DESIGNING AND COMPILING THE WRITTEN SUB-CORPUS OF THE BIMODAL ITALIAN LEARNER CORPUS OF CHINESE (BILCC): METHODOLOGICAL ISSUES

Alessia Iurato
Ca' Foscari University of Venice; Bremen University

1. Introduction[1]

The definition of what can be considered a 'learner corpus' has been a matter of debate since early studies in Learner Corpus Research (LCR) in the late 1980s (Meunier 2021; Tracy-Ventura and Miles 2015). When the field of LCR emerged, it aligned itself with the methodology and theoretical framework of corpus linguistics. It therefore adopted corpus linguistics' definition of what is intended by a (learner) corpus, i.e., a collection of naturally occurring, authentic, continuous, spontaneous spoken or written (learner) language samples (Callies and Götz 2015; Meunier 2021). Many corpus linguists consequently refused to consider a collection of learner data obtained through

[1] I would like to thank the two anonymous reviewers for their insightful comments and suggestions, and Serena Zuccheri for her patience and commitment as editor of this volume. I would also like to thank Bianca Basciano for reading a draft of this paper; naturally, all mistakes and errors are solely my responsibility.

elicitation methods as a 'learner corpus' because it lacked spontaneity and authenticity (Gilquin and Gries 2009; Lozano and Mendikoetxea 2013; Sinclair 2005). Later, several studies (Gilquin 2021; Norris and Ortega 2003; Tracy-Ventura and Myles 2015) showed that in the field of Second Language Acquisition (SLA), when the object of study is a rare structure, construct underrepresentation is a problematic issue that frequently occurs in general-purpose learner corpora. Tracy-Ventura and Myles (2015: 60) argued that "it is imperative to ensure the corpus contains multiple examples of the feature(s) under investigation" to meet SLA needs. Although the definition of learner corpus remains a hot topic in LCR and SLA, today researchers in both fields agree that two types of learner corpora can exist: 1) corpora as collections of authentic and natural data, i.e., "naturally occurring samples" (Granger 2012: 8); 2) corpora as the result of open-ended tasks (e.g., picture description, role-play) allowing learners to choose their own wording and whatever linguistic resources they want – or are able – to use. This type of learner data is what Granger (2012: 8) calls "clinically elicited samples".

Today, most of the available corpora collect data from L2 English learners (Gráf 2017). Although LCR has also spread to the Chinese context since the 1990s and an increasing number of L2 Chinese corpora are being compiled[2], there is a lack of L2 Chinese corpora that collect data from learners whose L1s are European languages (Iurato 2022a; Zhang and Tao 2018). In the Italian context, for instance, the growing number of students and the widespread interest in Chinese language teaching (Romagnoli and Conti 2021) have not been matched by an equally flourishing research on corpus compilation to support research on the acquisition of L2 Chinese by Italian learners (Iurato 2022a). The compilation of a learner corpus is a challenging issue due to the strict criteria that need to be observed for corpus design and data collection (Castillo Rodríguez *et al*. 2020; Dutra and Gomide 2015; Lozano 2021). This paper addresses these issues and presents the methodological steps necessary for the compilation of a written Italian

[2] The purpose of this paper is not to discuss the development of Chinese LCR, nor to provide an overview of L2 Chinese learner corpora. For an overview of these issues, see Iurato (2022a), Iurato (2022b), Xu (2019), and Zhang and Tao (2018).

learner corpus of L2 Chinese. The aim of this paper is threefold: first, illustrating the methodological stages involved in the compilation of a target-oriented corpus when the object of study is a specific (rare) structure that may be underrepresented in general-purpose learner corpora; second, describing a well-structured methodology grounded in LCR for compiling Italian learner Chinese corpora that can be reproduced in future studies, given the growth of L2 Chinese studies in the Italian context and the lack of Italian learner Chinese corpora; three, promoting the standardization of the proposed methodology, as the compilation of the presented corpus implements rigorous design principles and attempt to address some of the gaps in past LCR studies.

First, the paper will discuss the rigor and transparency required in the compilation process, explaining the criteria to be applied in the corpus design. Second, it will present the case study of a written L2 Chinese corpus specifically designed to explore the pragmalinguistic knowledge of the "*shì* 是...*de* 的 focus proper cleft" construction (Paul and Whitman 2008: 424) in L1 Italian learners' production. Here, the corpus is intended both as the result of open-ended tasks (Gilquin 2021; Tracy-Ventura and Myles 2015) and as a collection of contextualized data produced by L2 learners (Callies and Götz 2015). Corpus features, corpus typology, as well as environment, learner and task variables will be described. The collection procedure will also be explained. This corpus, presented as a case-study in the present work, constitutes the written sub-corpus of a larger project: the 'Bimodal Italian Learner Corpus of Chinese' (BILCC). It is named 'bimodal' because it collects two types of data (written and spoken data) from L1 Italian learners of L2 Chinese. The strength of such a bimodal mode corpus is that it allows us to get a deeper insight into the L1 Italian learners' pragmalinguistic knowledge and acquisition process of L2 Chinese language from different perspectives. Further details on BILCC and the design and collection of the written sub-corpus will be explored in Sections 3 and 4.

2. Design criteria

A random collection of heterogenous learner data is not a learner corpus (Granger 2012). A learner corpus is compiled according to strict

design criteria, and the usefulness of a learner corpus is directly proportional to the attention that has been paid to controlling the design criteria. These criteria primarily concern the participants and the task design, i.e., the two specific variables of learner corpora (Gilquin 2015; Meunier 2021). Careful selection, documentation, explanation, and justification of all criteria also increase "the likelihood that the resulting corpus is methodologically-sound" (Bell and Payant 2021: 56).

2.1 Learner corpus typology

Defining the corpus typology is the first step in designing a corpus. The typology of a learner corpus depends on several aspects, i.e., medium, size, text type, time of collection, target language (L2), learners' mother tongue (L1), and scope of collection.

2.1.1 Medium

Learner corpora can consist of written texts or phonetic/prosodic transcriptions of spoken discourse. The number of existing written L2 Chinese learner corpora is significantly higher compared to the number of oral corpora (Iurato 2022a; Iurato 2022b; Xu 2019; Zhang and Tao 2018). New types of corpora are multi-modal corpora (see, for example, Gao and Wang 2017; Huang 2018; Kong 2013), which usually contain collections of photo-pictorial elements, video, and speech recordings accompanied by transcriptions and gesture annotations. Multimodal corpora allow us to study how two or more modalities interface with one another in human communication.

2.1.2 Size

A distinction can be drawn between global and local learner corpora. Global corpora are large-scale projects and collect a vast amount of data from students from multiple universities/research centres; local corpora are small-scale projects and collect a minor amount of data from small groups of learners, who are usually both contributors and users of the corpus (Gilquin 2015). A further type of corpora is 'in-house learner corpora', which lie somewhere between global and local corpora. In this case, the contributors do not correspond to the users, but they come from the same population of learners (generally the same university) (Gilquin 2015).

2.1.3 Text type

Theoretically, any text type, also referred to as 'genre', may be represented in a learner corpus (Bell and Payant 2021). However, in practice, the two most common text types are argumentative essays for writing and informal interviews for speaking (Callies and Götz 2015; Granger 2012). This selection reflects the need to sample the least constrained types of production data (Granger 2012). Nonetheless, diversification in terms of textual genres is desirable. Recently, SLA and LCR researchers have attempted to include a variety of genres (i.e., task types) to ensure a balanced representation of learner interlanguage (see, for example, Campillo Llanos 2014; Lozano 2021).

2.1.4 Time of collection

Learner corpora can be cross-sectional or longitudinal. The former collect samples of learner production from different categories of learners at a particular point in time; the latter include data from the same learners produced at different stages in their development (Granger 2012; Meunier 2021). Quasi-longitudinal learner corpora (sometimes referred to as 'pseudo-longitudinal learner corpora') are also quite common; they contain data collected from learners at different proficiency levels at a single point in time (Granger 2012). In LCR and SLA, cross-sectional and quasi-longitudinal corpora are the most common, as they allow researchers to gather more data in a short period of time (Gilquin 2015). In pseudo-longitudinal corpora, the developmental stages of learners are classified according to external criteria, such as proficiency test or grade level. This can be problematic, as proficiency level is often assessed according to different parameters in different school systems, and it does not always reflect the actual learners' proficiency[3], especially if it is calculated on external unreliable variables (Tono 2003).

2.1.5 Target language

Learner corpora can be classified on the basis of the target language they sample. English is still the predominant target language, as re-

[3] For an in-depth discussion about learners' language proficiency assessment in LCR, see Leclercq *et al.* (2014) and Callies and Götz (2015).

vealed by the *Learner Corpora Around the World* database[4]. However, over the last few years, other L2s have gradually "joined the learner corpus bandwagon" (Granger 2012: 12). Most learner corpora are monolingual, as they contain data from only one target language, such as the *Jinan Chinese Learner Corpus* (JCLC; Wang *et al.*, 2015). On the other hand, a small but increasing percentage of learner corpora is multilingual, like the *Multilingual Corpus of Second Language Speech* (MuSSeL; Rubio *et al.* 2021), which collects texts produced in four languages: Chinese, French, Portuguese, and Spanish.

2.1.6 Learners' mother tongue (L1)

Mono-L1 learner corpora include data from learners from one and the same L1 background, i.e., a single L1 population (Gilquin 2015). Differently, multi-L1 learner corpora include data from learners from different mother-tongue backgrounds (Granger 2012), like the *Guangwai-Lancaster Chinese Learner Corpus*[5]. It is a collection of written and spoken data produced by learners of L2 Chinese from 80 different countries studying at the Guangdong University of Foreign Studies in China. Multi-L1 learner corpora are very useful for investigating the effect of L1 crosslinguistic influence.

2.1.7 Scope of collection

Learner corpora can be distinguished according to the purpose for which they are compiled. Commercial learner corpora, such as the *Cambridge Learner Corpus*[6], are compiled by publishers with the aim of creating learning materials based on learner outputs (Granger *et al.* 2015). Academic learner corpora are generally compiled by researchers interested in exploring learners' language use and interlanguage. However, if existing available corpora do not suit one's research purpose, or if there is a shortage of corpora from which to extract the data needed for a particular research, there remains the option of compiling one's

[4] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (visited 2023/02/20).

[5] https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fguangwai (visited 2023/02/20).

[6] https://www.sketchengine.eu/cambridge-learner-corpus/ (visited 2023/02/20).

own local learner corpus (Giqluin 2015). The biggest advantage of such a bespoke corpus is that it is fully controllable (Millar and Lehtinen 2008).

2.2 Environment, learner and task variables

Three main variables play a role in the corpus compilation process: the environment in which the data are gathered, the learners whose performances are being collected, and the tasks that participants are asked to complete (Bell and Payant 2021; Gilquin 2015).

As for the environment, a learner corpus can be compiled in different linguistic contexts, and each linguistic context has different implications for the collection process and data analysis. For instance, in second language contexts, learners are exposed to the target language in daily activities, while in foreign language contexts opportunities for interaction in the target language are limited because the context of the common target language use is the classroom (Bell and Payant 2021). Furthermore, a distinction can be made between data collected in educational settings (at school/university) and in natural settings (mundane activities outside school/university). This distinction is particularly significant because second languages can be used in several varieties of contexts, but foreign languages can also be used outside educational settings (Gilquin 2015), for example to send e-mails to colleagues.

As for the learner variables, it is important to stress that the type and number of participants, the criteria, and the rationale for recruiting participants will affect the analyses of the data (Mackey and Gass 2021). Gathering and making available a rich set of metadata is therefore fundamental to increasing the rigor and transparency of learner corpora (Bell and Payant 2021; Tono 2003). This information can be obtained through the "learner profile questionnaire" (Gilquin 2015: 18), which collects: 1) personal information about the learner (e.g. age, gender, nationality, mother tongue, level of education, level of proficiency); 2) information about the learner's knowledge of other languages (e.g. additional language(s) studied and related level of proficiency, extensive experience of living abroad); 3) information about the learner's educational background (e.g. length of time studying the target language, universities and countries where the target language

was studied, where he/she went to school and university). The learner profile questionnaire is accompanied by the informed consent form that learners are required to complete if they allow their data to be used for research purposes (Bell and Payant 2021).

As for the tasks[7], they can involve different variables, like timing constraints (the learner may have a limited time to write the text; timing can be controlled while performing computer-based tasks), availability of reference tools (grammar books, dictionaries), intertextuality (allowing or not the consultation of secondary sources such as articles, sample texts), computerization (writing by hand or using a computer) (Gilquin 2015). Topic (complex or sensitive themes to be discussed, for example) can also affect learners' performance (Mackey and Gass 2021). In addition, the researcher should take into account that if the composition that the participants are asked to write is part of an examination, the pressure to perform may alter the final results. Finally, motivation can affect the quality of learner data; motivated learners are more likely to complete the texts carefully and not to leave the paper blank. Participants should therefore be volunteers, and recruitment should be through general online advertisements, rather than through individual solicitations (Mackey and Gass 2021). Selecting the most appropriate tasks for data collection is also a crucial issue. There are innumerable types of tasks that can be created to compile a written corpus[8]. The choice of one task over another is highly dependent on the research questions outlined and may also be related to the theoretical framework within which the research is being developed (Mackey and Gass 2021).

3. The *Bimodal Italian Learner Corpus of Chinese* (BILCC)

The written sub-corpus that I will describe as a case-study in Section 4 is a portion of a larger ongoing corpus project: the compilation

[7] Here, I will only focus on variables and the design of written tasks. For an overview of oral tasks, see Faitaki and Murphy (2020), Prior (2018), Rolland *et al.* (2020).

[8] For an overview of tasks for written data elicitation, see Mackey and Gass (2021), Gass (2018).

of the *Bimodal Italian Learner Corpus of Chinese* (BILCC). BILCC, which is methodologically grounded in the LCR framework, corresponds to the concept of learner corpus as a collection of contextualized data produced by L2 learners. 'Bimodal' describes the corpus mode of BILCC. It is defined as 'bimodal' because the medium of the data it collects, i.e., one of the aspects of the corpus typology (see Section 2.1), has a dual nature that allows us to explore L1 Italian learners' pragmalinguistic knowledge of L2 Chinese from two different perspectives (written and spoken production). In fact, similar to the *Arabic Learner Corpus*[9] and the *YKI National Certificate Corpus*[10] listed in the CLARIN digital infrastructure[11] (Hinrichs and Krauwer 2014; Jong *et al.* 2020), BILCC comprises written and spoken data from Italian learners of L2 Chinese. Specifically, the spoken data consist of recordings of speech and their transcriptions. The data collection was conducted from December 2020 to March 2021. The written data were gathered from 103 BA (N=56) and MA (N=47) beginner (N=19), intermediate (N=50), and advanced (N=34) L1 Italian learners of L2 Chinese with an average age of 23, studying at the Ca' Foscari University of Venice. The spoken data were collected from 58 BA (N=30) and MA (N=28) students, divided in beginner (N=16), intermediate (N=21), and advanced (N=21) levels, who had previously completed the written tasks for the compilation of the written corpus. Since in LCR external proficiency measures are considered the only reliable criteria[12], the learners were grouped into three different proficiency levels according to their HSK language proficiency test scores[13]. The written sub-corpus of BILCC includes

[9] https://www.arabiclearnercorpus.com/about-the-corpus-en (visited 2023/02/20).

[10] https://metashare.csc.fi/repository/browse/the-national-certificates-corpus/944099dafccc11e18b49005056be118efc2ef6e1f96241b681c1d9bec0e9033a/ (visited 2023/02/20).

[11] https://www.clarin.eu/ (visited 2023/02/20).

[12] In LCR, proficiency based on external factors (e.g., institutional level, age) and self-assessment practices are considered unreliable and problematic (see Callies and Götz 2015; Leclerq *et al.*; Tono 2003).

[13] Although the HSK language proficiency test has been criticised (Fu *et al.* 2013; Peng *et al.* 2021), it has been adopted to assess learners' Chinese language proficiency for practical reasons and because more reliable criteria have not been found.

53,248 Chinese characters, 38,793 word tokens, and 693 word types. The spoken sub-corpus consists of 25-hour recordings, while the corresponding transcriptions, which were manually performed, consist of 14,321 Chinese characters, 10,414 word tokens, and 285 word types. The bimodal corpus mode is one of BILCC's strengths, since most of the existing (L2 Chinese) corpora are mainly written, and spoken corpora are rare (see Iurato 2022a; Iurato 2022b; Zhang and Tao 2018). Another strength is that it includes multiple sources of data from the same group of learners, a design feature generally absent and thus highly encouraged in LCR (Tracy-Ventura *et al.* 2021). The corpus is accompanied by a control corpus of 30 L1 Chinese speakers for comparative purposes. The written data consist of 19,073 Chinese characters, 10,414 word tokens, and 285 word types. The spoken data consist of 7-hour recordings and the related transcription include 11,872 Chinese characters, 9,048 word tokens, and 113 word types. All participants voluntarily completed the tasks.

Another distinctive feature of BILCC is that it is a specific-purpose learner corpus compiled to explore the (explicit/implicit) pragmalinguistic knowledge of a particular syntactic structure: the Chinese "*shì* ... *de* proper focus cleft" (Paul and Whitman 2008: 424) with [V 的*de* O] order (henceforth the terms '*shì*...*de* cleft construction' and 'proper cleft' will be used interchangeably). It is used to highlight a specific information (agent, time, place, manner, instrument, cause, etc.) of a concluded event that is generally given as presupposition in the discourse (Cui and Sung 2021; Li 2008; Li and Thompson 1981; Lü 1982; Paris 1979; Paul and Whitman 2008), as illustrated in (1):

(1) 我们是昨天去的图书馆。

wǒmen	*shì*	*zuótiān*	*qù-de*	*túshūguǎn*
IPL	SHI	yesterday	go-DE	library

'It was yesterday that we went to the library.'

Descriptively, the *shì*...*de* proper cleft is signalled by two morphemes: *shì* in pre-verbal position marking the clefted element; *de* in post-verbal position between the verb and the object. Unfortunately, in the

literature there is no consensus on the syntactic roles of *shì*[14] and *de*[15]. In the analysis of BILCC, following Cheng (2008), Hole (2011), Paul and Whitman (2008), Xu (2014), *shì* is considered a copula serving as focus marker. On the other hand, following Lü (1982), Shi (1994), Paul and Whitman (2008), *de* is identified as an aspect marker that leads to the mandatory past-tense interpretation of the sentence (Paul and Whitman 2008; Cui and Sung 2021). In fact, one of the most striking language-specific features of Chinese clefts is that, unlike English *it*-clefts, they have a default past tense reading. As a matter of fact, material contradicting the past-tense interpretation (e.g., future-oriented temporal adverbials) of the sentence cannot occur in Chinese proper clefts (Cui and Sung 2021; Hole 2011; Li and Thompson 1981; Paul and Whitman 2008).

From a discourse-pragmatic point of view, the *shì…de* proper cleft is generally considered a focalizing device that shows a bipartioning between the focus (i.e., the clefted constituent) and the presupposed content (Jing-Schmidt 2017). Similarly to English *it*-clefts, the Chinese cleft includes a narrow focus signaled by the copula. The syntactic constituents that can be clefted, and thus occupy the post-copular position, are subjects and adjuncts (Paul and Whitman 2008), whereas, due to Chinese word order constraints, post-verbal elements such as objects and verbal complements cannot[16] (see Luo 2009).

Following Li (2008), Xu (2014), and Cui and Sung (2021), we argue

[14] Variously analyzed as a copula by Paris (1979), Ross (1983), Li and Thompson (1981), Cheng (2008), Paul and Whitman (2008), as a copula marking the focus by Li (2008), Xu (2014), as an intensifier adverb functioning as an emphasis marker by Shi (1994).

[15] Variously categorized as a nominalizer of a headless relative clause by Paris (1979), Li and Thompson (1981), Cheng (2008), Li (2008) and Xu (2014), as an aspect marker by Zhao (1979), Lü (1982), Shi (1994) and Cui and Sung (2021), as a head of an aspectual phrase (AspP) projection by Paul and Whitman (2008), as an enclitic past tense marker by Simpson and Wu (2002).

[16] It must be pointed out that post-verbal constituents can receive focus by means of phonological prominence (Lü 1982; Cheng 2008; Cui and Sung 2021), as in the case of "object focus clefts" (Paul and Whitman 2008: 424; Hole 2011: 1712), where the object is not the "cleft focus", but the prosodically "marked focus" (Hole 2011: 1712).

that the *shì...de* proper cleft has both contrastive and non-contrastive discourse-pragmatic functions. We can distinguish two types of *shì... de* proper clefts: contrastive clefts and non-contrastive clefts. In contrastive *shì...de* proper clefts, the focal element conveys contrast because the information it expresses is opposed to other relevant information in the discourse. Such sentences are used to correct, expand, or clarify the listener's assumptions (see Berretta 1994). Therefore, sentences such as (2) contain a corrective contrastive focus (Jing-Schmidt 2017; Cui and Sung 2021).

(2) 他不是开车来的，是坐火车来的。 (Zhao 1979: 62)

tā	*bú*	*shì*	*kāi-chē*	*lái-de*
3SG	NEG	SHI	drive-car	come-DE
shì		*zuò-huǒchē*		*lái-de*
SHI		by-train		come-DE

'It was not by car that he came, but by train.'

Conversely, in non-contrastive clefts, although the clefted constituent is syntactically focalized, it does not convey contrast because the information it expresses is not opposed to other information already given in the discourse (Berretta 1994; Xu 2014; Garassino 2014). For example, *wh*-interrogative cleft sentences such as (3A) are mainly used non-contrastively (Li 2008; Cui and Sung 2021). Here, the non-contrastive focus has the full "original" focus marking function (Korzen 2014: 232), as it is used to direct the listener's attention to a specific piece of information of a concluded event, without any intention to create contrast (Berretta 1994; Cui and Sung 2021). This also applies to corresponding responses such as (3B).

(3) A: 你是怎么去的中国？ (Zhao 1979: 61)

nǐ	*shì*	*zěnme*	*qù-de*	*Zhōngguó*
2SG	SHI	how	go-DE	China

'How did you go to China?'

B: 我是坐飞机去的。

wǒ	*shì*	*zuò-fēijī*	*qù-de*
1SG	SHI	by-plane	go-DE

'I went by plane.'

Furthermore, non-contrastive clefts have a textual function, serving anaphoric recovery: the clefted constituent recalls or summarizes what has been said before (Berretta 1994). In other words, the cleft sentence brings an element from the "background" to the "foreground" of the text (Prince 1978: 891). For example, in sentence (4), the *shì…de* cleft served to draw the listener's attention to a detail that had remained in the background, i.e., *lái táonàn* 来逃难 ('come to be a refugee'), which had already appeared as a complete predicate. The example in (4) and the corresponding translation are taken from Xu (2014: 174).

(4) 我来逃难我都什么也不管。头发也不管，[...] 衣服没有买过一件[...]，先天上就觉得说我是来逃难的，我就应该很吃苦耐劳。

wǒ	*lái*	*táonàn*		*wǒ*	*dōu*	*shénme*
1SG	come	take.refugee		1SG	all	whatever
yě	*bù*	*guǎn*	*tóufǎ*	*yě*	*bù*	
also	NEG	regard	hair	also	NEG	
guǎn		*yīfú*	*méiyǒu*	*mǎi-guo*		*yī*
regard		clothes	NEG	buy-EXP		one
jiàn	*xiāntiān-shàng*			*jiù*	*juédé*	
CLF	in.nature-upon			just	feel	
shuō	*wǒ*	*shì*	*lái*	*táonàn-de*		
say	1SG	SHI	come	take.refugee-DE		
wǒ	*jiù*	*yīnggāi*	*hěn*	*chī-kǔ-nàiláo*		
1SG	just	should	very	be.able.to.bear.hardships		

'I came to be a refugee, so I have regards for nothing. I don't care about my hair, [...] and I have not bought one piece of clothing [...]. Internally I feel that I am here to be a refugee, and I should be able to bear all hardships.'

Based on the assumption that the *shì…de* proper cleft has both corrective contrastive and non-contrastive functions, BILCC, inspired by the working models adopted by Callies (2009), identifies two different pragmatic functions for annotating the corpus data at the pragmatic level: 'intensification' and 'corrective contrast'. Intensification refers to *shì…de* cleft sentences that do not convey a contrastive focus, such as the non-contrastive clefts in (3)-(4), where the focus has the original

function of highlighting a piece of information. Corrective contrast refers to sentences in which the *shì...de* pattern signals a corrective contrastive focus, such as the contrastive clefts in (2).
To summarize, the data collected for the compilation of BILCC have been annotated at the grammatical and pragmatic levels to explore L1 Italian learners' knowledge of the syntactic and pragmatic properties of the *shì...de* cleft construction[17]. The corpus annotation process started in 2021 and is still ongoing. BILCC will be made freely available once the compilation and annotation are completed.

4. A case study: The written sub-corpus of BILCC

In this section, I will address the methodological issues relating to the design and collection of the written sub-corpus of BILCC. It was assembled according to strict specific design criteria and is the result of "theoretically motivated" (Tracy-Ventura and Paquot 2021: 4) open ended tasks. In what follows, I will describe: a) general corpus design principles and SLA-motivated features; b) the corpus typology; c) environment, learner, and task variables; d) the data collection procedure.

4.1 Corpus design features

Based on the design criteria recommended by Tracy-Ventura *et al.* (2021) to fill current gaps in corpus compilation in LCR, the written sub-corpus of BILCC shows the following features:

1. *It focuses on L2s other than English*. Since most learner corpora collect data on L2 English, and in LCR there is a general shortage of corpora collecting data from other L2s (Gráf 2017; Tracy-Ventura *et al.* 2021), this corpus starts to fill the gap in LCR by collecting data on L2 Chinese, an underexplored language variety in LCR (Iurato 2022a; Iurato 2022b).

[17] The analysis of the *shì...de* cleft construction on syntactic, semantic and discourse-pragmatic levels goes beyond the scope of this paper. For a literature overview of the topic, see Cheng (2008), Hole (2011), Iurato (in preparation), Jing-Schmidt (2017), Li (2008), Lü (1982), Paul and Whitman (2008) and Xu (2014).

2. *It includes data from learners at all proficiency levels*. Unlike most corpora that collect data from intermediate and advanced learners (Tracy-Ventura *et al.* 2021), this corpus includes data from beginner, intermediate, and advanced learners. The inclusion of beginner learner data is important, as the purpose of SLA research is to explain acquisitional development from beginning to end (Tracy-Ventura *et al.* 2021).
3. *It includes a control corpus of L1 speakers for comparative purposes*. Based on the Integrated Contrastive Model (Granger 1996) for the study of cross-linguistic influence, the learner corpus is accompanied by a control corpus of 30 Chinese native speakers as a benchmark of the (variety of) language learners are exposed to (Lozano 2021)[18]. Moreover, following one of the most important corpus design criteria outlined by Sinclair (2005), the two corpora are comparable because the tasks that were administered to the learners and the control group were identical. In other words, the same design across the two corpora ensures comparability, a key issue particularly emphasized by Lozano (2021).
4. *It contains a rich set of metadata on learner variables*, as it is important to document learner variables accurately both to support data interpretation (Bell and Payant 2021) and to increase reliable comparability across studies (Tracy-Ventura *et al.* 2021). Adhering to another corpus design principle defined by Sinclair (2005) on the documentation of variables, the present corpus contains systematically collected metadata of learners' and Chinese native speakers' sociolinguistic variables.
5. *It includes a pilot study of the data collection*. Since piloting is not a common practice in learner corpus design and should rather become part of it (Tracy-Ventura *et al.* 2021; Bell and Payant 2021), this corpus went through piloting to a) check the effectiveness of the tasks; b) ensure that the instructions and tasks were understandable for the participants; c) measure the time that the participants needed to complete the tasks; d) check whether the expected

[18] Control corpora are justified in LCR; see the 'comparative fallacy' vs. 'comparative hypocrisy' debate (Meunier 2021; Tracy-Ventura *et al.* 2021).

findings could emerge from the collected data; e) correct/eliminate mistakes left in the instruments (Dörney and Csizér 2012).

6. *It is freely available to the research community*. Making corpora freely available is a highly encouraged practice in LCR. This is especially the case for (error)-annotated corpora, which are unfortunately rarely shared, but definitely necessary for the development of NLP tools (Tracy-Ventura *et al.* 2021). To fill this gap in the LCR literature, the present corpus will be made freely available to the research community once completed. Furthermore, it will be accompanied by the documentation on metadata and effects on piloting, as this practice will help researchers understand the necessary decisions to be made when compiling a corpus (Bell and Payant 2021).

4.2 Corpus typology

Following the learner corpus typology dimensions outlined by Bell and Payant (2021), Gilquin (2015), Granger (2012), and Meunier (2021), the written sub-corpus of BILCC is:

1. *In-house*. The contributors and the users are not the same students, but they belong to the same population of learners, i.e., L1 Italian learners of L2 Chinese studying in Italy.
2. *Pseudo-longitudinal*. It collects data at a specific point in time (December 2020-March 2021) from different learners at different stages in their development (beginner, intermediate, and advanced learners).
3. *Mono-L1*. It contains data produced by a single L1 population.
4. *Academic*. It is compiled for research purposes.
5. *Specific purpose designed*. Due to the lack of data from Italian learners in existing L2 Chinese learner corpora, the corpus was specifically designed to analyze the use of the *shì…de* cleft construction in L1 Italian learners' production.
6. *Representative*. The language samples in the corpus are representative of learners' contextualized language use at three different proficiency levels, as the data are produced through open-ended tasks that allow learners to choose their own wording (Callies and Götz 2015). The feature of representativeness distinguishes the corpus from common data collections (Meunier 2021).

7. *Error and pragmatically tagged*. A target-oriented error taxonomy and an error tagset with 20 labels for the annotation at the grammatical level were designed to spot learners' errors in the use of the *shì...de* cleft construction. A pragmatic annotation was also added to detect the misuse of the *shì…de* proper cleft construction at the discourse-pragmatic level. Following Díez-Bedmar (2015), the identification of errors was carried out simultaneously by a bilingual team consisting of two expert Chinese native speakers and the researcher, whose L1 is the same as that of the learners.

4.3 Environment, learner and task variables

The written section of BILCC was compiled in a foreign language context in an "educational setting" (Gilquin 2015: 16), i.e., Chinese language courses at the Ca' Foscari University of Venice.

The metadata of the written corpus collected all information about learner variables. In order to increase the rigor and transparency of the corpus (Bell and Payant 2021), the metadata will be published alongside the learner corpus. Based on the learner metadata scheme proposed by Wang *et al.* (2015), a detailed metadata set that gathers information on the learner profile[19] was collected. Information on learners' educational background[20], knowledge of further foreign language(s) other than Chinese (plus related proficiency level), was also included, as this is crucial for interpreting the role of L1 in learners' interlanguage (Tracy-Ventura *et al.* 2021). These variables were collected through a learner profile questionnaire and a language background questionnaire. Participants completed these two questionnaires before completing the tasks. Students were also asked to complete an informed consent form[21].

[19] Age, gender, nationality, L1(s), parents' L1(s), partner's L1(s), current program of study.

[20] Highest level of education, languages officially used at primary, high school and university, countries where the learner attended school and university, periods of Chinese language study in China, purpose of stay in China, experience of living in Chinese communities.

[21] The template of the informed consent form was provided by the Ca' Foscari University of Venice. The specific research objective, i.e., the knowledge of the *shì...de* cleft construction, was not made explicit in order not to influence the participants' output.

As for the design of open-ended tasks, it is important to bear in mind that different types of tasks affect the learner output, as learners' language use varies across tasks. The variety of lexical-grammatical aspect combinations used by learners may also be influenced by other variables such as task topic, prompts, and type of narrative. Moreover, different tasks tap different knowledge (explicit or implicit) (Tracy-Ventura and Myles 2015). Following Callies (2009) and Tracy-Ventura and Myles (2015), one way to get around the above-mentioned issues in creating the written tasks for the compilation of BILCC was to carefully design tasks that naturally create contexts for the features under investigation. As the corpus collects data to explore whether learners are aware of the two pragmatic functions of the *shì…de* cleft construction, specific open-ended tasks were created to elicit the data in a definite discourse context in which the use of the perfective verbal aspect could also emerge. Thus, the data production was contextualized in a scenario in which it was necessary to refer to concluded events. First, the tasks provided a background that allowed students to highlight particular details of a concluded action (approach used to explore learners' knowledge of the pragmatic meaning of intensification), and then a background that allowed them to clarify/correct incorrect information/assumptions related to a concluded event (approach used to explore learners' knowledge of the pragmatic function of the corrective contrastive focus). Four purpose-designed, theoretically motivated, open-ended written tasks were designed: two discourse completion tasks (DCTs) and two picture-based narratives. The tasks provided contexts for the time reference and the aspect as authentically as possible, by implementing the principles for ensuring the task effectiveness defined by Tracy-Ventura and Myles (2015). The tasks were also rich in background and foreground.

Based on Callies (2009), the DCTs consisted of two sections introduced by different situational descriptions designed to create communicative contexts in which specific information needed to be highlighted for reasons of intensification (situation 1) and corrective contrast (situation 2). Each item was contextualized by a short passage extracted from the narrative text that the participants had read earlier. This text passage was followed by a semi-structured dialogue sequence that participants were asked to complete.

In section 1, participants were asked to provide utterances focusing on a detail of information highlighted in the preceding text passage, as exemplified in (5):

(5) 中午，我们在餐厅吃了午饭。
zhōngwǔ wǒmen zài cāntīng chī le wǔfàn
(At noon, we had lunch at the restaurant)
A (your friend's question): ...
B (your response): ...

In section 2, the items involved obvious cases of misunderstanding between two interlocutors: the informant (interlocutor A) and a fictitious fellow student (interlocutor B). Participants were asked to create a dialogue in which interlocutor A corrects or clarifies incorrect assumptions made by interlocutor B concerning a situation that occurred in the context described in the previous text, as illustrated in (6):

(6) 午饭过后，我们在电影院旁边的超市买了一些饮料。
wǔfàn guòhòu wǒmen zài diànyǐngyuàn pángbiān de chāoshì mǎi le yì xiē yǐnliào.
(After lunch, we bought some drinks from the supermarket next to the cinema)
A (your friend's wrong assumption): ...
B (your corrective response/clarification): ...

The DCTs contained ten items, five of which appeared in an intensifying, and five in a contrastive context. Participants were asked to highlight (situation 1) or correct/clarify (situation 2) information about specific details (time, place, manner, agent, etc.) of a concluded event described in the narrative text. No distractors were included in the DCTs. The instructions were highly detailed, and the communicative goal was explicitly stated, thus there was no need to conceal this by including distractors.
The first picture-based narrative also consisted of two sections. In section 1, participants were given a picture preceded by a context with instructions. Participants were asked to answer the question provided in order to highlight a detail of the event described in the picture, see Figure 1.

Tom has just bought a new book. He shows it to his friends Hannah and John. Hannah and John are curious to know where and when he bought that book. How could they ask Luca for such information? And how could Tom reply to them to provide this specific information?

Figure 1. Example of picture-based narrative, section 1.

In section 2, informants were asked to create a dialogue between the portrayed interlocutors, in which interlocutor A was required to correct/clarify incorrect assumptions made by interlocutor B about specific details of a concluded event, see Figure 2.

Layla received a bouquet of flowers. Her friend Nicole thinks it was Oliver who gave her those flowers, but Layla corrects this wrong assumption, clarifying that it was Ismael who actually gave them. Write the dialogue between Layla and Nicole.

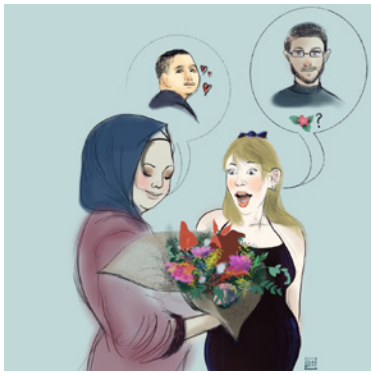

Figure 2. Example of picture-based narrative, section 2.

The second picture-based narrative, 'My trip to China', consisted of an open role-play (Mackey and Gass 2021). Students were given a picture (see Figure 3) representing two interlocutors accompanied by an attack line (*wǒ qù le Zhōngguó* 我去了中国, 'I went to China') from which to create two dialogues in two different contexts. In the first dialogue, interlocutor B was required to obtain specific information from interlocutor A about the concluded trip. In the second dialogue, interlocutor A was required to clarify/correct incorrect information provided by interlocutor B about the concluded event.

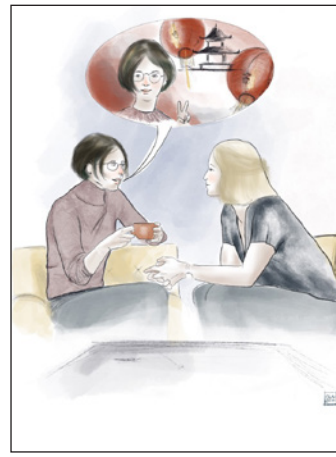

Figure 3. Illustration used to complete the second picture-based narrative, open role-play.

An artist[22] was commissioned to draw the pictures included in the tasks. Instructions were given in Italian for the learners and in Chinese for the control group. The use of teaching materials and dictionaries was not allowed. Since the research did not focus on vocabulary knowledge, an entry level vocabulary based on the official HSK2 vocabulary list was included in the tasks; those words that the students might not know were provided. Time on tasks was not controlled, as completion of the tasks was the priority. Before administration, all

[22] I would like to thank Francesca Biundo for drawing the illustrations used for data collection, some of which appear in this paper.

Chinese sentences in the tasks were checked by three native speakers who teach Chinese language at the Ca' Foscari University of Venice. Moreover, the tasks were piloted to ensure that they were manageable for all participants, that the vocabulary was appropriate, and that the instructions were clear (Bell and Payant 2021). The tasks were administered from the least to the most structured (1. picture-based narratives; 2. DCTs) to avoid the completion of the most guided tasks affecting the production (Mackey and Gass 2021).

4.4 Data collection and corpus description

The learner questionnaires, the consent form, and the tasks were administered via Google Form[23]. Three out of the 103 learners who had completed the written tasks were ruled out because their L1 was Chinese. As for the control group, 30 Chinese native speakers (Chinese language teachers: N=24; students: N=6) who live in Italy, Germany, USA, and China completed the tasks. Their average age was 36. Data and metadata from learners and native speakers were downloaded and stored in electronic format on Excel files, so that they could be retrieved and used with different software. The data of each learner and each native speaker were labelled with a univocal code (e.g., L1, NS1, etc.), since personal information revealing names and surnames of students was eliminated to preserve their privacy (Castillo Rodríguez *et al.* 2020). Data cleaning and character counting were carried out using Regex in Nisus Writer Pro[24]. SegmentAnt (Anthony 2017) was used for basic word segmentation (Chinese Jieba Engine); AntConc (Anthony 2019) for word tokens and word types counting. Basic information on the learner corpus and the control corpus size are illustrated in Table 1.

	Learner corpus (100 Ls)	**Control corpus (30 NSs)**
Sentences	4,985	1,504
Chinese characters	53,248	14,321
Word tokens	38,793	10,414
Word types	693	285

Table 1. Size of the written sub-corpora of BILCC.

[23] The selection of data collection tools was dictated by the restrictions caused by the Covid-19 emergency during the lockdown period in Italy.

[24] https://www.nisus.com/pro/ (visited 2023/02/20).

The tasks proved to be effective as the expected results were achieved. Inferential analyses of the learner and the native speaker corpora reveal and confirm the initial hypothesis that the *shì…de* cleft construction is used much more frequently by native speakers than by learners (χ^2=143.307, df=1, *p*=.00001). Moreover, significant differences in the proportion of the *shì…de* cleft construction are observed between advanced and beginner (χ^2=104.637, df=1, *p*=.00001), between advanced and intermediate (χ^2=16.52625, df=1, *p*=.000047), and between elementary and intermediate learners (χ^2=62.8262, df=1, *p*=.00001). Therefore, statistics show that the *shì...de* proper cleft construction is used significantly more frequently by intermediate and advanced than by beginner learners (see Table 2).

	Beginner (16)	**Intermediate (50)**	**Advanced (34)**	**Total Ls**	**NSs (30)**
Frequency of correct *shì...de*	120	824	729	1,673	920
Total sentences in the corpus	815	2,462	1,708	4,985	1,504
Proportion of *shì...de* (%)	14.7	33.46	42.68	33.56	61.17

Table 2. Frequency rate of the *shì…de* cleft construction in learner and native speaker data.

Statistics also reveal that the *shì…de* cleft construction conveying the pragmatic meaning of intensification is more frequently used by native speakers than by learners (χ^2=213.1123, df =1, *p*=.00001), and that it is significantly more frequently used by advanced than by intermediate and beginner learners. Similarly, the *shì…de* proper cleft construction conveying the pragmatic meaning of corrective contrast is more frequently used by native speakers than by learners (χ^2=25.1626, df=1 *p*=.00001), and it is used significantly more frequently by advanced than by intermediate and beginner learners. However, it is interesting to note that the frequency rate of contrast in learners' written production is closer to that in native speakers' production, compared to the frequency rate of intensification in learner data which drastically deviates from the frequency rate of intensi-

fication in native speaker data (see Table 3). The analysis therefore reveals that the learners use the *shì...de* cleft construction more to correct/clarify wrong assumptions than to highlight the speaker's evaluation or explanation regarding a concluded event.

	Intensification		**Contrast**	
	Native Speakers (30)	**Learners (100)**	**Native Speakers (30)**	**Learners (100)**
Frequency of correct *shì...de*	645	913	335	776
Total sentences in the corpus	1,504	4,985	1,504	4,985
Proportion of *shì...de* (%)	42.88	18.31	22.27	15.56
Significant difference between NSs and Ls: χ^2=213.1123, df =1, p<.05			Significant difference between NSs and LSs: χ^2=25.1626, df =1 p<.05	

Table 3. Frequency rate of the *shì...de* cleft construction conveying intensification and contrast in learner and native speakers data.

Since the focus of this article is outlining the methodological steps in corpus design and data collection, no further information on the accuracy rate and error rate of the *shì…de* cleft construction in learner and native speaker data will be provided. This information, as well as details on corpus annotation, corpus exploitation, data analysis, and data interpretation will be presented in future research.

5. Conclusions

This paper defined the methodological steps to be followed when compiling a written learner corpus specifically designed to analyse the morpho-syntactic and pragmatic knowledge of Chinese syntactic features in L1 Italian learners' output. The compilation of the written sub-corpus of BILCC was presented as a case study. The paper demonstrated the importance of methodological decisions in defining the corpus typology, learner and environment variables, and task types, as

these aspects impact on the learner output, the validity, and the generalizability of findings emanating from the corpus (Bell *et al*. 2021). The rationale of the corpus compilation and the importance of methodological transparency in documenting all steps in the preparation of the learner corpus were also discussed. The purpose of this paper was presenting a protocolized methodology for the compilation of a learner corpus, which may also be suitable for future corpus projects, given the continuous expansion of L2 Chinese studies in Italy and the current lack of Italian corpora of L2 Chinese. If the methodology is to be replicated for future studies, the methodological steps must be well-structured and as flexible as possible. Only in this way can they be used for different research purposes (Bell *et al*. 2021). The compilation process presented here can be repurposed for future research on L2 Chinese acquisition, since it is grounded on the main methodological stages of LCR (Granger 2012), and it is based on specific design principles (Tracy-Ventura *et al*. 2021). In addition, the corpus size and statistical results demonstrated the effectiveness of the compilation process of BILCC. The standardization of the methodology for compiling a corpus, especially in an expanding field in which few studies have been carried out (as in the case of L2 Chinese corpora collecting data from Italian learners) (Iurato 2022a; Iurato 2022b), is essential to avoid different studies on the same topic generating different, sometimes even contradictory, results. Thus, since so far in Chinese LCR "there is no synthesis of research findings, making it difficult to outline a full picture of [L2 Chinese] learners' development" (Zhang and Tao: 58), the aim of this contribution is encouraging the replicability of the methodology presented here to support future studies on the acquisition of L2 Chinese by L1 Italian learners.

References

Anthony, L. (2017) SegmentAnt (Version 1.1.3) [Computer Software]. Tokyo: Waseda University. https://www.laurenceanthony.net/software (visited 2022/04/09).

Anthony, L. (2019) AntConc (Version 3.5.8) [Computer Software]. Tokyo: Waseda University. https://www.laurenceanthony.net/software (visited 2022/04/09).

Bell, P., L. Collins and E. Marsden (2021) "Building an Oral and Written Learner Corpus of a School Programme: Methodological Issues", in B. Le Bruyn and M. Paquot (Eds.) *Learner Corpus Research Meets Second Language Acquisition*. Cambridge, Cambridge University Press: 214-242.

Bell, P. and C. Payant (2021) "Designing Learner Corpora. Collection, Transcription, and Annotation", in N. Tracy-Ventura and M. Paquot (Eds.) *The Routledge Handbook of Second Language Acquisition and Corpora*. London/New York, Routledge: 53-67.

Berretta, M. (1994) "Ordini marcati dei costituenti di frase in italiano. La frase scissa". *Vox Romanica*, 53: 79-105.

Callies, M. (2009) *Information Highlighting in Advanced Learner English. The syntax–pragmatics interface in second language acquisition*. Amsterdam, John Benjamins.

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

Campillo Llanos, L. (2014) "A Spanish Learner Oral Corpus for Computer-Aided Error Analysis". *Corpora*, 9(2): 207-238.

Castillo Rodríguez, C., J.M. Díaz Lage and B. Rubio Martínez (2020) "Compiling and Analyzing a Tagged Learner Corpus: A Corpus-Based Study of Adjective Uses". *Círculo de Lingüística Aplicada a la Comunicación*, 81: 115-136. http://dx.doi.org/10.5209/CLAC.67932.

Cheng, L. (2008) "Deconstructing the *shi…de* Construction". *The Linguistic Review*, 25: 235-266.

Cui, S. and K. Sung (2021) *A Reference Grammar for Teaching Chinese. Syntax and Discourse*. Singapore, Springer.

Díez-Bedmar, M.B. (2015) "Dealing with Errors in Learner Corpora to Describe, Teach and Assess EFL Writing: Focus on Article Use", in E. Castello, K. Ackerley and F. Coccetta (Eds.) *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*. Bern, Peter Lang: 37-69.

Dörney, Z. and K. Csizér (2012) "How to Design and Analyze Surveys in Second Language Acquisition Research", in A. Mackey and S.M. Gass (Eds.) *Research Methods in Second Language Acquisition. A Practical Guide*. Oxford, Wiley-Blackwell: 74-94.

Dutra, D.P. and A.R. Gomide (2015) "Compilation of a University Learner Corpus". *BELT*, *Brazilian English Language Teaching Journal*, 6: 21-33. http://dx.doi.org/10.15448/2178-3640.2015.s.21311.

Faitaki, F. and V.A. Murphy (2020) "Oral Language Elicitation Tasks in Applied Linguistics Research", in J. McKinley and H. Rose (Eds.) *The Routledge Handbook of Research Methods in Applied Linguistics*. London/New York, Routledge: 360-369.

Fu, J. 符华均, P. Zhang 张晋军, Y. Li 李亚男, P. Li 李佩泽 and T. Zhang 张铁英 (2013) "新汉语水平考试HSK（五级）效度研究". *Kaoshi Yanjiu*, 3: 65-69.

Gao, F. and B. Wang (2017) "A Multimodal Corpus Approach to Dialogue Interpreting Studies in the Chinese Context: Towards a Multi-Layer Analytic Framework". *The Interpreters' Newsletter*, 22: 17-38.

Garassino, D. (2014) "Cleft Sentences. Italian-English in Contrast", in A.M. De Cesare (Ed.) *Frequency, Forms and Functions of Cleft Constructions in Romance and Germanic: Contrastive, Corpus-Based Studies*. Berlin, De Gruyter: 101-138.

Gass, S.M. (2018) "SLA Elicitation Tasks", in K. Phakiti, P. De Costa, L. Plonky and S. Starfield (Eds.) *The Palgrave Handbook of Applied Linguistics Research Methodology.* London, Palgrave Macmillan: 313-338.

Gilquin, G. (2015) "From Design to Collection of Learner Corpora", in S. Granger, G. Gilquin and F. Meunier (Eds.) *The Cambridge Handbook of Learner Corpus Research.* Cambridge, Cambridge University Press: 9-34.

--- (2021) "Combining Learner Corpora and Experimental Methods", in N. Tracy-Ventura and M. Paquot (Eds.) *The Routledge Handbook of Second Language Acquisition and Corpora*. London, Routledge: 133-144.

Gilquin, G. and S.T. 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.

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

Granger, S. (1996) "From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora", in K.

Ajimer, B. Altenberg and M. Johansson (Eds.) *Languages in Contrast. Text-Based Cross-Linguistic Studies*. Lund, Lund University Press: 37-51.

--- (2012) "How to Use Foreign and Second Language Learner Corpora", in A. Mackey and S. M. Gass (Eds.) *Research Methods in Second Language Acquisition. A Practical Guide.* Oxford, Wiley-Blackwell: 7-29.

Granger, S., G. Gilquin and F. Meunier (Eds.) (2015) *The Cambridge Handbook of Learner Corpus Research*. Cambridge, Cambridge University Press.

Hinrichs, E. and S. Krauwer (2014) "The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, May 2014: 1525-1531.

Hole, D. (2011) "The deconstruction of Chinese shi...de clefts revisited". *Lingua*, 121: 1707-1733.

Huang, L. (2018) "Issues on Multimodal Corpus of Chinese Speech Acts: A case in Multimodal Pragmatics". *Digital Scholarship in the Humanities*, 33(2): 316-326.

Iurato, A. (2022a) "Learner Corpus Research Meets Chinese as a Second Language Acquisition: Achievements and Challenges". *Annali di Ca' Foscari. Serie Orientale*, 58(1): 709-742.

--- (2022b) "Analyzing Chinese Learner Corpus Research and Chinese learner corpora: Main Features, Critical Issues and Future Pathways". *Kervan. International Journal of African and Asian Studies*, 26(2): 533-564.

--- (in preparation) *The Acquisition of the Chinese* 是*shì*...的*de Cleft Construction by L1 Italian Learners*: *Triangulating Learner Corpus and Experimental Data*. PhD Dissertation. Venice, Ca' Foscari University of Venice; Bremen, University of Bremen.

Jing-Schmidt, Z. (2017) "Grammatical Construction and Chinese Discourse", in C. Shei (Ed.) *Routledge Handbook of Chinese Discourse Analysis*. London, Routledge: 209-230.

Jong, F. de, B. Maegaard, D. Fišer, U. Dieter Van and A. Witt (2020) "Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN", in *Proceedings LREC 2020, 12th International Conference on Language Resources and Evaluation*. ELRA. https://www.aclweb.org/anthology/2020.lrec-1.417/ (visited 2023/02/20)

Kong, K.C. (2013) "A Corpus-Based Study in Comparing the Multimodality of Chinese- and English- Language Newspapers". *Visual Communication*, 12(2): 173-196. https://doi.org/10.1177/1470357212471594.

Korzen, I. (2014) "Cleft sentences. Italian-Danish in contrast", in A.M. De Cesare (Ed.) *Frequency, Forms and Functions of Cleft Constructions in Romance and Germanic: Contrastive, Corpus-Based Studies*. Berlin, De Gruyter: 217-275.

Leclercq, P., A. Edmonds and H. Hilton (2014) (Eds.) *Measuring L2 Proficiency: Perspectives from SLA*. Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781783092291.

Li, C.N. and S.A. Thompson (1981) *Mandarin Chinese: a functional reference grammar*. Berkeley, University of California Press.

Li, K. (2008) "Contrastive Focus Construction in Mandarin Chinese", in M.K.M. Chan and H. Kang (Eds.) *Proceedings of the 20th North American Conference on Chinese Linguistics* (NACCL-20), vol. 2. Columbus (Ohio), The Ohio State University: 759-774.

Lozano, C. (2021) "CEDEL 2: Design, Compilation, and Web Interface of an Online Corpus of L2 Spanish Acquisition Research". *Second Language Research*, 1-19. https://doi.org/10.1177/02676583211050522.

Lozano, C. and A. Mendikoetxea (2013) "Corpus and Experimental Data: Subjects in Second Language Research", in S. Granger, G. Gilquin and F. Meunier (Eds.) *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead*. Louvain-la-Neuve, Presses Universitaires de Louvain: 313-323.

Lü, B. 吕必松 (1982) "关于 "是……的"结构的介个问题". *Yuyan Jiaoxue yu Yanjiu*, 4: 21-37.

Mackey, S. and S.M. Gass (2021) *Second Language Research. Methodology and Design*. New York/London: Routledge.

Mai, Z. and B. Yuan (2016) "Uneven Reassembly of Tense, Telicity and Discourse Features in L2 Acquisition of the Chinese shì…de Cleft Construction by Adult English Speakers". *Second Language Research*, 32 (2): 247-276.

Meunier, F. (2021) "Introduction to Learner Corpus Research", in N. Tracy-Ventura and M. Paquot (Eds.) *The Routledge Handbook of Second Language Acquisition and Corpora*. London, Routledge: 23-36.

Millar, N. and B. Lehtinen (2008) "DIY Local Learner Corpora: Bridging Gaps between Theory and Practice". *JALT CALL Journal*, 4(2): 61-72.

Norris, J. and L. Ortega (2003) "Defining and measuring SLA", in C.J. Doughty and M.H. Long (Eds.) *The Handbook of Second Language Acquisition*. MA-US, Blackwell Publishing Ltd: 717-761.

Paris, M.-C. (1979) *Nominalization in Mandarin Chinese. The Morpheme 'de' and the 'shi…de' Constructions*. Paris, Université de Paris VII.

Paul, W. and J. Whitman (2008) "Shi...de Focus Clefts in Mandarin Chinese". *The Linguistic Review*, 25(3/4): 413-451.

Peng, Y., W. Yan and L. Cheng (2021) "Hanyu Shuiping Kaoshi (HSK): A Multi-Level, Multi-Purpose Proficiency Test". *Language Testing*, 38(2): 326-337. https://doi.org/10.1177/0265532220957298.

Prince, E.F. (1978) "A comparison of wh-clefts and it-clefts in discourse". *Language*, 54: 883-906.

Prior, M.T. (2018) "Interviews and Focus Froups", in K. Phakiti, P. De Costa, L. Plonky and S. Starfield (Eds.) *The Palgrave Handbook of Applied Linguistics Research Methodology*. London, Palgrave Macmillan: 225-248.

Rolland, L., J. Dewaele and B. Costa (2020) "Planning and Conducting Ethical Interviews: Power, Language and Emotions", in J. McKinley and H. Rose (Eds.) *The Routledge Handbook of Research Methods in Applied Linguistics*. London/New York, Routledge: 279-289.

Romagnoli, C. and S. Conti (Eds.) (2021) *La lingua cinese in Italia. Studi su didattica e acquisizione*. Roma, Roma TrE Press.

Ross, C. (1983) "On the Functions of Mandarin 'de'". *Journal of Chinese Linguistics*, 11(2): 214-246.

Rubio, F., E. Kia, E. Schnur and J. Hacking (2021) *Multilingual Corpus of Second Language Speech* (MuSSeL). https://l2trec.utah.edu/learner-corpora/mussel/ (visited 2023/02/20).

Shan, C. 杉村博文 (1999) ""的"字结构与分类", in L. Jiang 江蓝生 and J. Hou 侯精一 (Eds.) 汉语现状与历史的研究. Beijing, China Social Sciences Press.

Shi, D. (1994) "The Nature of Chinese Emphatic Sentences". *Journal of East Asian Linguistics*, 3: 81-100.

Simpson, A. and Z. Wu (2002) "From D to T- Determiner Incorporation and the Creation of Tense". *Journal of East Asian Linguistics*, 11: 169-209.

Sinclair, J. (2005) "Corpus and Text – Basic Principles", in M. Wynne (Ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford, Oxbow Book Company: 1-16.

Tono, Y. (2003) "Learner Corpora: Design, Development and Applications", in D. Archer, P. Rayson, A. Wilson and T. McEnery (Eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL 16, Lancaster University: 800-809.

--- (2016) "What is Missing in Learner Corpus Design?", in M. Alonso-Ramos (Ed.) *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam, John Benjamins: 33-52.

Tracy-Ventura, N. and F. 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.

Tracy-Ventura, N. and M. Paquot (Eds.) (2021a) *The Routledge Handbook of Second Language Acquisition and Corpora*. London/New York, Routledge.

--- (2021b) "Second Language Acquisition and Corpora. An Overview", in N. Tracy-Ventura and M. Paquot (Eds.) *The Routledge Handbook of Second Language Acquisition and Corpora*. London/New York, Routledge: 1-8.

Tracy-Ventura, N., M. Paquot and F. Myles (2021) "The Future of Corpora in SLA", in N. Tracy-Ventura and M. Paquot (Eds.) *The Routledge Handbook of Second Language Acquisition and Corpora*. London/New York, Routledge: 409-424

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

Xu, Y. (2014) "A Corpus-Based Functional Study of *shi…de* Constructions". *Chinese Language and Discourse*, 5(2): 146-184.

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

Zhang, J. and H. Tao (2018) "Corpus-Based Research in Chinese as a Second Language", in C. Ke (Ed.) *The Routledge Handbook of Chinese Second Language Acquisition*. New York, Routledge: 48-62.
Zhao, S. 赵淑华 (1979) "关于"是……的"句". *Yuyan Jiaoxue yu Yanjiu*, 1: 57-66.