# Investigating Chinese learner corpus research and learner corpora

## Main features, critical issues and future pathways

*Alessia Iurato*

Despite the increasingly wide-ranging accessibility of L2 Chinese learner corpora and achievements in Chinese Learner Corpus Research (CLCR), there are no studies which provide a critical analysis of the key features and limitations of this expanding field. This study therefore aims to fill this lack in the literature by both investigating the actual state of CLCR and existing L2 Chinese learner corpora and outlining their main features and critical issues. First, the paper introduces the development and current trends in CLCR, particularly emphasizing the widespread use of multi-method approaches in this field. Second, it focuses on design issues of L2 Chinese learner corpora, identifying main characteristics and limitations. The paper shows that L2 Chinese learner corpora present many gaps concerning language-related, task-related, and learner-related criteria. Gaps in learner corpus analysis and annotation are also discussed. Third, the paper offers the first analysis conducted to date that categorizes existing L2 Chinese learner corpora according to mode (written, spoken and multimodal) and size (large-scale and small-scale). Finally, the article directs attention toward challenges in this field, concluding with future directions for CLCR and its intersections with Second Language Acquisition (SLA) to support L2 Chinese teaching and learning. Suggestions on direct and indirect applications of L2 Chinese learner corpora data are also provided.

Keywords: Chinese Learner Corpus Research, Chinese learner corpora, corpus design, corpus annotation, Chinese as a second/foreign language

## 1. Introduction[1]

A learner corpus is a specific type of corpus which can be broadly defined as a collection of machine-readable texts consisting in "continuous, spontaneous, contextualized, representative (near-)natural

[1] This research has been conducted within the project 'The acquisition of Chinese resultative verbal complexes by L1 Italian learners: combining learner corpus and experimental data' directed by Prof. Bianca Basciano, which has received funding by

(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" (Iurato 2022: 714-5; Granger 2002; Callies and Götz 2015; Meunier 2021). One of the main potentials of learner corpora is allowing us to observe the frequency, distribution, and the contexts of use of specific linguistic features in learner language use from both quantitative and qualitative perspectives (Granger 2012; Meunier 2021; Iurato 2022).[2] Learner corpus data generally serve two main purposes: first, informing Second Language Acquisition (SLA) research, including, for instance, usage-based approaches (cf. Wulff 2021), generative approaches (cf. Lozano 2021), variationist approaches (cf. Gudmestad 2021), pragmatic approaches (cf. Fernández and Staples 2021); second, providing "useful impact for applied projects (including the creation or improvement of teaching materials/approaches, or the training/development of Natural Language Processing tools)" (Meunier 2021: 23).

The application of learner corpus data gave rise to a flood of studies that have been grouped under the umbrella term of 'Learner Corpus Research' (LCR) (Granger, Gilquin and Meunier 2015). LCR can be considered as a 'young field of study' (Alonso-Ramos 2016: 3), since studies in LCR as a field independent from corpus linguistics began only in the late 1980s (Granger 2002; Meunier 2021; Tracy-Ventura and Paquot 2021). Although the application of learner corpora has also developed in the context of Chinese as a Second/Foreign language (CSL/CFL) research, in our opinion it is more appropriate to speak of infancy rather than youth when referring to Chinese Learner Corpus Research (CLCR). In fact, although CLCR has witnessed significant growth in the last twenty years (Zhang and Tao 2018), in this paper we will show that there are still several limitations concerning, among others, corpus design, annotation procedure, and pedagogical implications. Almost thirty years have now passed since the release of the first Chinese learner corpus project (Chu and Chen 1993), and the number of L2 Chinese corpora compiled in and outside of China has increased significantly since then (see Section 5). Nonetheless, we noticed that there is a lack of research that analyzes the current state of CLCR, discussing both key features and gaps to be filled. This article therefore stems from the idea that an analysis of this growing discipline, including a well-structured overview of corpora compiled to date, is needed. We think that several areas of research might benefit from this analysis. First, the LCR community. Most LCR focuses

the Italian Ministry of University and Research (MUR), PRIN 2020, project number n. 20209M3Z77. Information on the project can be found at: https://pric.unive.it/projects/achieve/home.

For detailed comments and suggestions on this paper, I would like to thank the three anonymous reviewers. I would also like to thank Prof. Mauro Tosco for his kindness and patience as editor of *Kervan*.

[2] For a discussion on potentials of learner corpora, see Meunier (2021) and Iurato (2022).

on the analysis of L2 English and research is generally developed through the application of L2 English learner corpora, while Chinese is a generally understudied language in this field (Iurato 2022). Thus, an investigation of CLCR would allow the entire LCR community to benefit from an up-to-date analysis on the current state and trends of this still under-explored branch of study that is growing within the LCR macro-area. CSL/CFL research might also benefit from an investigation of main features and limitations in CLCR. In fact, scholars might have a deeper awareness of how and where to direct their studies, and might expand the use of corpora (e.g., in Chinese teaching) to support learners' acquisition process of L2 Chinese.

This article also stems from the desire to point out that the number of existing L2 Chinese learner corpora is higher than that specified in the *Learner corpora around the world* database.[3] The database currently includes only three L2 Chinese learner corpora, the *Jinan Chinese Learner Corpus*[4] (JCLC; Wang *et al*. 2015), the *Multilingual Corpus of Second Language Speech*[5] (MuSSeL; Rubio *et al*. 2021), and the *Spoken Chinese Corpus of Informal Interaction*[6] (Li 2021). This means that Chinese learner corpora constitute only 1,5% of the total number of corpora (197) listed in the database; 52,3% of the learner corpora in the database are L2 English learner corpora and the remaining 46,2% represents other target languages (Spanish, French, Italian, German, Dutch, Finnish, Portuguese, Russian, Korean, Arabic, Czech, Croatian, etc.). These results suggest that Chinese is not well ranked among "non-English" learner corpora; nonetheless, this overlooks the consistent growth of Chinese learner corpora and CLCR over the past two decades that has been shown by Zhang and Tao (2018), Xu (2019), and Iurato (2022).

The aim of this contribution is threefold. First, to scrutinize the development and achievements in the field of CLCR and extend the work by Iurato (2022) by providing a more comprehensive and structured overview of existing L2 Chinese learner corpora; second, to identify and discuss the main features and limitations that characterize the design of existing L2 Chinese learner corpora (e.g., types of data collection and types of annotation). Three, to determinate and investigate weaknesses and critical issues in CLCR to enable future studies to fill the gaps in this field.

[3] The *Learner corpora around the world* database is maintained by the Center for English Corpus Linguistics of the University of Louvain. Related information can be found at: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (last access: June 2022).

[4] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (last access: 12 June 2022).

[5] https://l2trec.utah.edu/learner-corpora/mussel/ (last access: 26 June 2022).

[6] https://github.com/blculyn?tab=repositories (last access: 26 June 2022).

In this article, we will first define the status of Chinese as a second/foreign language. Second, we will introduce a bird's-eye view of CLCR and examine the main features and limitations related to learner corpus design. Third, we will categorize and analyze all L2 Chinese learner corpora compiled in the last twenty years according to mode (written, oral, and multimodal) and size (large-scale and small-scale). To the best of our knowledge, this is the first and most extensive exploration of L2 Chinese corpora which classifies them by mode and size. Lastly, we will establish challenges in CLCR and outline suggestions for further applications of L2 Chinese learner corpora to improve L2 Chinese teaching and learning.

## 2. The status of Chinese as a second/foreign language

Chinese[7] is the largest spoken language in the world, with a total of 1.12 billion speakers, of which 921 million are native speakers and 199 million are non-native speakers (Eberhard *et al.* 2022). Differently from Chinese, English is more widely spoken as a lingua franca by non-native speakers (978 million), and less widespread among native speakers (370 million) (Eberhard *et al.* 2022). Although the number of non-native speakers of Chinese is lower than that of Chinese native speakers and to the number of non-native speakers of English, it is worth highlighting that the population of learners of L2 Chinese is constantly expanding. The results of statistical surveys published by the Institute of International Education (2019) show that the international student enrolment trend in Chinese universities has witnessed an increase of 49,9% from 2013 to 2019, as illustrated in Table 1.

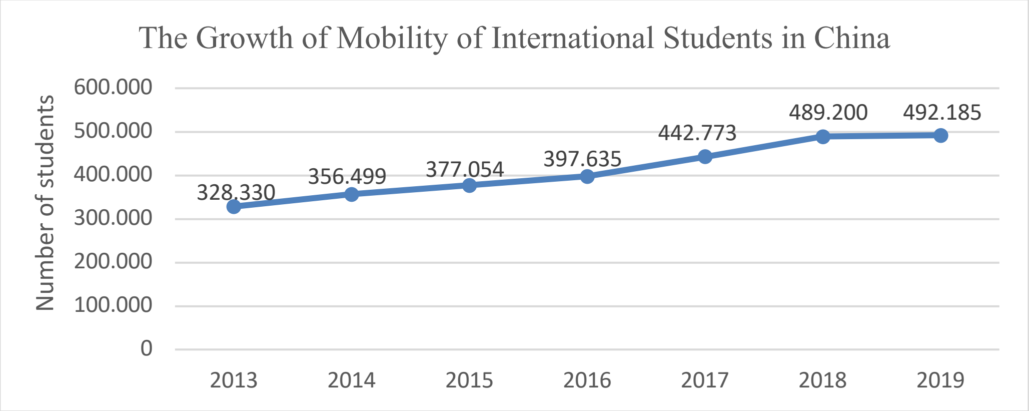

*Table 1.* The growth trend of international student enrolment in universities in China, 2013-2019[8]

[7] In this paper, the term 'Chinese' refers to the official standard language of the People's Republic of China.

[8] Data published by the Institute of International Education (2019). Consulted online on 5 July 2022 at:

According to Biney and Cheng (2021: 305), China is "an emerging preferred study location by most international students in recent times," and it has the third largest population of foreign students. Moreover, education programs established around the world for teaching Chinese and the professional development of Chinese language teachers has grown significantly in recent years, partially "due to a massive infusion of human and material resources, soft power diplomacy, and advocacy by the Chinese Government" (Duff *et al.* 2013: 1).

In addition to an overall increase in interest in learning L2 Chinese, there has been an evolution in learners' motivations for studying Chinese (Wen 2020). Up until approximately thirty years ago, the reason for a learner to study Chinese was the desire to become a sinologist; in recent years, students are approaching the study of this language to become socially and occupationally more competitive and attractive. In fact, L2 Chinese learners are aware that studying this language will enable them to enjoy benefits, such as higher post-graduation employment prospects, the establishment of business partnerships, and the advantage of directly experiencing China's rapid economic growth (Biney and Cheng 2021; Duff *et al.* 2013; Sung 2013; Wen 2011). Given the increasing number of learners, more tools and learner corpora are required for teaching and learning Chinese. However, despite the rapid expansion of CSL/CFL studies, research on CSL/CFL acquisition is still limited "and clearly lags behind the research progress of general second language acquisition (SLA)" (Wen 2019: 2). Likewise, we will show that CLCR calls for more studies. In fact, there are still several limitations, compared to the research achievement of general LCR, such as, to name a few, the limited number of learner corpora, the contexts of data collection, and the types of findings, which usually report mainly descriptive analyses, thus hindering the access of a full picture of the CSL/CFL learners' acquisitional development.

## 3. Chinese learner corpus research: A field on the move

Research in CLCR first appeared in the 1990s (Zhang and Tao 2018; Xu 2019) and many Chinese learner corpus projects have been carried out over the past decades (see Lee *et al.* 2018; Tsang and Yeung 2012; Wang *et al.* 2015; Wang *et al.* 2021; Wu and Shih 2014, among others). With an increasing number of corpora available, the scope of CLCR studies has also expanded considerably. In the last two decades, there has been a surge of interest in research on CSL/CFL (Lu and Chen 2019; Wen 2019). A considerable body of literature in L2 Chinese studies has been produced (Cui 2005; Lee *et al.* 2019; Li *et al.* 2020; Xu *et al.* 2019; Yang 2016; Zhang 2014); through the analysis of available learner corpora, researchers "have

https://www.iie.org/en/Research-and-Insights/Project-Atlas/Explore-Data/China.

explored a wide range of inquiries regarding how learners acquire the different levels and aspects of Chinese" (Zhang and Tao 2018: 49). Moreover, with the increasing availability of Chinese learner corpora, "investigations of patterns of as well as individual variation in learner language use, acquisition, and development have flourished" (Lu and Chen 2019: 6). The biennial Chinese learner corpus research conference series, which was first convened in 2012, as well as related conference proceedings, published by the *Journal of Chinese Language Teachers Association*, and the first *International Conference on Corpora of Chinese Spoken Interlanguage* in 2015, also bear witness to the progress of this discipline.

In contrast to early research in CLCR, whose predominant subject was the description of learners' language based on the canonical identification of error taxonomies (Tono 2003; Zhang and Tao 2018), the trend of current research is "to look at language in its totality" (Zhang and Tao 2018: 50) to present an overall picture of learner language use. Therefore, recent studies also contemplate the investigation of acquisitional and developmental patterns (see Lu and Chen 2019), as well as the comparison of the frequency of use of specific linguistic features by learners with different L1 backgrounds (see Li *et al.* 2020; Xu *et al.* 2019; Zhang 2014). However, Zhang and Tao (2018), Istvanova (2021), and Iurato (2022) point out that there is a significant limitation in the scenario of existing L2 Chinese learner corpora: currently available corpora seem unbalanced, as they collect data mainly from Asian or English-speaking learners. The analysis conducted for the present article confirms that there is a lack of L2 Chinese corpora for learners whose L1s are European languages other than English, which limits and affects research on Chinese as a second language acquisition. This lack is addressed, for example, by Istvanova (2021), who created a specific corpus of L2 Chinese with data from Slovak learners, since no analyzable data could be found in existing Chinese learner corpora. An analogous situation also arises in the context of the acquisition of Chinese by Italian-speaking learners. In fact, Iurato (2022) emphasizes the need for corpora that collect data from Italian learners of L2 Chinese to investigate the acquisition of Chinese by Italian learners, given the present remarkable increase in L2 Chinese teaching and learning in Italy (Romagnoli and Conti 2021).

A significant advancement in LCR research in recent years has been the emergence of the multi-method approach (Gilquin and Gries, 2009; Gilquin 2021; Lozano and Mendicoetxea 2015; Mendicoetxea and Lozano 2018), consisting of the combination of two types of analyses. On the one hand, the comparison of learner corpus data with native speakers' data, which highlights learners' performance in terms of overuse or underuse of linguistic features compared to native speakers. On the other hand, the use of experimental data, which reveals learners' competence. The need for learner corpus data to be supplemented and verified by elicited data had already been established decades ago in CLCR, as also highlighted by Zhang and Tao (2018). In fact, Shi (1998) combined the learner corpus method with elicited data to investigate CSL/CFL learners' acquisition order of 22 syntactic structures. Recent studies in CLCR have continued to advance toward this methodological approach. Qu (2013), for instance, adopts a mixed approach based on the use of corpus data supplemented by grammaticality tests and interviews to study CSL/CFL learners' acquisition of *gěi* 给 as a preposition or a verb. Similarly, Iurato (2021a; 2021b) adopts a triangulated multi-method approach combining corpus and experimental data to study the acquisition of the *shì* 是*...de* 的 syntactic cleft construction by L1 Italian learners. The usefulness of a combined 'corpus plus experimental data' approach has already been shown in LCR and CLCR studies, especially when a small learner corpus is used to analyze a specific (rare) syntactic feature (Römer *et al.* 2014). Furthermore, methodological pluralism helps to prevent or avoid phenomena such as the underrepresentation of a linguistic feature, when a specific rare structure is the object of a study (Tracy-Ventura and Myles 2015).

## 4. Learner corpus design issues in CLCR: Features and limitations

Following the main categories proposed by Tono (2003) and Alonso-Ramos (2016) concerning the design of learner corpora, we illustrate in Table 2. the three major features that distinguish the design of L2 Chinese learner corpora compiled to date. In what follows, we will outline and discuss the features and limitations related to L2 Chinese learner corpora that emerged from our exploration of them. Specifically, we analyzed existing L2 Chinese learner corpora according to a) language-related criteria, including mode, genre (text type), style, and topic of the corpus data; b) task-related criteria, including information on data collection, data elicitation, use of references and time limitation; c) learner-related criteria, such as age, learning context, L1 background, and Chinese language proficiency level. To the best of our knowledge, this type of examination is the first to be conducted on L2 Chinese learner corpora.

| *Language-related criteria* | *Task-related criteria* | *Learner—related criteria* |
|---|---|---|
| *Mode*: written corpora are more numerous than spoken and multimodal corpora | *Data collection*: cross-sectional | *Age*: mostly young adults (college and undergraduate students) |
| *Genre*: essays and oral presentations | *Data elicitation*: most are written and spoken open-ended compositions based on topics included in exams or proposed by teachers/researchers as assignments | *Learning context*: Chinese as a foreign/second language in university context |
| *Style*: narrative, argumentative, descriptive are the most common | *Use of references*: not indicated | *L1 background*: Japanese, Korean (and in general Asian languages), and English are the most common L1s |
| *Topic*: generally related to personal life experiences, such us travelling, holidays, festivities, job, etc. | *Time limitation*: generally based on the duration of exams; sometimes no time limitation | *Proficiency level*: sometimes inadequately assessed, because based on external factors. Sometimes based on HSK test[9] standards, which, however, are not officially recognized as comparable to CEFR[10] standards |

*Table 2.* Features in the design of L2 Chinese learner corpora

As far as language-related criteria are concerned, we found that there is a significant scarcity of spoken and multimodal learner L2 Chinese corpora. Moreover, we found that there are no L2 Chinese academic learner corpora, such as the *Corpus of Academic Learner English* (CALE; Callies and Zaytseva 2013), a

[9] 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.

[10] *Common European Framework of Reference for Languages*. Information on CEFR (Council of Europe 2022) is available at: https://www.coe.int/en/web/common-european-framework-reference-languages (last access: 15 September 2022).

specialized learner corpus comprising academic texts produced by learners of L2 English in university courses, and the *Varieties of English for Specific Purposes dAtabase*[11] (VESPA; Paquot *et al.* forthcoming), a more comprehensive learner corpus project that aims to build a large collection of L2 texts in a wide range of disciplines (linguistics, business, medicine, law, biology, etc), registers (papers, reports, MA dissertations), and degrees of writer expertise in academic settings (from first-year students to PhD students).

As for task-related criteria, our exploration reveals that there are no longitudinal Chinese learner corpora: this is certainly a big gap compared to the amount of existing longitudinal learner corpora of L2 European languages, such as LONGDALE[12] (Meunier 2016), LANGSNAP[13] (Tracy-Ventura *et al.* 2016), LEONIDE[14] (Glazniesk *et al.* 2022) and LoCCLI[15] (Spina and Siyanova in preparation). The analysis of learner-related criteria shows that L2 Chinese learner corpora mainly collect data from learners whose L1s are English or Asian languages, as evidenced above. We agree with Istvanova (2021) and Iurato (2022) that this lack in CLCR makes the study of Chinese acquisition by learners with L1 backgrounds other than English or Asian languages more difficult. Furthermore, we found out that most L2 Chinese learner corpora do not incorporate L1 data as an integral part of the design, and this makes it difficult to identify specific features of L1-related errors or mis/over/underuse patterns, as also highlighted by Tono (2003).

Moreover, we found that language proficiency, "which is not always optimally identified in English LCR" (Alonso-Ramos 2016: 7), is also a critical issue in CLCR, as sometimes it seems to be inadequately assessed. We noticed that not all corpora provide information on how learners' language proficiency level is defined. Sometimes a placement test is used to define learners' proficiency, but the level of language proficiency identified does not always correspond to official standards (e.g., CEFR standards). It is indeed based on learner's external factors, such as institutional status, length of study, course of study, which cannot be considered reliable criteria, as stressed by Callies *et al.* (2014), Leclercq and Edmonds (2014), and Tono (2003).

[11] https://uclouvain.be/en/research-institutes/ilc/cecl/vespa.html (last access: 31 August 2022).

[12] https://uclouvain.be/en/research-institutes/ilc/cecl/longdale.html (last access: 15 September 2022).

[13] http://langsnap.soton.ac.uk/ (last access: 26 June 2022).

[14] https://www.porta.eurac.edu/lci/leonide/ (last access: 25 June 2022).

[15] https://ricerca.unistrapg.it/retrieve/handle/20.500.12071/12083/6717/poster_spina_siyanova.pdf (last access: 27 June 2022).

Following general LCR practice, we observed that two approaches are usually employed in CLCR to analyze corpus data: Contrastive Interlanguage Analysis (CIA; Granger 1996) and Computer-aided Error Analysis (CEA; Dagneaux *et al.* 1998). Specifically, we noted that the CEA approach is more frequently used compared to the CIA approach.

As far as CIA is concerned, CLCR follows the terminological practice of (English) LCR adopting the terms 'overuse' and 'underuse' to refer to quantitative differences between native speakers and learners (see, for example, Xu *et al.* 2019). However, as Alonso-Ramos (2016) states, this terminology should be interpreted prescriptively, and not descriptively. In light of this, since CIA has come under criticism for its "lack of recognition of learner language as a variety in its own right, not as a faulty or deficient variety" (Alonso-Ramos 2016: 7-8), Granger (2015) describes a renewed version of CIA which promotes the concepts of *reference language varieties* and *interlanguage varieties*, where the first term substitutes 'native language,' and the second term substitutes 'learner language.' As stated by Alonso-Ramos (2016: 8), instead of considering the "idealized imagine of a native speaker," we think that CLCR, as well as general LCR, might consider the principle proposed by Callies (2015), according to which native-like proficiency should be interpreted as a gradual phenomenon which goes beyond the sharp separation between native and non-native speakers.

Our study reveals that the CEA approach is the most frequently used in CLCR, since the most common type of corpus annotation used in Chinese learner corpora is error annotation (as, e.g., in Chang 2013; Tsang and Yeung 2012; Wang *et al.* 2021). Part-of-speech (POS) tagging (e.g., in Chu and Chen 1993; Ming and Tao 2008; Zhang 2003), parsing, and semantic tagging (e.g., in Chu and Chen 1993) are less common in Chinese learner corpora, as Table 4. illustrates (see Section 5). We also found that error taxonomies adopted in error annotation in available L2 Chinese corpora are designed to cater for the anomalous nature of learner language. Error types commonly are taken from predefined error tagset based on the error categorization proposed by Lu (1994) and Lü (1993), such as omission (*yíluò* 遗落), word ordering error (*cuò xù* 错序), substitution (*wù dài* 误代), misuse (*wùyòng* 误用), and overuse (*wù jiā* 误加). Our investigation also shows that error analysis is mainly conducted at the character and lexical levels (as, for example, in Teng *et al.* 2007; Tsang and Yeung 2012); analysis at the grammatical and punctuation level is less frequent (see, for example, Ming and Tao 2008; Lee *et al.* 2018), whereas annotation at the discourse level is very rare (as, e.g., in Chu and Chen 1993). The error annotation in CLCR is generally manually developed by a group of annotators specifically trained for the error tagging procedure (as, e.g., in Lee *et al.* 2018; Wang *et al.* 2021). Following this type of error annotation, numerous acquisitional studies in China have been carried out (see, for example, Li *et al.* 2020; Liu and Ming 2015; Xie 2010; Zhang 2016, among others); however, in our opinion they can be criticized for focusing mainly on the description of errors. In contrast, we found that recent studies move from purely descriptive analyses to the interpretation of the data, supported by SLA theories (see, for example, Chen and Xu, 2019; Xu *et al.* 2019; Zhang 2014).

Another limitation we noticed is that error annotated L2 Chinese learner corpora do not include any target hypothesis. Following van Rooy (2015), we argue that problem-oriented annotation systems would be preferable, since the identification of grammar errors and appropriateness errors needs linguistic and extra-linguistic contexts, which can only be found with the support of a target hypothesis (Lüdeling and Hirschmann 2015). Moreover, it is important to bear in mind that errors could be analyzed differently depending on which target hypothesis is adopted (Lüdeling and Hirschmann 2015). However, we believe it is worth pointing out that a multi-layered annotation can be found in the *Yet Another Chinese Learner Corpus*[16] (YACLC; Wang *et al.* 2021), where each annotator provides a variety of interpretations of grammatical errors, including grammatical and fluency corrections. As far as we know, this is a unique case in CLCR. Nevertheless, it has its limits, as it provides analysis only at the grammatical level.

Despite these limitations, it is important to clarify that corpus data must be interpreted in order to be useful, given that an annotation layer does not code the absolute truth – rather it exemplifies one way of interpreting the corpus data. Therefore, it can be inferred that "[t]he 'correct' version against which a learner utterance is evaluated is simply a necessary methodological step in identifying an error" (Lüdeling and Hirschmann 2015: 141).

In light of the above, we conclude that it would be desirable for CLCR to explore new directions in the corpus design, so that richer and more reliable amount of data would be available to conduct further acquisition analyses from different perspectives.

## 5. Existing Chinese learner corpora: An analysis based on corpus mode and size

As more and more L2 Chinese learner corpora compilation projects continue to flourish, it is difficult to keep up to date with the increasing amount of corpus projects around the world. However, our purpose in this paper is to present an overview of existing learner corpora that aims to be as comprehensive as possible, further expading the overview work proposed by Iurato (2022). We have grouped existing L2 Chinese learner corpora according to their mode (written, spoken and multimodal)

[16] http://cuge.baai.ac.cn/#/ (last access: 15 September 2022).

and size (large and small-scale), as illustrated in Table 3. As for the corpus size, here a corpus is categorized as 'small-scale' or 'large-scale' according to the following criteria: in the case of written corpora, corpora that collect data from up to 1,000 learners and include maximum 500,000 Chinese characters fall under the 'small-scale' category, while corpora that collect data from more than 1,000 students and include more than 500,000 characters are considered 'large-scale' corpora. As for spoken corpora, corpora that include up to 50,000 characters and 60 hours of recordings of speech fall into the 'small-scale' category, while corpora that include more than 50,000 characters and 60 hours of recordings of speech fall into the 'large-scale' category. Finally, as for multimodal corpora, a corpus is considered 'small-scale' if it collects less than 100,000 Chinese characters and less than 60 hours of recordings of speech, while it is labelled as a 'large-scale' corpus if it contains more than 100,000 characters and more than 60 hours of recordings of speech. As far as we know, this is the most extensive analysis of available L2 Chinese learner corpora, including written, spoken, and multimodal corpora, carried out to date that categorizes existing corpora according to their mode and size.

| *Written corpora* | *Spoken corpora* | *Multimodal corpora* |
|---|---|---|
| *Large-scale* | *Large-scale* | *Large-scale* |
| L2 Chinese Interlanguage Corpus (Chu and Chen 1993) | Chinese as a Second Language Spoken Corpus (Chang 2016) | Guangwai-Lancaster Chinese Learner Corpus[17] |
| HSK Dynamic Composition Corpus (Zhang 2003) | Spontaneous Chinese Learner Speech Corpus (Wu and Shih 2014) | *Small-scale* |
| TOCFL Learner Corpus (Chang 2013) | Multilingual Corpus of Second Language Speech (MuSSeL; Rubio *et al.* 2021) | Mandarin Interlanguage Corpus (MIC; Tsang and Yeung 2012) |
| Jinan Chinese Learner Corpus (JCLC; Wang *et al.* 2015) | *Small-scale* | Bimodal Italian Learner Corpus of L2 Chinese (BILCC; Iurato forthcoming) |
| Yet Another Chinese Learner Corpus (YACLC; Wang *et al.* 2021) | Spoken Chinese Corpus of Informal Interaction (Li 2021) | |
| *Small-scale* | COPA corpus (Zhang 2009) | |
| Chinese Character Errors Corpus (CCEC; Teng *et al.* 2007) | HKPU corpus (Chan *et al.* 2013) | |
| Hanzi Pianwu Biaozhu de Hanyu Lianxuxing Zhongjieyu Yuliaoku (Zhang 2017) | | |
| UCLA Heritage Language Learner Corpus (Ming and Tao 2008) | | |

*Table 3.* Available Chinese learner corpora analyzed in this paper and grouped by their mode and size

Beginning with the written corpora, the first group consists of large-scale learner corpora. The earliest interlanguage Chinese learner corpus is the *L2 Chinese Interlanguage Corpus* (*Hànyǔ zhōngjièyǔ yǔliàokù xìtǒng* 汉语中介语语料库系统; Chu and Chen 1993), which was compiled between 1993 and 1995 at the Beijing Language Institute, now Beijing Language and Culture University. This first project was carried out independently of the research in LCR conducted in Europe and America (Xu 2019). It contains 5,774 written essays with 3,528,988 Chinese characters produced by 1,365 CSL/CFL learners from 96 different countries studying L2 Chinese at nine universities in China. Data are POS tagged, parsed, and error

[17] https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fguangwai (last access: 15 September 2022).

annotated, and they are accompanied by rich ethnographic learners' metadata, documenting learners' sociolinguistic variables. This corpus is not available for public use.

The *HSK Dynamic Composition Corpus*[18] (HSK *dòngtài zuòwén yǔliàokù* HSK 动态作文语料库; Zhang 2003) is one of the most frequently cited L2 Chinese learner corpus (Xu 2019). It is freely available for public use. The Corpus Version 1.0 launched in 2006 has been recently upgraded to Corpus Version 2.0. It comprises 11,569 essays with 4,24 million characters produced by L2 Chinese learners who took the HSK Chinese language proficiency test between 1992 and 2005. The type of essays is principally narrative or argumentative. Nearly 90% of contributors to the corpus are from Asia, and 64% of the data is gathered from Korean and Japanese learners. The corpus also includes scanned copies of learners' original compositions, learners' rich metadata, the results of each section of the HSK test (listening, writing, speaking), and the HSK examination overall score. The corpus is POS and error-tagged at levels of punctuation, character, lexicon, grammar, and discourse. Discourse annotation is also included. An important new feature in the Corpus Version 2.0 is that users can develop graphs for statistical analyses and add or edit error annotations to the corpus.

A special case is the *TOCFL Learner Corpus*[19] (Chang 2013): it is the first learner corpus of traditional Chinese characters which includes grammatical error annotation (Lee *et al.* 2018). It collects written essays completed by students from 42 different L1 backgrounds (mainly from Asian regions and English-speaking countries) who took the TOCFL test[20] since 2016. Metadata provide information on learners' L1 backgrounds, CEFR level, as well as information relating to the text genre, text function, text length, and TOCFL test score. As for the corpus size, it consists of 5,092 essays, 1,740,000 characters and 1,140,000 words. 33,835 grammatical errors and their corresponding corrections have been manually added by Chinese native-speaking annotators which were specifically trained to apply annotation guidelines provided by the research team (Lee *et al.* 2018). The corpus is available online to support future research.

Another large-scale corpus is the *Jinan Chinese Learner Corpus* (JCLC; Wang *et al.* 2015), which collects written texts produced by university students at beginner, intermediate, and advanced levels with 59 different L1s. It currently contains 5,91 million Chinese characters across 8,739 texts,

[18] The *HSK Dynamic Composition Corpus (Version 2.0)* and related information can be found at: http://yuyanziyuan.blcu.edu.cn/en/info/1043/1501.htm (last access: 15 September 2022).

[19] http://nlp.ee.ncu.edu.tw/resource/tocfl.html (last access: 15 September 2022).

[20] The *Test of Chinese as a Foreign Language* (TOCFL) is the Mandarin language proficiency test adopted in Taiwan, which cannot be obtained in mainland China.

accompanied by a rich set of metadata. JCLC is an ongoing project. New data continues to be collected and added to the corpus. At present, it has not been annotated yet. The corpus is freely available upon request to the research team.

The *Yet Another Chinese Learner Corpus* (YACLC; Wang *et al.* 2021), to the best of our knowledge, is one of the most recent Chinese learner corpus projects. Researchers collected 441,670 sentences from 29,595 essays, provided by approximately 50,000 learners. After the data cleaning, the team worked on 32,124 sentences from 2,421 essays. The corpus presents a multilayered annotation: for each sentence, annotators (183 specifically trained individuals to annotate the corpus) provided a variety of revisions consisting of grammatical and fluency corrections. Grammatical correction verifies whether sentences are conformed to grammar, while fluency correction verifies whether sentences are fluent or not and native sounding (Wang *et al.* 2021). The YACLC corpus is available online.

Some small-scale written corpora were compiled specifically to analyze Chinese characters. An example is the *Chinese Character Errors Corpus* (CCEC; Teng *et al.* 2007), which is the first learner corpus collecting data to analyze learner errors in the writing of traditional characters (Xu 2019). It collects data from 124 students at beginner, intermediate, and advanced levels with 15 different L1 backgrounds. It also includes the scanned version of learners' original composition. Misspelled characters are tagged and errors are categorized into nine different groups. This corpus is not available for public use.

A similar small-scale project is the *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[21] (Zhang 2017). It includes texts written in simplified Chinese characters. The texts were tokenized and POS tagged. Similarly to the CCEC, copies of the original hand-written texts are stored along-side each entry in the corpus; misspelled characters are accompanied by error annotation.

Also worth mentioning is the *UCLA Heritage Language Learner Corpus* (Ming and Tao 2008). It is a unique collection of data from Chinese heritage learners with Chinese family background. It was developed at the University of California and contains approximately 1,000 written essays and compositions, completed as homework assignments, produced by learners at the intermediate level attending heritage Chinese classes in 2006 and 2007. The text genres are argumentative, narrative, and

[21] 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/ (last access: 15 September 2022).

descriptive. POS and error tagging are both included in the corpus and a coding system was specifically designed for heritage learner error annotation (Zhang and Tao 2018).

Spoken corpora represent a small percentage of the corpora compiled to date. Among the large-scale corpora, the *Chinese as a Second Language Spoken Corpus*[22] (Chang 2016) features prominently. It is a database of spoken data collected from the TOCFL proficiency test. The corpus includes data only from English, Japanese, and Korean learners, and it contains 450 tests with 773,000 characters (Zhang and Tao 2018).

Another large-scale spoken corpus is the *Spontaneous Chinese Learner Speech Corpus* (Wu and Shih 2014), which was compiled at the University of Illinois at Urbana-Champaign. It collects 185 audio and video recordings, which were gathered during Chinese speech training classes from 2004 to 2009. 11 Chinese language teachers, 11 Korean-speaking learners, 23 English-speaking learners, and 86 Chinese heritage learners took part in this project as speakers; they completed two different oral open-ended tasks, each of which was designed to fit in a 50-minute class. The data were transcribed through a transcription website and, according to Wu and Shih (2014), this corpus is a rich resource with speech samples for various research topics.

The *Multilingual Corpus of Second Language Speech* (MuSSeL; Rubio *et al.* 2021) deserves a separate discussion. It is being developed by researchers at the University of Utah's Second Language Teaching and Research Center. Once completed, this large-scale corpus will include samples from three learning contexts (child classroom, adult classroom, and adult post-immersion) across six languages: Chinese, French, German, Portuguese, Russian, and Spanish. The corpus provides users with a varied set of transcribed and tagged L2 speech samples as well as access to the original MP3 recordings. The transcripts are tagged according to the CHAT protocols established by CHILDES (MacWhinney 2000). The corpus is searchable using various metadata filters, e.g., language, age group, gender, learning context, topic, and proficiency level.

Small-scale spoken corpora include the *Spoken Chinese Corpus of Informal Interaction* (Li 2021), compiled at the Massey University in New Zealand. It collects spoken data from English-speaking intermediate and advanced learners from New Zealand and Australia. The data are collected from informal conversation between 14 learners of L2 Chinese and Chinese native speakers. Another small-scale spoken corpus is the *COPA corpus*[23] (Zhang 2009) which collects speech recordings from 120 college

[22] http://140.122.83.243/mp3c (last access: 14 September 2022).

[23] Information about the *COPA corpus* and the link to it can be found at: https://www.clarin.eu/resource-families/L2-corpora (last access: 14 September 2022).

students learning Chinese in Hong Kong. Corpus data are gathered from conversation with Chinese native speakers. This corpus is included in the *SLABank* database,[24] which is a component of *TalkBank*,[25] an online platform committed to providing corpora to support the study of second language acquisition.

Similarly, the small-scale *HKPU corpus*[26] (Chan *et al.* 2013) is included in the *SLABank* collection and is available via *TalkBank*. It contains speech recordings of 20 college students learning Chinese in Hong Kong collected through oral interviews.

Finally, the smallest group is multimodal corpora. The first of this kind is the *Guangwai-Lancaster Chinese Learner Corpus* (GWLCLC), compiled at the Guangdong University of Foreign Studies (GDUFS) in China in collaboration with Lancaster University. It is a collection of written and spoken data produced by 886 learners from 80 different countries studying at GDUFS. Learners are grouped into beginner, intermediate, and advanced proficiency levels, according to the HSK Chinese Proficiency Test score standards. The corpus consists of 1,664,237 tokens and 1,289,060 words; it is POS and error tagged. Both spoken and written data were collected during exams at GDUFS. Written tasks consist of essays on a given topic, whereas oral tasks comprise informal conversations between native and non-native speakers of Chinese. Learners' data are also accompanied by metadata. It is a balanced corpus that has often been used by researchers to explore theoretical and practical issues on the acquisition of L2 Chinese (see Chen and Xu 2019; Gablasova 2021; Xu *et al.* 2019). This corpus is available online on *Sketch Engine*.[27]

The *Mandarin Interlanguage Corpus* (MIC; Tsang and Yeung 2012) is a small-scale learner corpus compiled at the University of Hong Kong which collects written and spoken data from pre-intermediate to intermediate Chinese learners with different L1s. Data were collected in the form of coursework and examinations, and the corpus contains approximately 50,000 characters and 60 hours of oral output. The MIC annotates the errors at the character level. Unfortunately, it is not available online.

[24] 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/ (last access: 14 September 2022).

[25] *TalkBank* is a project organized at the Carnegie Mellon University committed to foster fundamental research in the study of spoken communication. Data in *TalkBank* are provided by researchers working in over 34 languages internationally. Further information is searchable at: https://talkbank.org/ (last access: 14 September 2022).

[26] The *HKPU corpus* and related information can be found at:
https://slabank.talkbank.org/access/Mandarin/HKPU.html (last access: 14 September 2022).

[27] Information on the *Guangwai-Lancaster Chinese Learner Corpus* (GWLCLC) can be found at:
https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fguangwai (last access: 26 June 2022).

Other Chinese learner corpora have been compiled with a specific aim in mind, such as the study of a particular construction or word. For instance, the *Bimodal Italian Learner Corpus of Chinese* (BILCC; Iurato forthcoming) was specifically compiled to explore the pragmalinguistic knowledge of the Chinese *shì* 是...*de* 的 cleft construction by L1 Italian learners. It is a target-oriented small-scale learner corpus that includes written and spoken data (including both recordings of speech and related transcriptions) from 103 beginner, intermediate and advanced L1 Italian learners enrolled at Ca' Foscari University of Venice. A control corpus containing written and spoken data from 30 L1 Chinese speakers was also included. The learner corpus consists of 57,600 Chinese characters and 25 hours of recordings of speech, whereas the control corpus includes 31,000 Chinese characters and 7 hours of recordings of speech. BILCC was assembled according to strict specific design criteria and it is the result of theoretically motivated open-ended tasks. All data in BILCC are accompanied by a rich set of metadata. A target-oriented error taxonomy was developed to manually annotate the grammatical errors; a pragmatic annotation was also added to detect the inappropriate use of the pragmatic functions of the *shì...de* cleft construction. A similar project is ACHIEVE,[28] directed by Bianca Basciano (Ca' Foscari University of Venice) and funded by the Italian Ministry of University and Research (MUR), which is aimed at compiling a corpus to explore the acquisition of Chinese resultative verbal complexes by L1 Italian learners. This corpus has not yet been compiled; however, similarly to BILCC, it will be made freely available once the compilation and annotation processes are completed.

Table 4. summarizes the main descriptive features of some of the corpora illustrated above which we have selected for the sake of their representativeness. Drawing inspiration from the entries in the *Learner corpora around the world* database used for the categorization of learner corpora, we have listed L2 Chinese learner corpora according to the following entries:

a. corpus project name;
b. learners' Chinese language proficiency;
c. learners' L1 background;
d. size of the learner corpus;
e. typology of the data collection;
f. text type;
g. information provided on the learners' metadata;
h. type of annotation.

[28] https://pric.unive.it/projects/achieve/home

| Corpus project | Chinese L2 proficiency | L1 | Size | Data collection | Text types | Learner metadata | Annotation |
|---|---|---|---|---|---|---|---|
| L2 Chinese Interlanguage Corpus (Chu and Chen 1993) | HSK language proficiency levels | Students from 96 different countries | 3,528,988 Chinese characters | Cross-sectional | Written essays | Biographic information, Chinese learning experience | POS; error |
| HSK Dynamic Composition Corpus (Zhang 2003) | HSK language proficiency levels | Students from Asian countries and English-speaking regions | 4,24 million Chinese characters | Cross-sectional | Narrative and argumentative written essays | Gender, age, country, L1 background, HSK total score, HSK awarded certificate, results of HSK tests. | POS; error; discourse-pragmatic |
| TOCFL Learner Corpus (Chang 2013) | TOCFL language proficiency levels | Students from Asian countries and English-speaking regions (42 L1 backgrounds) | 1,740,000 Chinese characters and 1,140,000 words | Cross-sectional | Written essays | L1, text function, text length, TOCFL score | Error |
| Jinan Chinese Learner Corpus (JCLC; Wang *et al.* 2015) | Beginner, intermediate and advanced, according to the length of study | Students from 59 different nationalities | 5,91 million Chinese characters | Cross-sectional | Written exams and assignments | Gender, age, educational level, L1, other acquired languages, proficiency level, | - |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | length of Chinese language study | |
| Yet Another Chinese Learner Corpus (YACLC; Wang *et al.* 2021) | 50,000 learners; HSK language proficiency levels | Learners from 60 different countries studying in China | 32,124 sentences | Cross-sectional | Written essays | - | Error |
| UCLA Heritage Language Learner Corpus (Ming and Tao 2008) | Intermediate level | Chinese heritage learners with Chinese family backgrounds | 200,000 Chinese characters | Cross-sectional | Argumentative, narrative, descriptive written essays and compositions completed as homework assignments | - | POS; error |
| Spontaneous Chinese Learner Speech Corpus (Wu and Shih 2014) | Third and fourth-year Chinese language classes + L1 Chinese native speakers (control group) | Korean, English-speaking learners, and Chinese heritage learners | 185 hours of audio and video recordings | Cross-sectional | Argumentative and narrative oral tasks | - | - |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Guangwai-Lancaster Chinese Learner Corpus[29] | HSK language proficiency levels | Learners from 80 different countries studying in China | 1,2 million words | Cross-sectional | Argumentative written essays and descriptive oral tasks from exams and tutorial sessions | Nationality, L1, gender, proficiency, test score | POS; error |
| Mandarin Interlanguage Corpus (MIC; Tsang and Yeung 2012) | Pre-intermediate and intermediate level, based on 2-year certificate course on mandarin Chinese at a tertiary institution in HK | English, French, Spanish, German, Dutch, Japanese, Korean, Thai, Indonesian, Tamil | 50,000 Chinese characters and 60 hours of recordings of speech | Cross-sectional | Narrative written essays and oral presentations from end-of-course examination | Nationality, age, L1, other languages spoken, length of Chinese study | POS; error |
| Bimodal Italian Learner Corpus of Chinese (BILCC; Iurato forthcoming) | HSK language proficiency levels | 103 learners studying at Ca' Foscari University of Venice + 30 L1 Chinese native speakers (control corpus) | 67,600 Chinese characters and 25 hours of recordings + control corpus consisting of 31,000 Chinese characters and 7 | Cross-sectional | Open-ended tasks; written essays; role-plays, interviews | Age, L1, educational level, other acquired languages, proficiency level, length of Chinese language study, | Error; discourse-pragmatic |

[29] https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fguangwai (last access: 26 June 2022).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | hours of recordings of speech | | | length of stay in China | |

*Table 4.* Most representative Chinese learner corpora and related main features

The corpora listed in Table 4. cover three modes (oral, written and multimodal), two sizes (small-scale and large-scale), one type of data collection (cross-sectional), and three types of annotation (POS annotation, error annotation, discourse-pragmatic annotation).

## 6. Future directions and conclusions

CLCR has achieved a great deal in the short course of 20 years (Zhang and Tao 2018; Xu 2019; Iurato 2022). Nevertheless, we argue that this field has to face some challenges, since there are still many gaps concerning the learner corpus design, analysis of learner corpora, the direct and indirect uses of learner corpora from a pedagogical perspective, and the interactions between CLCR and SLA.

We have identified four main challenges concerning the learner corpus design in CLCR. First, given the scarcity of corpora with data from learners whose L1 is other than English or Asian languages, we hope that in the future (more) digitalized Chinese corpora will be created including data from learners whose L1 is a non-English European language. The availability of such corpora will facilitate research on L2 Chinese acquisition in those contexts where there is a progressive increase in L2 Chinese language learners. Second, we believe that future studies should consider the creation of different modes of learner corpora (academic, multimodal, multilingual, multi-layered), as they would allow more research across a wider spectrum of learner development. Third, we think that the compilation of publicly available Chinese longitudinal corpora should be stimulated, considering that the research conducted so far allows us to gain insight into the learner acquisition process only from a cross-sectional rather than a developmental perspective. We hope that in the future studies based on longitudinal corpora may offer more accurate and broader perspectives on learner variations. Fourth, we strongly argue that particular attention should also be paid to the inclusion of target hypotheses in the error annotation process in future CLCR research, since 1) the identification of errors requires linguistic and extra-linguistic contexts, and 2) the categorization of errors varies according to the target hypothesis identified for specific research purposes (Díez-Bedmar 2015; Lüdeling and Hirschmann 2015).

Our study revealed that another major issue in CLCR is the lack of consistency in the results obtained from corpus analysis. This is caused by the fact that scholars in this field conduct their research using only the available corpora to which they have access, since many L2 Chinese learner corpora are not available for public use (see Section 5), as our investigation reveals. In accordance with Zhang and Tao's position (2018), we believe that due to the significant differences in corpora sizes, learners' proficiency levels, contexts and procedures of data collection, types of tasks, and so forth, different studies on the same topic may generate not only different, but even contradictory results. We also think that this lack of homogeneous and coherent synthesis in research results makes it difficult to outline a global picture of CSL/CFL learners' development. Moreover, it hinders the use of these corpora in the development of teaching materials, since it causes difficulties for those who want to obtain information about the acquisition of particular linguistic features.

As for the so-called 'indirect uses' of corpora (Zhang and Tao 2018: 57), i.e., the application of corpus data, it must unfortunately be noted that Chinese corpora currently have little impact on the creation of Chinese coursebooks. Learner corpora can be useful to guide the writing of textbooks, pedagogical materials, and dictionaries. The rich understandings gained from LCR research, including the acquisition orders and developmental patterns of different linguistic features, the common errors learners tend to produce at different proficiency levels, and crosslinguistic influences, should all be considered by researchers, CSL/CFL textbook writers, and pedagogical materials developers (Zhang and Tao 2018). On the other hand, it is important to stress that the direct use of learner corpora in classrooms by teachers and learners might support the traditional instructional approaches for vocabulary, grammar, and language use (Crosthwaite 2012). We therefore believe it is essential to encourage Chinese language teachers to use L2 Chinese learner corpora as useful tools to support the teaching of L2 Chinese.

There are, in addition, obvious benefits of using learner corpora for the design of language testing, not simply for placement purposes, but also for proficiency assessment (Callies and Götz 2015). This also applies to the Chinese context, in which learner corpora may be used to inform CSL/CFL assessment and establish the benchmarks of students' language proficiency levels in both writing and speaking (Zhang and Tao 2018), since we found that in current L2 Chinese learner corpora learner proficiency assessment does not adopt homogeneous criteria and sometimes the proficiency is even inadequately assessed.

Finally, another important challenge concerns the interaction between CLCR and SLA. Similarly to general LCR, we believe that more collaboration between researchers in CLCR and SLA is needed. Although the convergence of (C)LCR and SLA paradigms has not yet been achieved (Granger 2021; Iurato 2022), both disciplines could reciprocally benefit from important advantages, if there was a concrete integration between the application of the methodological framework of LCR and the implementation of the theoretical interpretation of data of SLA research in the design of acquisitional studies.

In this paper we have demonstrated that in CLCR several areas remain in need of further development. However, it is necessary to be aware that the computational tools used for English cannot be readily adopted in Chinese, given the unique characteristics of Chinese language and script. CLCR therefore requires greater effort and longer working time to achieve the results already achieved in parallel areas, e.g., English LCR. We hope that in the future scholars in the fields of CLCR, linguistics, and computational linguistics will collaborate by joining forces, so that L2 Chinese learner language development will be explored more effectively.

## References

Alonso-Ramos, Margarita. 2016. “Spanish learner corpus research. Achievements and challenges.” In: *Spanish Learner Corpus Research: Current trends and future perspectives*, edited by Margarita Alonso-Ramos, 3-32. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.78.01alo

Callies, Marcus. 2015. “Using Corpora in Language Testing and Assessment: Current Practice and Future Challenges.” In: *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*, edited by Erik Castello, Katherine Ackerley and Francesca Coccetta, 21-35. Bristol: Peter Lang.

Callies, Marcus, María Belén Díez-Bedmar and Ekaterina Zaytseva. 2014. “Using learner corpora for testing and assessing L2 proficiency.” In: *Measuring L2 proficiency: Perspectives from SLA*, edited by Pascale Leclercq, Heather Hilton and Amanda Edmonds, 71-90. Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781783092291

Callies, Marcus and Ekaterina Zaytseva. 2013. “The Corpus of Academic Learner English (CALE) – A new resource for the study and assessment of advanced language proficiency.” In: *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1*, edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier, 49-59. Louvain-la-Neuve: Presses universitaires de Louvain.

Callies, Marcus and Sandra Götz. 2015. “Learner Corpora in Language Testing and Assessment. Prospect and Challenges.” In: *Learner Corpora in Language Testing and Assessment*, edited by Marcus Callies and Sandra Götz, 1-9. Amsterdam and Philadelphia: John Benjamins. https://doi.org/10.1075/scl.70.001int

Chan, Angel, Zheng H. Feng, Wei C. Yang (2013, June). *A new multimedia shared L2 spoken Mandarin Chinese corpus: construction and linguistic analyses*. Paper presented at the *Annual Meeting of the International Association of Chinese Linguistics (IACL)*, 21. Taiwan.

Chang, Lee (2013). “TOCFL Zuowen yuliaoku de jianzhi yu yingyong” [Compilation and applications of the TOCFL Composition Corpus]. In: *Di’er jie Hanyu zhongjieyu yuliaoku jianshe yu yingyong xueshu taolunhui lunwen xuanji* [Selected papers from the 2nd International Conference on the Construction and Applications of Chinese Learner Corpora], edited by Cui Xiang and Baolin Zhang, 141-152. Beijing: Beijing Language and Culture University Press.

Chang, Lee (2016). "TOCFL xuexizhe yuliaoku de pianwu biaoji" [Error annotation for the TOCFL learner corpus]. In: *Disanjie Hanyu zhongjie yuliaoku jianshe yu yingong guoji xueshu taolunhui lunwen xuanji* [Selected papers from the 3rd International Conference on the Construction and Applications of Chinese Learner Corpora], edited by Xiao Lin, Xinfeng Xiao and Baolin Zhang, 131-159. Beijing: Beijing Language and Culture University Press.

Chen, Heng and Hai Xu. 2019. "Quantitative linguistics approach to interlanguage development: a study based on the Guangwai-Lancaster Chinese Learner Corpus." *Lingua* 230: 102736-102751. https://doi.org/10.1016/j.lingua.2019.102736

Chu, Chengde and Xiao Chen. 1993. "Constructing a Chinese Interlanguage Corpus." *Shijie Hanyu Jiaoxue*, 7/3: 199-205.

Crosthwaite, Peter. 2012. "Learner corpus linguistics in EFL classroom." *PASAA - A Journal of Language Teaching and Learning in Thailand* 44: 133-147.

Cui, Xiaojun (2005). "Oumei xuesheng hanyu jiecixide de tedian ji pianwu fenxi" [The acquisition of Chinese prepositions by European and American learners and analysis of their errors]. *Shijie Hanyu Jiaoxue* 19/3: 83-95.

Dagneaux, Estelle, Sharon Denness and Sylviane Granger. 1998. "Computer-aided Error Analysis." *System* 26: 163-174. https://doi.org/10.1016/S0346-251X(98)00001-3

Díez-Bedmar, María Belén. 2015. "Dealing with Errors in Learner Corpora to Describe, Teach and Assess EFL Writing: Focus on Article Use." In: *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*, edited by Erik Castello, Katherine Ackerley and Francesca Coccetta, 37-69. Bristol: Peter Lang.

Duff, Patricia, Tim Anderson, Roma Ilnyckyj, Ella VanGaya, Rachel Tianxuan Wang and Elliott Yates. 2013. *Learning Chinese: Linguistic, Sociocultural, and Narrative Perspectives*. Berlin: De Gruyter. https://doi.org/10.1515/9781934078778

Fernández, Julieta and Shelley Staples. 2021. "Pragmatic Approaches." In: *The Routledge Handbook of Second Language Acquisition and Corpora*, edited by Nicole Tracy-Ventura and Magali Paquot, 240-251. New York and London: Routledge.

Gablasova, Dana. 2021. "Corpora for second language assessments." In: *The Routledge Handbook of Second Language Acquisition and Language Testing*, edited by Paula Winke and Tineke Brunfaut, 45-53. New York and London: Routledge. https://doi.org/10.4324/9781351034784

Gilquin, Gaëtanelle. 2021. "Combining Learner Corpora and Experimental Methods." In: *The Routledge Handbook of Second Language Acquisition and Corpora,* edited by Nicole Tracy-Ventura and Magali Paquot, 1-13. New York and London: Routledge. https://doi.org/10.4324/9781351034784

Gilquin, Gaëtanelle and Stefan Th. Gries. 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

Glaznieks, Aivars, Jennifer-Carmen Frey, Maria Stopfner, Lorenzo Zanasi and Lionel Nicolas. 2022. “LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English.” *International Journal of Learner Corpus Research* 8/1: 97-120. https://doi.org/10.1075/ijlcr.21004.gla.

Granger, Sylviane. 1996. “From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora.” In: *Languages in Contrast. Text-based cross-linguistic studies,* edited by Karin Aijmer, Bengt Altenberg and Mats Johansson, 37-51. Lund: Lund University Press.

Granger, Sylviane. 2002. “A Bird’s-Eye View of Learner Corpus Research.” In: *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching,* edited by Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, 3-33. Amsterdam: John Benjiamins.

Granger, Sylviane. 2015. “Contrastive interlanguage analysis: A reappraisal.” *International Journal of Learner Corpus Research* 1/1: 7-24. doi:10.1075/ijlcr.1.1.01gra.

Granger, Sylviane. 2021. “Commentary: Have Learner Corpus Research and Second Language Acquisition Finally Met?” In: *Learner Corpus Research Meets Second Language Acquisition*, edited by Bert Le Bruyn and Magali Paquot, 258-273. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108674577.012

Granger, Sylviane, Gaëtanelle Gilquin and Fanny Meunier. 2015. *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

Granger, Sylviane, Gaëtanelle Gilquin and Fanny Meunier. 2015. “Introduction: Learner Corpus Research – past, present and future.” In: *The Cambridge Handbook of Learner Corpus Research,* edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier, 1-5. Cambridge: Cambridge University Press.

Gudmestad, Aarnes. 2021. “Variationist Approaches.” In: *The Routledge Handbook of Second Language Acquisition and Corpora*, edited by Nicole Tracy-Ventura and Magali Paquot, 228-239. New York and London: Routledge.

Istvanova, Maria. 2021. “Chinese Learner Corpora and Creation of Slovak Learner Corpus of Chinese.” *The Silk Road. Language and Culture* 2021: 48-55.

Iurato, Alessia. 2021a, October. *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, Inland Norway University of Applied Sciences, Norway.

Iurato, Alessia. 2021b, July. *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), George Washington University, USA.

Iurato, Alessia. 2022. “Learner Corpus Research meets Chinese as a Second Language Acquisition: Achievements and Challenges.” *Annali di Ca’ Foscari. Serie Orientale* 58/1: [1-34] 709-742.

Iurato, Alessia. Forthcoming. “Designing and compiling the written sub-corpus of the Bimodal Italian Learner Corpus of Chinese (BILCC): Methodological issues.” In: *Chinese Linguistics in Italy*, edited by Serena Zuccheri. Bologna: Bologna University Press.

Leclercq, Pascale and Amanda Edmonds. 2014. “How to assess L2 proficiency? An overview of proficiency assessment research.” In: *Measuring L2 proficiency: Perspectives from SLA*, edited by Pascale Leclercq, Heather Hilton and Amanda Edmonds, 3-23. Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781783092291

Lee, Lung-Hao, Yuen-Hsien Tseng and Li-Ping Chang. 2018. “Building a TOFCL Learner Corpus for Chinese Grammatical Error Diagnosis.” *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2298-2304. Miyazaki: European Language Resource Association.

Lee, Lung-Hao, Yuen-Hsien Tseng and Li-Ping Chang. 2019. “Resources and Evaluations of Automated Chinese Error Diagnosis for Language Learners.” In: *Computational and Corpus Approaches to Chinese Language Learning,* edited by Xiaofei Lu and Berlin Chen, 235-252. Singapore: Springer. DOI: 10.1007/978-981-13-3570-9_12

Li, Cixin, Angel Tu and He Zhao (2020). “Jiyu HSK. Zhenti kaocha de Hanyu fuju xide yanjiu—yi HSK si ji yuyandian de fuju weili” [A Study of the Acquisition of Chinese Complex Sentences Based on HSK Test Papers: The Case of Chinese Complex Sentences in Level-4 HSK]. *Yunnan shifan daxue xuebao* 18/5: 12-18. DOI:10.16802/j.cnki.ynsddw.2020.05.006.

Li, Lin. 2021. *A spoken Chinese corpus: Development, description, and application in L2 studies* [Unpublished PhD diss.]. Manawatū: Massey University. https://github.com/blculyn

Liu, Tao and Zisi Ming. 2015. “Jiyu huayu gongneng de Hanguo liuxuesheng ‘shi…de’ ju pianwu” [Error analysis of the shì…de sentence by Korean learners of L2 Chinese]. *Zhongguo Kuangye Daxue Xuebao (Shehui Kexue Ban)* 3: 106–110.

Lozano, Cristóbal. 2021. “Generative Approaches.” In: *The Routledge Handbook of Second Language Acquisition and Corpora*, edited by Nicole Tracy-Ventura and Magali Paquot, 213-227. New York and London: Routledge.

Lozano, Cristóbal and Amaya Mendikoextea. 2013. “Learner corpora and second language acquisition: the design and collection of CEDEL2.” In: *Automatic Treatment and Analysis of Learner Corpus Data,* edited by Ana Díaz-Negrillo, Nicolas Ballier and Paul Thompson, 65-100. Amsterdam and Philadeplhia: John Benjamins. DOI: 10.1075/scl.59.06loz

Lu, Xiaofei and Berlin Chen. 2019. “Computational and Corpus Approaches to Chinese Language Learning: An Introduction.” In: *Computational and Corpus Approaches to Chinese Language Learning*, edited by Xiaofei Lu and Berlin Chen, 3-11. Singapore: Springer. DOI: 10.1007/978-981-13-3570-9_1

Lu Shuxiang. 1994. “Waiguoren xue Hanyu de yufa pianwu fenxi” [An analysis of grammatical errors of foreign learners of Chinese]. *Yuyan jiaoxue yu yanjiu* 1: 49-68.

Lü Bisong. 1993. "Lun han zhongjieyu de yanjiu" [On the study of Chinese Interlanguage]. *Yuyan wenzi yinyong* 2: 27-31.

Lüdeling, Anke and Hagen Hirschmann. 2015. "Error annotation systems." In: *The Cambridge Handbook of Learner Corpus Research*, edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier, 135-157. Cambridge: Cambridge University Press.

MacWhinney, Brian. 2000. "The CHILDES Project: Tools for Analyzing Talk." *Child Language Teaching and Therapy* 8/2. DOI: 10.1177/026565909200800211.

Mendikoextea, Amaya and Cristóbal Lozano. 2018. "From Corpora to Experiments: Methodological Triangulation in the Study of Word Order at the Interfaces in Adult Late Bilinguals (L2 learners)." *J Psycolinguistic Res* 47/4: 871-898. https://doi.org/10.1007/s10936-018-9560-0.

Meunier, Fanny. 2016. "Introduction to the LONGDALE project." In: *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*, edited by Erik Castello, Katherine Ackerley and Francesca Coccetta, 123-126. Berlin: Peter Lang. DOI :10.3726/978-3-0351-0736-4/17

Meunier, Fanny. 2021. "Introduction to Learner Corpus Research." In: *The Routledge Handbook of Second Language Acquisition and Corpora*, edited by Nicole Tracy-Ventura and Magali Paquot, 23-36. New York and London: Routledge. https://doi.org/10.4324/9781351137904

Ming, Tao and Hongyin Tao. 2008. "Developing a Chinese Heritage Language Corpus: Issues and a Preliminary Report." In: *Chinese as a Heritage Language: Fostering Rooted World Citizenry*, edited by Agnes W. He and Yun Xiao, 167-178. Honolulu, HI: National Foreign Language Resource Center, University of Hawai'I.

Paquot, Magali, Damien De Meyere, Signe O. Ebeling, Hilde Hasselgard, Tove Larsson, Natalie J. Laso, Larry Valentin, Isabel Verdaguer and Sanne van Vuuren (forthcoming). "The Varieties of English for Specific Purposes dAtabase (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing." *Research in Corpus Linguistics*.

Qu, Mingxian. (2013). "Jiyu 'HSK Dongtai Zuowen Yuliaoku' de 'gei' zi ju xide yanjiu [Acquisition of the "gei" sentences based on the "HSK Dynamic Composition Corpus"]. In: *Di'er jie Hanyu zhongjieyu yuliaoku jianshe yu yingyong tolunhui lunwen xuanji* [Selected papers from the 2nd International Conference on the Construction and Applications of Chinese Learner Corpora], edited by Cui Xiang and Baolin Zhang, 201-211. Beijing: Beijing Language and Culture University Press.

Romagnoli, Chiara and Sergio Conti. 2021. *La Lingua Cinese in Italia. Studi su Didattica e Acquisizione*. Roma: Roma Tre Press.

Römer, Ute, Audrey Roberson, Matthew Brook O'Donnell and Nick Ellis. 2014. "Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions." *ICAME Journal* 38: 115-135. https://doi.org/10.2478/icame-2014-0006

Rubio, Fernando, Elnaz Kia and Jane F. Hacking. 2021. Multilingual Corpus of Second Language Speech (MuSSeL). https://l2trec.utah.edu/learner-corpora/mussel/.

Shi Jianxin. (1998). "Waiguo liuxuesheng 22 lei xiandai Hanyu jushi de xide sjunxu yanjiu" [Foreign students' acquisition order of 22 modern Chinese sentence structures]. *Shijie Hanyu Jiaoxue* 46/4: 77-98.

Spina, Stefania and Anna Siyanova-Chanturia. In preparation. *The Longitudinal Corpus of Chinese Learners of Italian*.

Teng Siling, He Deng, Xiaoli Wang and Ping Feihong (2007). "Huayuwen xuexizhe hanzi pianwu shuju ziliaoku jianli ji pianwu leixing fenxi" [The Construction of Chinese Learners' Character Writing Error Databse and the Analysis of Error Types]. *Proceedings of the 2007 National Linguistics Conference*, 313–325. Taiwan: National Taiwan University Press.

Tono, Yukio. 2003. "Learner Corpora: Design, Development and Applications." In: *Proceedings of the Corpus Linguistics 2003 Conference. UCREL 16,* edited by Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery, 800-809. Lancaster: Lancaster University Press.

Tracy-Ventura, Nicole. And Flores Myles. 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. https://doi.org/10.1075/ijlcr.1.1.03tra

Tracy-Ventura, Nicole and Magali Paquot. 2021. "Second Language Acquisition and Corpora. An Overview." In: *The Routledge Handbook of Second Language Acquisition and Corpora*, edited by Nicole Tracy-Ventura and Magali Paquot, 1-8. New York and London: Routledge.

Tracy-Ventura, Nicole, Rosamond Mitchell and Kevin McManus. 2016. "The LANGSNAP longitudinal learner corpus." In: *Spanish Learner Corpus Research: Current trends and future perspectives*, edited by Margarita Alonso-Ramos, 117-142. Amsterdam and Philadelphia: John Benjamins. https://doi.org/10.1075/scl.78.05tra

Tsang, Wai-Ian and Yuk Yeung. 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

van Rooy, Bertus. 2015. "Annotating learner corpora." In: *The Cambridge Handbook of Learner Corpus Research,* edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier, 79-105. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.005

Wang, Maolin, Shervin Malmasi and Mingxuan Huang. 2015. "The Jinan Chinese Learner Corpus." *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 118-123. Denver, Colorado: Association for Computational Linguistics.

Wang, Yingying, Cunliang Kong, Liner Yang, Yijun Wang, Xiaorong Lu, Renfen Hu, Shan He, Zhenghao Liu, Yun Chen, Erhong Yang and Maosong Sun. 2021. "YACLC: A Chinese Learner Corpus with Multidimensional Annotation." *Computer Science – Computation and Language* 1-5. arXiv:2112.15043. https://doi.org/10.48550/arXiv.2112.15043

Wen, Xiaohong. 2019. “Research in second language acquisition of Chinese. An introduction.” In: *Studies on Learning and Teaching Chinese as a Second Language*, edited by Xiaohong Wen and Xin Jiang, 1-13. New York and London: Routledge.

Wu, Chen-huei and Chilin Shih. 2014. “A Design of the Spontaneous Chinese Learner Speech Corpus.” *Learner Corpus Studies in Asia and the World* 2: 115-124. https://doi.org/10.24546/81006694.

Wulff, Stefanie. 2021. “Usage-based Approaches.” In: *The Routledge Handbook of Second Language Acquisition and Corpora*, edited by Nicole Tracy-Ventura and Magali Paquot, 175-188. New York and London: Routledge.

Xie Fu. 2010. “Jiyu yuliaoku de liuxuesheng ‘shi...de’ ju xide yanjiu” [The acquisition of the “shì...de” sentence by foreign students: A study based on the HSK Interlanguage corpus]. *Yuyan Jiaoxue yu Yanjiu* 2: 48-55.

Xu, Jiajin. 2019. “The corpus approach to the teaching and learning of Chinese as an L1 and an L2 in retrospect.” In: *Computational and Corpus Approaches to Chinese Language Learning*, edited by Xiaofei Lu and Berlin Chen, 33-53. Singapore: Springer.

Xu, Hu, Xiaofei Lu and Vaclav Brezina. 2019. “Acquisition of the Chinese Particle *le* by L2 learners: A corpus-Based approach.” In: *Computational and Corpus Approaches to Chinese Language Learning*, edited by Xiaofei Lu and Berlin Chen, 197-216. Singapore: Springer. DOI: 10.1007/978-981-13-3570-9_10

Yang Pei. 2016. “Jiyu HSK dongtai zuowen yuliaoku de ‘shi’ ziji xide yanjiu—yi hanguo liuxuesheng weili” [The acquisition of “shi” sentence by Korean students: A study based on HSK dynamic composition corpus]. *Xiandai Yuwen. Modern Chinese* 10: 130-132.

Zhang Baolin. 2003. “HSK Dongtai Zuowen Yuliaoku Jianjie” [Introducing Chinese Proficiency Test Dynamic Essay Corpus]. *Ceshi Yanjiu* 1/4: 37-38.

Zhang, Jie. 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, Jie and Hongyin Tao. 2018. “Corpus-Based Research in Chinese as a Second Language.” In: *The Routledge Handbook of Chinese Second Language Acquisition,* edited by Chuanren Ke 48-62. New York and London: Routledge.

Zhang Ruoying. 2017. “Hanyu zhongjieyu yuliaoku zhong de hanzi pianwu chuli yanjiu” [The Character Errors in Chinese Interlanguage Corpora]. *Yuliaoku Yuyanxue* 3/2: 50-59.

Zhang, Ying. 2009. *A Tutor for Learning Chinese Sounds through Pinyin* [Unpublished doctoral dissertation]. Pittsburgh: Carnegie Mellon University.

Zhang Zaoxing. 2016. “Liuxuesheng “shi...de” juxing de xidexing pianwu” [The acquisition of “shi...de” construction by foreign learners of L2 Chinese]. *Xiamen Ligong Xueyuan Xuebao* 24/2: 48–53.

Alessia Iurato is a PhD Candidate in Chinese Linguistics and Learner Corpus Linguistics. She is attending a Joint Doctoral Program between the Department of Asian and North-African studies at Università Ca' Foscari Venezia in Italy and the Faculty of Linguistics and Literature at Universität Bremen in Germany. She is currently writing her PhD dissertation titled "The Acquisition of the Chinese 是 shì...的 de cleft construction by L1 Italian learners: Triangulating corpus and experimental data." She received the "2022 HICCS Graduate Student Award" for the best research paper from the scientific committee of the Hawai'i International Conference on Chinese Studies at University of Hawa'i, Mānoa. Iurato is a member of the research project of national interest 2020 (PRIN 2020), allocated by MUR (Italian Ministry of University and Research), titled "The acquisition of the resultative compounds in Chinese: Combining learner corpus and experimental data." Her research interests include Chinese as a second language acquisition, learner corpus linguistics, Chinese linguistics, syntax and pragmatics.
Alessia Iurato can be reached at: alessia.iurato@unive.it