conti, S. (a cura di) (2021). *La lingua cinese in Italia. Studi su didattica e acquisizione*. Roma: Roma TrE-Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stewart, D.; Bernardini, S.; Aston, G. (eds) (2004). "Introduction". Aston, G.; Bernardini, S.; Stewart, D., *Corpora and Language Learners*. Amsterdam: John Benjamins, 1-18. https://doi.org/10.1075/scl.17.01ste.

Stubbs, M.; Halbe, D. (2012). "Corpus Linguistics: Overview". Chapelle, C.A. (ed.), *The Encyclopedia of Applied Linguistics*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/book/10.1002/9781405198431.

Su Xinchun 苏新春 et al. (2010). "Jiaocai yuyan tongji yanjiu de duoweidu gongneng" 教材语言统计研究的多维度功能 (The Multi-Dimensional Function of Statistical Research on Textbook Language). *Proceedings of the Innovation of International Chinese Teaching Theories and Models Conference*. Xiamen: Xiamen Chubanshe, 128-41.

Teng, S. et al. (2007). "Huayuwen xuexizhe hanzi pianwu shuju ziliaoku jianli ji pianwu leixing fenxi" 华语文汉字偏误数据资料库建立及偏误类型分析 (The Construction of Chinese Learners' Character Writing Error Databse and the Analysis of Error Types). *Proceedings of 2007 National Linguistics Conference*, 313-25. Tainan: National Cheng Kung University.

Tognini Bonelli, E. (2010). "Theoretical Overview of the Evolution of Corpus Linguistics". McCarthy, O'Keeffe 2010a, 14-28.

Tono, Y. (2003). "Learner Corpora: Design, Development and Applications". Archer, D. et al. (eds), *Proceedings of the Corpus Linguistics 2003 Conference*. *UCREL Technical Paper*, 16, 800-9.

Tracy-Ventura, N.; Myles, F. (2015). “The Importance of Task Variability in the Design of Learner Corpora for SLA Research”. *International Journal of Learner Corpus Research*, 1(1), 58-95.

Tracy-Ventura, N.; Paquot, M. (eds) (2021a). *The Routledge Handbook of Second Language Acquisition and Corpora*. London: Routledge. https://doi.org/10.4324/9781351137904.

Tracy-Ventura, N.; Paquot, M. (2021b). “Second Language Acquisition and Corpora. An Overview”. Tracy-Ventura, Paquot 2021a, 1-8.

Tsang, W.; Yeung, Y. (2012). “The Development of the Mandarin Interlanguage Corpus (MIC) – A Preliminary Report on a Small-Scale Learner Database”. *JALT Journal*, 34(2), 187-208. https://doi.org/10.37546/JALTJJ34.2-1.

Tucci, T. (2021). “Il *Principle of Temporal Sequence* nell’apprendimento del sintagma locativo ‘*zài* 在 + luogo’: un’indagine preliminare su discenti italofoni”. Romagnoli, Conti 2021, 115-39.

Wallace Robinett, B.; Schachter, J. (eds) (1983). *Second Language Learning: Contrastive Analysis, Error Analysis, and Related Aspects*. Ann Arbor (MI): University of Michigan Press.

Wang, M.; Malmasi, S.; Huang, M. (2015). “The Jinan Chinese Learner Corpus”. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO): Association for Computational Linguistics, 118-23. https://doi.org/10.3115/v1/W15-0614.

Wang, Y. et al. (2021). “YACLC: A Chinese Learner Corpus with Multidimensional Annotation”. *Computer Science – Computation and Language*, 1-5. https://doi.org/10.48550/arXiv.2112.15043.

Wu, C.; Shih, C. (2014). “A Design of the Spontaneous Chinese Learner Speech Corpus”. *Learner Corpus Studies in Asia and the World*, 2, 115-24. https://doi.org/10.24546/81006694.

Xiao Xiqiang 肖奚强; Zhou Wenhua 周文华 (2014). “Hanyu zhongjieyu yuliaoku biaozhu de quanmianxing ji leibie wenti” 汉语中介语语料库标注的全面性及类别问题 (The Exhaustiveness and Taxonomy of Chinese Interlanguage Corpus Annotation). *Shijie Hanyu Jiaoxue*, 28(3), 368-77.

Xu, J. (2015). “Corpus-Based Chinese Studies: A Historical Review from the 1920s to the Present”. *Chinese Language and Discourse*, 6(2), 218-44.

Xu, J. (2019). “The Corpus Approach to the Teaching and Learning of Chinese as an L1 and an L2 in Retrospect”. Lu, X.; Chen, B. (eds), *Computational and Corpus Approaches to Chinese Language Learning*. Singapore: Springer, 33-53.

Zhan, W. et al. (2006). “Recent Developments in Chinese Corpus Research”. *Proceedings of the 13th NIJL International Symposium*, *Language Corpora: Their Compilation and Application*. Tokyo: University of Tokyo Press, 315-36.

Zhang Baolin 张宝林 (2003). “HSK dongtai zuowen yuliaoku jianjie” HSK 动态作文语料库简介 (Introducing Chinese Proficiency Test Dynamic Essay Corpus). *Ceshi Yanjiu*, 1(4), 37-8.

Zhang Baolin 张宝林 (2010). “Guanyu tongyongxing Hanyu zhongjieyu yuliaoku biaozhu moshi de zai renshi” 关于通用型汉语中介语语料库标注模式的再认识 (Re-Considering the Models of Annotation of All-Purpose Chinese Interlanguage Corpus). *Shijie Hanyu Jiaoxue*, 27(1), 128-40.

Zhang Baolin 张宝林 (2013). “Huibi yu fanhua – Jiyu HSK Dongtai Zuowen Yuliaoku de ‘ba’ zi ju xide kaocha” 回避与繁华 – 基于 HSK 动态作文语料库的“把”字句习得考察 (Avoidance and Overgeneralization – An Investigtion of Acquisition of the Ba-Sentence Based on the HSK Dynamic Composition Corpus). *Shijie Hanyu Jiaoxue*, 24(2), 263-78.

Zhang Baolin 张宝林; Cui Xiliang 崔希亮 (2013). "'Quanqiu hanyu zhongjieyu yuliaoku jianshe he yanjiu' de sheji linian" "全球汉语中介语建设和研究" 的设计理念 (Design Concepts of "The Construction and Research of the Interlanguage Corpus of Chinese from Global Learners"). *Yuyan Jiaoxue Yu Yanjiu*, 24(5), 27-34.

Zhang, J. (2014). "A Learner Corpus Study of L2 Lexical Development of Chinese Resultative Verb Compounds". *Journal of the Chinese Language Teachers Association*, 49(3), 1-24.

Zhang, J.; Tao, H. (2018). "Corpus-Based Research in Chinese as a Second Language". Ke, C. (ed.), *The Routledge Handbook of Chinese Second Language Acquisition*. London; New York: Routledge, 48-62.

Zhang Ruipeng 张瑞朋 (2017). "Hanyu zhongjieyu yuliaoku zhong de hanzi pianwu chuli yanjiu" 汉语中介语语料库中的汉字偏误处理研究 (The Character Errors in Chinese Interlanguage Corpora). *Yuliaoku Yuyanxue*, 3(2), 50-9.

Zhang, Y. (2009). *A Tutor for Learning Chinese Sounds through Pinyin* [PhD Dissertation]. Pittsburgh (PA): Carnegie Mellon University.

Zhou Xiaobing 周小兵 et al. (2017). "Guoji hanyu jiaocai yuliaoku de jianshe yu yingyong" 国际汉语教材语料库的建设与应用 (The Construction and Application of International Chinese Textbook Corpus). *Yuyan Wenzi Yingyong*, 25(1), 125-35.