# A spoken Chinese corpus: Development, description, and application in L2 studies

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Applied Linguistics

at Massey University, Manawatū
New Zealand

Lin Li

2021

# Abstract

This thesis introduces a corpus of present-day spoken Chinese, which contains over 440,000 words of orthographically transcribed interactions. The corpus is made up of an L1 corpus and an L2 corpus. It includes data gathered in informal contexts in 2018, and is, to date, the first Chinese corpus resource of its kind investigating non-test/task-oriented dialogical interaction of L2 Chinese. The main part of the thesis is devoted to a detailed account of the compilation of the spoken Chinese corpus, including its design, the data collection, and transcription. In doing this, this study attempts to answer the question: what are the key considerations in building a spoken Chinese corpus of informal interaction, especially in building a spoken L2 corpus of L1–L2 interaction? Then, this thesis compares the L1 corpus and the L2 corpus before using them to carry out corpus studies. Differences between and within the two subcorpora are discussed in some detail. This corpus comparison is essential to any L1–L2 comparative studies conducted on the basis of the spoken Chinese corpus, and it addresses the question: to what extent is the L1 corpus comparable to the L2 corpus? Finally, this thesis demonstrates the research potential of the spoken Chinese corpus, by presenting an analysis of the L2 use of the discourse marker 就是 *jiushi* in comparison with the L1 use. Analysis considers mainly the contribution 就是 *jiushi* makes as a reformulation marker to utterance interpretation within the relevance theoretic framework. To do this, it seeks to answer the question: what are the features that characterise the L2 use of the marker 就是 *jiushi* in informal speech?

The results of this study make several useful contributions to the academic community. First of all, the spoken Chinese corpus is available to the academic community through the website, so it is expected the corpus itself will be of use to researchers, Chinese teachers, and students who are interested in spoken Chinese. In addition to the obtainable data, this thesis presents transparent accounts of each step of the compilation of both the L1 and L2 corpora. As a result, decisions and strategies taken with regard to the procedures of spoken corpus design and construction can provide some valuable suggestions to researchers who want to build their own spoken Chinese corpora. Finally, the findings of the comparative analysis of the L2 use of the marker 就是 *jiushi* will contribute to research on the teaching and learning of interactive spoken Chinese.

# Acknowledgements

This work would not have been possible without the support and help of many people. I would like to begin by thanking my supervisors Gillian Skyrme, Cynthia White, Tony Fisher, and Michael Li at Massey University. Without their unswerving encouragement, expert guidance, and dedicated involvement in every step throughout the process, this study would have never been accomplished. Thanks are due also to the University Scholarships Committee at Massey University for funding and supporting my studies in New Zealand. I also extend my gratitude to the staff at the School of Humanities, Media and Creative Communication, the librarians, and the Graduate Research School at Massey University for their help and support. I must also thank all the participants in this study, for their time, kindness, and willingness to talk with me so that I could gather valuable data. I wish to express my gratitude to Bowen for his assistance with the release of my corpus.

Thanks are due also to my colleagues: Siti, Hong, Anne, Panithi, Kwan, Noor, Barira, and Somayyeh for the shared lunch, laughter, complaints, discussions, and the overall warmth we shared in New Zealand. I would also like to thank Huan and Chujie at Massey University for sharing research ideas, and always being there to listen to and support me.

Finally, I must thank my parents for providing food, love, and support, thereby allowing me to adjust to working from home based in China due to the outbreak of the pandemic, my elder sister and my cute niece for reminding me to have fun when I have been stressed out. Thank you all! 后会有期, New Zealand!

# Contents

# List of Tables

# List of Figures

# Chapter 1 Introduction

## 1.1 A New Corpus Resource: The Spoken Chinese Corpus

This study introduces a spoken Chinese corpus of conversational interaction which is made up of two parts: an L1 corpus which includes L1–L1 interaction, and an L2 corpus which contains L1–L2 interaction. My interest in conversational interaction arose out of my personal experience of participating in English interactions in daily life. As a learner and user of English, who has gained English language skills primarily from classroom English teaching exercises and textbooks, when I started my doctoral life in New Zealand I was clearly aware that the English expressions I had learnt from textbooks differed from the norms adopted by native English speakers in everyday conversation. The lack of authentic conversational input in English hampered me from conversing sufficiently in daily life. In my case, there was no doubt that there existed a communicative dimension to English that I had little experience of; as a result, there was a need for me to develop knowledge about real conversational English to improve my communicative skills. L2 learners of Chinese who learn Chinese in a country where Chinese is not spoken are likely to encounter the same problems as I did when they talk with L1 speakers of Chinese in everyday conversation. Given this, the value of informal conversational interaction became so evident that I decided to gather interaction between L1 and L2 speakers of Chinese and to compile the data into a spoken L2 corpus. The creation of the spoken L1 corpus was motivated by the consideration of employing the L1–L2 comparative approach to study L2 production. In what follows, I will give an account of the background information in some detail.

### 1.1.1 Chinese Conversational Interaction

In this thesis, the conversational interaction in the spoken L2 corpus comprises informal L1–L2 interaction between L2 participants and the researcher (a native Chinese speaker) conducted in non-academic settings by adopting an unstructured interviewing method. It is noticeable that L2 speech in formal or academic settings, such as oral tests and classroom discussions, remains the most common sources of data for spoken L2 corpus creation and L2 studies (for examples of spoken L2 Chinese corpora, see Appendix A). A notable example in the English academic

community recently is the Trinity Lancaster Corpus (TLC), which contains spoken interaction between exam candidates (L2 speakers of English) and examiners (L1 speakers of English) (Gablasova et al., 2019b). The TLC represents interaction in an institutional setting in which the language "is semi-formal in nature and is close to academic interaction" (Gablasova et al., 2019b, p. 142). Unlike in some other corpus studies on L2 production, classroom talk between teachers and students, oral tests between examiners and examinees, and any oral pedagogical tasks (e.g., picture descriptions) are outside the scope of conversational interaction as outlined in this study. There are many good reasons for the popularity of using test/task-oriented data: they are much easier to collect in comparison to informal interactions; the complexity of L2 language use is reduced; variables affecting L2 language use can be controlled in some way, and so forth (Ädel, 2015). Yet, as Biber and Conrad (2019) note, daily conversation, for most people, "is the most common type of spoken language that they produce" (p. 1), and the popularity of the well-known spoken L1 corpora, such as the London Lund Corpus of Spoken English (LLC, Svartvik, 1990) and the Spoken British National Corpus 2014 (the Spoken BNC2014, Love et al., 2017), has proved the considerable value of spontaneous conversations in studies on spoken L1. Unfortunately, spontaneous interactions between L2 and L1 speakers have not so far received much attention in the development of spoken L2 corpora. Although the data gathered for the spoken L2 corpus in this thesis differ from spontaneous conversation collected for these existing spoken L1 corpora (see 3.2.4 in Chapter 3), here the first step forward has been taken to broaden the traditional databases for L2 studies. It is expected that this thesis, as a preliminary study, can reflect on the L2 use of spoken Chinese in a previously under-researched setting, and that this kind of L2 data can complement evidence from existing corpus resources by providing an opportunity to reveal aspects of L2 language.

There are many ways to gather conversational interaction. For example, the compilers of the Spoken BNC2014 encourage participants to use their smart phones to record themselves chatting with whoever they choose (Love et al., 2017). This method without the involvement of corpus compilers to a large extent can ensure the data are as natural as possible. In this study, due to some practical considerations, such as costs and time, as we will see in Chapter 3, I adopted an unstructured interviewing method which involved some degree of structuring to gather informal conversations between L2 speakers of Chinese and me. Further discussions in terms of data collection will be given in Chapter 3.

### 1.1.2 The Spoken L2 Corpus of L1–L2 Interaction

In the two decades since the compilation of the first L2 Chinese corpus–the Chinese Interlanguage Corpus (汉语中介语语料库系统)[1] (Chen, 1996; Chu & Chen, 1993), there has been a tremendous increase in interest and activity in the area of L2 Chinese corpus research[2]. Table 1 shows some spoken L2 Chinese corpora[3] created and mentioned in previous studies at the time of writing (for detailed introductions to these corpora, see Appendix A).

**Table 1**

*Some Existing Spoken L2 Chinese Corpora*

| Spoken L2 Chinese corpus | Availability |
|---|---|
| the Learner Corpus of Spoken Chinese (汉语学习者口语语料库) | Not available |
| the HSK Dynamic Spoken Corpus (HSK 动态口语语料库) | Not available |
| the Jinan Learner Corpus of Spoken Chinese (暨南大学华文学院口语语料库) | Available |
| the Soochow L2 Corpus | Not Available |
| the Longitudinal Chinese Interlanguage Corpus (外国留学生汉语口语纵向语料库) | Not Available |
| the spoken component of the Guangwai–Lancaster Chinese Learner Corpus (GLCLC) | Available |
| the spoken component of the International Corpus of Learner Chinese (全球汉语学习者语料库) | Available |
| the Country–Specific Corpus of Spoken Chinese Interlanguage (Korean Learners) (韩国学习者汉语中介语口语语料库) | Not available |
| the Multimodal Corpus of L2 Chinese (外国学生多模态口语语料库) | Not available |

One of the strengths of corpora is the availability of data, that is, corpora can be shared by the research community in order for the results to be replicable (Brezina, 2018). However, due to commercial and other considerations, only three spoken L2 Chinese corpora are

[1] Most of the L2 Chinese corpora presented in this thesis neither have official English names (i.e., English names of these corpora are not widely accepted by the Chinese academic community, different authors tend to use different English names refer to the same corpus) nor have been introduced to the English research community. For those which have not been introduced in the English research literature, I used my own English translations to represent them. For the sake of clarity, both Chinese names and English translations are provided throughout this thesis.

[2] Corpora compiled by Chinese researchers, but which target other languages (e.g., English spoken by Chinese learners) were outside my account of L2 Chinese corpora in this thesis.

[3] I specifically exclude from this thesis all consideration of speech corpora which are mainly concerned with the sound of speech and typically collected for the purposes of improving technology. This however does not imply that research into these corpora is any less important; simply that these types of corpora are distinguished from spoken corpora discussed in this study, making it important to consider them separately.

available to the public at the time of writing: the Jinan Learner Corpus of Spoken Chinese (暨南大学华文学院口语语料库), the Guangwai–Lancaster Chinese Learner Corpus (GLCLC), and the International Corpus of Learner Chinese (全球汉语学习者语料库).

Users can access the websites of these three corpora, which allow searches, concordances and so on to be carried out. However, the Jinan Learner Corpus of Spoken Chinese (暨南大学华文学院口语语料库) has limited contributions to the academic community due to a number of constraints. First of all, it is unclear when this corpus was built, and according to which design criteria (e.g., the data collection method). This corpus also fails to provide such basic information as sample size, speaker metadata, and text metadata. In any corpus study, researchers normally take empirical examples from a corpus and analyse the data quantitatively and/or qualitatively, then relate the results back to the corpus which represents a target population in order to make reliable interpretations. Without the above fundamental but crucial information in terms of speakers and texts contained in the corpus, researchers can only show the result that a specific pattern is found in that L2 corpus, for example, but are not able to give any meaningful explanations of the result, such as who (e.g., female or male learners, Japanese or Korean learners of Chinese) used this pattern in which settings. Therefore, it is clear that the extent of the usefulness of this corpus to the research community is questionable.

The Guangwai–Lancaster Chinese Learner Corpus (GLCLC) is a rather new L2 Chinese corpus which consists of both a written part and a spoken part[4]. The GLCLC can be freely accessed via Sketch Engine with the downloading limitation of a maximum of 10,000 rows of data[5]. The disadvantage of accessing this corpus through Sketch Engine is that corpus users are limited to concordances. A concordance is "a collection of the occurrences of a word-form, each in its own textual environment" (Sinclair, 1991, p. 32). Figure 1 is an example of a concordance. The bold word forms under investigation—key word in context (KWIC) — appear in the centre of each line, and the length of the context is specified for various purposes (e.g., four or eight words on either side of the keywords). Even though KWIC saves researchers looking up each occurrence in a corpus (Sinclair, 1991) and allows researchers to observe the behaviour of a particular word-form in detail (McEnery et al., 2006), when carrying out a

[4] For a brief introduction to this project, see http://cass.lancs.ac.uk/tag/guangwai-lancaster-chinese-learner-corpus/.
[5] The GLCLC: https://www.sketchengine.eu/guangwai-lancaster-chinese-learner-corpus/.

corpus study (e.g., to analyse the function of discourse markers), it is always valuable to have access to the whole text rather than concordances only.

**Figure 1**

*An Extract of a Concordance from the GLCLC Via Sketch Engine*

cql <u country=="新西兰">[] **119** (71.5 per million)

| | Details | Left context | KWIC | Right context |
|---|---|---|---|---|
| 1 | S_B_M_SDW_NZ > 来，好，做一下自我介绍。</s><s> | | **呃** | ，我的全名是。</s><s> 但，呃，呃， |
| 2 | S_B_M_SDW_NZ 认，呃，但，但认……嗯。</s><s> | | **认识** | 我，我的人都叫我。</s><s> 嗯。</s |
| 3 | S_B_M_SDW_NZ 我的人都叫我。</s><s> 嗯。</s><s> | | **这** | ，这也是我的笔名。</s><s> 嗯哼。< |
| 4 | S_B_M_SDW_NZ ¦是我的笔名。</s><s> 嗯哼。</s><s> | | **呃** | ，我的专业……嗯。</s><s> 我的专业 |
| 5 | S_B_M_SDW_NZ s><s> 呃，我的专业……嗯。</s><s> | | **我的** | 专业是，呃，电脑工程和汉语。</s>< |
| 6 | S_B_M_SDW_NZ 3脑工程和汉语。</s><s> 嗯。</s><s> | | **虽然** | 我的汉语水平不如电脑工程。</s><s> |
| 7 | S_B_M_SDW_NZ 下如电脑工程。</s><s> 嗯哼。</s><s> | | **我** | ，嗯，可是，可是我，我更喜欢汉语 |
| 8 | S_B_M_SDW_NZ 戈更喜欢汉语。</s><s> 嗯哼。</s><s> | | **呃** | ，呃，希望，希望我将，将来……嗯 |
| 9 | S_B_M_SDW_NZ 望，希望我将，将来……嗯。</s><s> | | **能** | ，能够会看懂中，中文小说。</s><s> |
| 10 | S_B_M_SDW_NZ ɔ，中文小说。</s><s> 嗯哼。</s><s> | | **在** | ，在学习电脑工程的时候……嗯哼。 |

Similarly, the International Corpus of Learner Chinese (全球汉语学习者语料库) is also likely to be of limited usefulness with respect to my research focus and to those who want to do the kind of research that I am interested in. It is a new L2 Chinese corpus which consists of both a written part and a spoken part. Unlike the GLCLC which is not very well known among Chinese researchers, the International Corpus of Learner Chinese (全球汉语学习者语料库) has received much attention in the Chinese academic community. However, to my best knowledge, there is no article explicitly introducing the detailed procedures of the design and compilation of this L2 corpus. Although some corpus building principles of this corpus have been addressed by Cui and Zhang (2011b) and Zhang and Cui (2013a, 2013b, 2015), it remains unclear how those theoretical guidelines are being employed in practice. Without available descriptions of each step of the construction of the International Corpus of Learner Chinese (全球汉语学习者语料库), it is difficult to evaluate whether the findings of corpus studies on L2 Chinese reflect the real features of L2 production or are caused by the decisions made in the procedures of corpus construction.

In addition to the above concerns in the existing spoken L2 Chinese corpora, another issue which limits their usefulness is that compilers of these corpora have not produced detailed records of how transcription was undertaken. While spoken interaction may exhibit orderliness at deeper levels of organisation (see for example, Sacks et al., 1974), at a more superficial level it is often a rather messy affair, characterised by non-verbal vocalisations, pauses, coordinating conjunctions, repetitions, false starts, and so on (Kirk & Andersen, 2016). It thus is natural that transcribers as listeners have various strategies to understand and deal with these features. Therefore, transcription can be viewed as an act of interpretation which is influenced by transcribers' biases. Thus, without detailed accounts of the strategies employed to represent the features inherent in spoken Chinese, the credibility of the transcribed data, and the reliability of studies on the basis of the existing spoken L2 Chinese corpora, may be subjected to criticism. Detailed discussions on transcription for spoken L2 Chinese corpora will be provided in Chapter 2.

Overall, these three spoken L2 Chinese corpora have their virtues but at the same time due to the concerns identified above (appropriateness, accessibility, and availability of the spoken data), they are of limited utility to investigate language use in informal conversational interaction. Given that these corpus data were not well matched to my research interests, I decided to build my own corpus from scratch. It is important to emphasise that a transparent corpus design and creation procedure is called for and maintained throughout the whole thesis, due to the imperative role data plays in any corpus studies. To enable researchers to make appropriate use of the spoken L2 corpus built in this study, this thesis provides transparent and documented accounts of each step of the spoken L2 corpus compilation (for further discussion on this see Chapters 3 and 4).

### 1.1.3 The Spoken L1 Corpus of L1–L1 Interaction

One common way of using corpora to study L2 production is to compare L2 language with L1 language, the latter of which is seen as the ultimate attainment when learning a foreign/second language (Gablasova, Brezina, & McEnery, 2017; Granger, 1996, 2009, 2015). Corpus studies using this comparative approach are able to uncover distinctive features of and provide more valuable insights into L2 language use than those using L2 data alone (Gablasova, Brezina, & McEnery, 2017; Granger, 2009, 2015). This being the case, to study L2 production by employing the L1–L2 contrastive approach, it was essential to obtain a spoken L1 Chinese

corpus that would be comparable to the spoken L2 corpus of informal interaction. For the sake of comparability, it was decided to build an L1 corpus designed with the similar criteria to the L2 corpus (further discussions with regard to corpus comparability can be found in Chapter 2). Therefore, the spoken Chinese corpus introduced in this thesis was designed to include an L1 subcorpus and an L2 subcorpus[6]. Corpus design and data collection will be discussed in Chapter 3.

Based on the spoken Chinese corpus, I have also included in this thesis an example of how the spoken Chinese corpus can be employed to study L2 production by using the L1–L2 comparative approach, including an investigation into the use of a Chinese discourse marker in informal interaction (see Chapters 5 and 6). In the following section, I mainly provide some preliminary background to its relevance to Chinese discourse marker analysis.

## 1.2 Using the Spoken Chinese Corpus to Analyse L2 Production

Spoken corpora can be used to conduct a wide range of studies. In my original research plan, I expected in the first place to investigate the use of Chinese collocations (i.e., words that occur in combinations, such as 红茶 [red tea] *black tea*) by L2 speakers of Chinese to address the research potential of the spoken Chinese corpus. When carrying out the transcription, however, I encountered some unforeseen difficulties in certain contexts, such as what spoken features needed to be transcribed, and how to appropriately record features which could not be represented in standard orthograph (e.g., backchannels). These difficulties not only challenged and ultimately deepened my understanding of Chinese interaction, but also prompted me to shift my attention from lexical collocations to Chinese discourse markers[7] (e.g. 就是 *jiushi* 'be exactly', further explanations will be given in Chapter 6). When lexical items, such as 就是 *jiushi* 'be exactly', 然后 *ranhou* 'then', 这个 *zheige*/*zhege* 'this', and 那个 *neige*/*nage* 'that' in the examples given below, function as discourse markers in contexts, they do not have propositional content on some occasions and are used to signal the pragmatic relation of an utterance to the immediate context (e.g., Fang, 2000; Li, 2016; Shi & Hu, 2013; Wang, 2018; Zhang & Gao, 2012; Zhu, 2017), which have the similar pragmatic functions as *well*, *you know*,

[6] In the following chapters, the two parts of the spoken Chinese corpus will be presented as the L1 corpus and the L2 corpus rather than the L1 subcorpus and the L2 subcorpus.
[7] Although, earlier documentation reflects the original intentions, the changes have not had an impact on the ethical considerations of this study.

*like*, and *just* in spoken English. Here, I give two examples which were taken from the spoken Chinese corpus built in this study:

(1) An extract of the speech of a male L2 speaker of Chinese

<L15> 我**这个**来自**那个**在新西兰的北岛的西部 er **然后**我在那边儿长大**然后**长大以后就去**那个** er <university>在<city>的那个大学

I ***well (zheige[8])*** come from ***like (neige)*** the western side of New Zealand's North Island er ***then (ranhou)*** I grew up there ***then (ranhou)*** when I grew up went to ***like (neige)*** er <university> in <city> that university

<S00> eng eng[9]

eng eng

<L15> **然后**在那边儿学读过法律 er **然后这个**读了五年就毕业了**然后**毕业了之后我就觉得那时候我认识到中文和中国的**这个这个这个**市场的重要性

(I) ***then (ranhou)*** studied law at that university er ***then (ranhou) like (zheige)*** got my degree five years later ***then (ranhou)*** after I graduated I thought at that moment I realised the importance of Chinese and China ***well I mean you know (zheige zheige zheige)*** market

(2) An extract of the speech of a male L1 speaker of Chinese

<N04> **然后然后**去了之后**就是**那天的话风还挺大的就在湖边**那个**那个度假中心**然后**风还挺大的**然后就就是**大家吃完饭之后**然后**就开始**就是**分组**就是**大家选**那个**对什么比较感兴趣就选什么**然后**大概的话**就是** er 是有帆板有帆船还有**那个**单人的皮划艇还有双人的那种划船的

***then then (ranhou ranhou)*** (I) got there ***you know (jiushi)*** it was a windy day ***like (neige)*** that vacation centre was on the lake ***then (ranhou)*** it was quite windy ***then you know (ranhou jiu jiushi)*** after dinner ***then (ranhou)*** we started to ***you know (jiushi)*** choose groups ***you know (jiushi)*** we chose ***like (neige)*** activities which we were interested to attend ***then (ranhou)*** generally ***you know (jiushi)*** er they had windsurfing sailing and ***like (neige)*** single kayaks and double kayaks

[8] 这个 *zheige*/*zhege* 'this' and 那个 *neige*/*nage* 'that' in the examples can be interpreted as hesitational devices, which enable the speakers to buy production time and signal utterance continuation.

[9] In this thesis, *eng* can be referred to as a backchannel, indicates that the hearer is following the speaker or agrees with the speaker.

During the transcription process, my cursory observations showed that both L1 and L2 speakers of Chinese used these expressions in their conversations, and that L1 speakers used them more frequently than L2 speakers (see Chapter 6), which suggested that these expressions not only characterise Chinese in informal conversations but also serve some interactional functions in communication. With regard to the roles discourse markers play in communication, some four decades ago Svartvik (1980) notes that:

> [I]f a foreign learner says *five sheeps* or *he goed*, he can be corrected by practically every native speaker. If, on the other hand, he omits a *well*, the likely reaction will be that he is dogmatic, impolite, boring, awkward to talk to etc., but a native speaker cannot pinpoint an 'error'. (p. 171)

Likewise, Crystal (1988) states that markers such as *you know* can be seen "as the oil which helps us perform the complex task of spontaneous speech production and interaction smoothly and efficiently" (p. 48).

In the literature on English, numerous studies have attempted to specify the meaning and function of discourse markers with a wide range of frameworks reflecting divergent research interests and approaches (e.g., Blakemore, 2002; Fischer, 2006a; Fraser, 1990, 1999; Jucker & Ziv, 1998; Lenk, 1998; Redeker, 1991; Schiffrin, 1987; Schourup, 1999). Although there is little consensus about the characteristics of discourse markers, much of the research generally agrees that discourse markers contribute to the pragmatic meaning of utterances and fulfil integral functions in communication. Recently, with the development of spoken corpora, there is an increasing interest in characterising English discourse markers by making use of large-scale empirical corpus data (e.g., Aijmer, 2002, 2013; Beeching, 2016; Rühlemann, 2019). As a result, corpus studies on English discourse markers in L1 interaction have arguably led to a better understanding of the roles discourse markers have in speech. Meanwhile, the achievements of L1 studies on discourse markers have inspired L2 researchers to consider the characteristics of discourse markers in the L2 production of a number of languages. Compared with the myriad of studies on discourse markers in L1 production and the "jungle of publications" it is "almost impossible to find one's way through" (Fischer, 2006b, p. 1), relatively limited research has been undertaken on the range and variety of discourse markers used by L2 speakers in communication (Bardovi-Harlig, 2013; Callies, 2013; Fung & Carter, 2007; McEnery et al., 2019; Müller, 2005). The situation is particularly evident with reference to L2 research into Chinese discourse markers. It is noted that the characteristics of Chinese

discourse markers have received some coverage in the Chinese research literature, particular attention has been primarily paid to the origins (e.g., grammaticalization, or lexicalization) and functions of certain individual discourse markers in L1 production (e.g., Biq, 1990, 2001; Fang, 2000; Gao & Tao, 2021; B. Liu, 2009; Yue, 2020; Zhang & Gao, 2012; Zheng, 2020), and only very few analyses have been attempted for L2 production to date (e.g., Li, 2009; Shi, 2020). With little knowledge of the functions of Chinese discourse markers in L2 speech, it is not surprising to see that expressions such as 这个 *zheige*/*zhege* 'this' and 那个 *neige*/*nage* 'that' used by L2 speakers of Chinese tend to be thought of redundant or as evidence of a lack of speaking proficiency in some studies (e.g., Hu & Wang, 2011; Wang & Yang, 2011). In what follows, I shall give a brief account of the importance of studying Chinese discourse markers in L2 production.

It is widely accepted that learning a language involves more than learning grammar and vocabulary, pragmatic competence is also an essential component of being a successful L2 user (Bardovi-Harlig, 2013; Kasper & Rose, 1999; Taguchi & Roever, 2017). However, compared with other areas of L2 studies (such as grammar and lexis), the pragmatics of spoken communication is still an under-researched area in SLA research (Bardovi-Harlig, 2013; Callies, 2013; Gablasova, Brezina, McEnery, et al., 2017). In their discussion referring to East Asian pragmatics, Wang and Halenko (2019) point out that "Chinese is the second most studied East Asian language in L2 pragmatics next to Japanese" (p. 4), although only 14 data-based studies of Chinese learners' pragmatic competence and development had been published up to 2015 (Taguchi, 2015). Taguchi (2015) claims that the first and only book that devotes its entire attention to pragmatics of L1 and L2 speakers of Chinese is *Pragmatics of Chinese as native and target language* which is edited by Kasper (1995). There is thus an evident need for more research on L2 Chinese pragmatics if we are to know how L2 speakers develop their abilities to communicate effectively and appropriately in social settings. Given the dearth of studies on the L2 use of Chinese discourse markers, and the fact that Chinese discourse markers in L2 production are in need of description and interpretation, in developing the focus of this research it became evident that it would be worthwhile to investigate the roles expressions such as 就是 *jiushi* 'be exactly' have in interactions. Such an investigation could make a contribution to our knowledge of the function of discourse markers in L2 informal conversation. In order to get a fresh perspective on the language use of L2 speakers of Chinese in spoken contexts, this study is dedicated to investigating the use of the most frequently used discourse marker

observed in the spoken Chinese corpus: 就是 *jiushi* 'be exactly' (see Chapter 6 for detailed discussions). In this it contributes to a wider purpose of identifying how corpus linguistics can deepen the definition and description of Chinese discourse markers and contribute to our understanding of what they are doing in informal conversation. Although discourse markers reflect a very narrow view of pragmatics (McEnery et al., 2019), it is expected to provide new insights into L2 Chinese pragmatics, highlighting that this area that has not benefitted from the use of relatively large-scale quantitative analyses. Potential implications of the corpus analysis of the marker 就是 *jiushi* in the field of L2 pragmatics will be discussed in Section 7.6 in Chapter 7.

## 1.3 Research Questions

Given the above discussions, the aim of this thesis is twofold. It first attempts to compile a relatively well-designed spoken Chinese corpus consisting of an L2 corpus and a comparable L1 corpus. In the main part of this thesis, I provide a transparent and documented account of the practical decisions made in terms of the design and compilation of both the L1 and L2 corpora. The second part is an attempt to demonstrate how the spoken Chinese corpus may be used for linguistic research. To begin with, the second part first compares the L1 corpus and the L2 corpus with respect to corpus design and data collection, seeking to (i) evaluate to what extent the two corpora match each other and (ii) investigate whether certain decisions and compromises made during the corpus building procedure led to the observed differences between L1 and L2 use. Following this, this thesis then investigates the use of the discourse marker 就是 *jiushi* by L2 speakers of Chinese. This application as an example attempts to show how the spoken Chinese corpus can be used to investigate L2 production by employing the L1–L2 comparative approach.

With these concerns in mind, the present study asks:

RQ1: What are the main considerations in building a spoken Chinese corpus of informal interaction?

(i) What are the design features in creating a spoken L2 Chinese corpus of L1–L2 conversational interaction?

(ii) Which strategies should be employed to compile a spoken L1 corpus to ensure that it will be comparable to the spoken L2 corpus as far as possible?

RQ2: To what extent is the spoken L1 corpus comparable to the spoken L2 corpus?

RQ3: What are the features which characterise the L2 use of the Chinese discourse marker 就是 *jiushi* in the L1–L2 interactions?

(i) What is the role of 就是 *jiushi* when it is used as a reformulation marker in conversational interaction?

(ii) How frequent is the marker 就是 *jiushi* among L1 and L2 production?

(iii) How does the distribution of the marker 就是 *jiushi* differ between the two subcorpora?

The research questions above underpin an enquiry whereby the spoken Chinese corpus built in the present study can be a valuable resource for research into Chinese as well as a reliable basis for the contrastive study of L1 and L2 language in informal interaction. Moreover, the findings of the corpus analysis of the marker 就是 *jiushi* are expected to contribute to our knowledge of the nature of Chinese discourse markers.

## 1.4 Outline of the Thesis

Chapter 2 considers spoken L2 corpus design and compilation, and carefully discusses the methodological considerations in terms of conducting L1–L2 comparisons. The purpose is not to go into the very basics of corpus design, data and metadata collection, transcription or any other feature with respect to corpus construction, as relevant literature will be introduced and discussed in the following chapters where appropriate. Nor in Chapter 2 will I review the existing spoken L2 Chinese corpora to clarify the justification of the creation of the spoken Chinese corpus. The limits of the existing spoken L2 Chinese corpora have been clearly discussed above, meaning there is no need to repeat the contextualisation of the situation which has arisen whereby the creation of the spoken Chinese corpus is necessary. Furthermore, due to the scarcity of available information on these Chinese L2 corpora, it is doubtful whether a critical review can be successfully conducted. What Chapter 2 does do is to critically consider

both corpus design and compilation, and approaches to L1–L2 comparisons based on corpus data.

Chapters 3 and 4, as the foundation part of this thesis, provide a transparent and documented account of the practical decisions made in terms of the design and compilation of the spoken Chinese corpus developed in this study. Chapter 3 outlines key decisions and compromises made in terms of the spoken Chinese corpus design associated with data and metadata collection. Chapter 4 reflects on the practice of transcribing the data collected for the spoken Chinese corpus, describing the development of a bespoke transcription scheme for the corpus.

Chapters 5 and 6 demonstrate how the spoken Chinese corpus may be used for linguistic research. Chapter 5 draws extensively on the comparability of the L1 and L2 corpus, and covers the groundwork for the subsequent analysis. The aim of Chapter 6 is to investigate the use of the marker 就是 *jiushi* in L2 production in comparison to L1 use. In Chapter 6 I provide both qualitative and quantitative analyses of the marker 就是 *jiushi* within the relevance theoretic framework.

Chapter 7 presents a discussion on the design and management issues and decisions taken during the practice of the spoken Chinese corpus construction. It then considers the validity and reliability of the results of the L2 use of the marker 就是 *jiushi* as part of a wider interpretation of the significance of the findings of this study. At the end of this chapter, it makes some tentative suggestions for ways in which the findings of this study might contribute to research on spoken Chinese and contrastive interlanguage analysis. Finally, the concluding Chapter 8 summarises this thesis and discusses the major contributions and limitations of the work of this study. To conclude it suggests future research directions and the next steps that can be taken in the ongoing development of the corpus.

# Chapter 2 Literature Review

## 2.1 Introduction

This chapter reviews a range of theoretical considerations and methodological instructions that have been presented and discussed in the development and applications of spoken L2 corpora. The chapter is presented in three main parts. In the first part, the focus is on some crucial features of spoken L2 corpus design and compilation. By reviewing four existing spoken L2 corpora, this part first provides a critical discussion of certain selection criteria in terms of target L2 speakers in corpus design. Then, it turns to consider issues regarding spoken L2 data and collection. Of particular concern are some aspects of gathering L1–L2 interaction. Following this is a discussion of the matter of representativeness associated with sampling size in spoken L2 corpus design. At the end of this part, the chapter also touches on the operations of transcription and annotation in L2 corpus compilation. The second part is concerned with one of the most common ways of using spoken L2 corpora including comparisons with spoken L1 corpora in contrastive studies. To begin with, I critically review the disputes over the L1–L2 comparative approach in both SLA research and L2 corpus studies. Given that the criticisms of this approach mostly point to the L1 controls, I then discuss the requirements for spoken L1 corpora being involved in L1–L2 comparisons. Although the L1–L2 comparative approach has been popular in corpus studies on L2 production, there are many methodological issues which have so far received little attention in the academic community. Therefore, the third part considers some steps and methods which need to be taken to systematically conduct L1–L2 contrastive studies. It begins with a discussion of L1 and L2 group comparisons, then it places some emphasis on interspeaker variation associated with relevant statistical issues in L2 corpus research. Towards the end of this section, particular attention is paid to problems in terms of interpreting differences between L1 and L2 production.

## 2.2 Building a Spoken L2 Corpus

At the centre of corpus linguistics is the corpus (Kilgarriff & Grefenstette, 2003). There are varied accounts in the literature as to what exactly constitutes a corpus, but there is an increasing consensus that a corpus is a principled collection of machine-readable authentic

texts which is sampled to represent a particular language or language variety (e.g., Atkins et al., 1991; Conrad, 2002; Kennedy, 1998; McEnery et al., 2006; Meyer, 2002). Nevertheless, Kilgarriff and Grefenstette (2003) hold the view that the above definition for corpus-hood is stringent and mixes the question 'what is a corpus?' with 'what is a good corpus (for certain kind of linguistic study)?'. Their alternative definition is that a corpus is simply "a collection of texts" (p. 334). This broad definition leads researchers to welcome and use new resources, e.g., web as corpus or web for corpus building (for discussions on the use of the web in corpus research, see Hundt et al., 2007). At the same time, however, this more inclusive understanding of what constitutes a corpus risks making many of the defining features of a corpus, such as representativeness, sampling, and balance, redundant. A corpus is made for the study of language, therefore, it should be designed from a linguistic perspective with particular purposes in mind (Leech, 1992; Sinclair, 2005). This being the case, some collections of texts differ from corpora. For example, acknowledging the value of the web as a remarkable resource for researchers working on language, many linguists argue that the web is not a corpus, because "its dimensions are unknown and constantly changing" (Sinclair, 2005, p. 17), and it is not always possible to determine the authorship of the text (Kennedy, 2007). Consequently, the findings based on its contents cannot be generalised to the language under examination. Given this link between representativeness and generalisability, in this study I follow Sinclair (2005) who defines a corpus as:

> [A] collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. (p. 15)

The reason for providing the corpus definition at the outset of this chapter is to emphasise the fact that a spoken L2 corpus is not simply a collection of speech produced by L2 speakers[10]. Rather, it should be representative of a particular language under examination. As a notable pioneer in L2 corpus studies, Granger (2008) rightly points out that an L2 corpus has all the characteristics commonly attributed to an L1 corpus. Therefore, the significant considerations in the practice of L1 corpus design and construction, notably sampling, representativeness, and balance, should also be considered and evaluated carefully in the design and compilation of L2 corpora. Also, it is equally important to place some emphasis on the

[10] In this chapter, the term 'L2 speakers' covers L2 learners and L2 users.

specific features belonging to L2 corpora, including metadata relating to L2 participants (e.g., native language, proficiency level) and information on the settings (Granger, 2002, 2003, 2004).

In this section I discuss current trends in spoken L2 corpus design and compilation by reviewing the methodological decisions or strategies made in terms of four spoken L2 corpora (see Table 2): the Louvain International Database of Spoken English Interlanguage [11] (LINDSEI, Gilquin et al., 2010), the Trinity Lancaster Corpus (TLC) [12], the International Corpus of Learner Chinese (全球汉语学习者语料库), and the Guangwai–Lancaster Chinese Learner Corpus (GLCLC). These corpora were chosen according to a number of practical reasons. The LINDSEI is widely employed in studies on L2 production as in, for example, analyses of the discourse marker *well* in L2 speech (e.g., Aijmer, 2011; Buysse, 2015, 2017). Compared to the LINDSEI, the TLC is a relatively new and large-scale spoken L2 corpus which is available via the TLC Hub[13]. It is noticeable that the published work in terms of the TLC has given a detailed description of the procedure of the corpus design and creation (Gablasova et al., 2019a, 2019b), which provides a good model for subsequent corpus practices. The International Corpus of Learner Chinese (全球汉语学习者语料库) and the GLCLC are considered in this chapter as representative samples of spoken L2 Chinese corpora more broadly. By reviewing the practices of these four spoken L2 corpora in the literature, this section argues that the quality of the L2 data directly relates to the reliability of the inferences drawn from the corpus, so it is important that care must be taken in the design process of an L2 corpus. One benefit of a painstaking corpus design for L1–L2 comparative research as in this study is that a well-designed L2 corpus and a comparable L1 corpus have the ability to reflect the systematic differences in language use of two groups of speakers, and to largely mitigate the risk that the differences between L1 and L2 use are caused by the artefact of the corpus design and/or the data collection method (Ädel, 2008; Gablasova, Brezina, & McEnery, 2017).

[11] The LINDSEI: https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html. In this thesis, the LINDSEI mainly refers to the first version which was released in 2010.
[12] The TLC: http://cass.lancs.ac.uk/trinity-lancaster-corpus/.
[13] The TLC Hub: http://corpora.lancs.ac.uk/trinity/.

**Table 2**

*The Basics of the Four Spoken L2 Corpora*

| **Spoken L2 corpus** | **The Louvain International Database of Spoken English Interlanguage (LINDSEI)** | **The Trinity Lancaster Corpus (TLC)** | **The International Corpus of Learner Chinese (全球汉语学习者语料库)** | **The Guangwai-Lancaster Chinese Learner Corpus (GLCLC)** |
|---|---|---|---|---|
| **Reference** | Gilquin et al. (2010) | Gablasova et al. (2019a, 2019b) | Cui and Zhang (2011a); Zhang and Cui (2013b) | Chen and Xu (2019); Xu et al. (2019) |
| **L1 backgrounds** | 11 different L1 backgrounds: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish | the major subcorpora contain speakers from eight countries: Argentina, China, Italy, India, Mexico, Russia, Spain, Sri Lanka; it also contains data from other 22 language backgrounds, e.g., Arabic, French, Czech. | more than 120 countries and regions | 107 countries |
| **Proficiency level** | higher intermediate to advanced L2 speakers | L2 speakers from the B1 to the C2 levels of the Common European Framework of Reference for Languages (CEFR) | L2 speakers at different proficiency levels | L2 speakers at different proficiency levels |
| **Data type** | 554 interviews: each interview includes a warming-up activity (in which three topics were provided, i.e. travels, a film or a play), a free information discussion, and a four-picture description | semi-formal/institutional interaction between L1 and L2 speakers of English including presentation, discussion, interactive talk, and conversation | institutional data | oral tests administered by an L1 instructor to 1–3 test-takers; interviews conducted by an L1 speaker with individual advanced L2 speakers; free monologic talks on the topic “My Hometown” or “A Memorable Trip”; tutorials given by an L1 speaker to individual L2 speakers |

| Spoken L2 corpus | The Louvain International Database of Spoken English Interlanguage (LINDSEI) | The Trinity Lancaster Corpus (TLC) | The International Corpus of Learner Chinese (全球汉语学习者语料库) | The Guangwai-Lancaster Chinese Learner Corpus (GLCLC) |
|---|---|---|---|---|
| **Data collection method** | L2 speakers of English are interviewed by L1 speakers | one L2 speaker interacts with one L1 speaker of English in all the tasks | unclear | L1–L2 interaction: interactions typically between L1 and L2 speakers of Chinese and involve one, two or multiple speakers |
| **Number of speakers** | each subcorpus has about 50 L2 speakers | over 2,000 L2 speakers | unclear | 1,492 L2 speakers of Chinese |
| **Corpus size** | over one million words, of which about 800,000 were produced by L2 speakers | 4.2 million words | 5 million words | 621,990 words |
| **Transcription (available)** | transcription guidelines are provided | a bespoke Lancaster spoken language transcription scheme | unclear | unclear |
| **Annotation** | – | – | word segmented; error tagged | error tagged |

### 2.2.1 Target L2 Speakers

Drawing on L2 corpus practices and theoretical discussions on L2 corpus design (e.g., Breiteneder et al., 2006; Gablasova et al., 2019b; Gilquin et al., 2010; Granger, 1998a, 2003), Table 3 summarises some variables with respect to L2 informants that are of relevance to the compilation of a spoken L2 corpus, which should be documented as far as possible in metadata. As Burnard (2005) puts it, metadata restores and specifies the contexts, and in a typical corpus analysis, it relates the target language patterns that are taken out from the corpus to the contexts in which they originally occur. This in turn enables researchers to characterise patterns in linguistic production by certain groups in certain situations. According to specific research purposes, L2 corpus compilers normally prioritise certain variables, such as L1 backgrounds, gender, and language proficiency level, and use them as the criteria to recruit L2 participants. Bearing this in mind, this section is concerned with several key variables considered in spoken L2 corpus design based on the features of the four spoken L2 corpora.

**Table 3**

*Variables in Terms of L2 Speakers in Spoken L2 Corpus Design*

- L1 background
- nationality
- ethnicity
- language proficiency level
- gender
- age
- whether studied in the target language speaking country
- duration living there
- years of exposure to the target language
- educational level
- occupation
- prior acquaintance with one's interlocutor

The first variable that is taken for granted in spoken L2 corpus design and construction is the matter of L1 backgrounds for L2 speakers. Table 2 shows that the four spoken L2 corpora feature the speech of L2 informants with multiple L1 backgrounds, especially the two spoken Chinese corpora. From the point of view of corpus design, greater diversity in the L1 status of contributors to an L2 corpus does not necessarily result in greater representativeness or balance. As Biber (1993) puts it, to achieve representativeness it is necessary to undertake "a thorough definition of the target population and decisions concerning the method of sampling are prior

considerations" (p. 243). In spoken corpus design, it therefore seems much less important to cover L2 speakers with various L1 backgrounds in comparison to the task of delimiting the boundaries of the target L2 population (e.g., whether heritage speakers of a language should be included in the corpus). Moreover, it is equally important to pay attention to the organization of categories within the target L2 population, such as gender and L2 proficiency. Without a well-defined definition of what the samples are intended to represent, it therefore is difficult to evaluate the representativeness of such a spoken L2 corpus.

From the perspective of implications, corpora consisting of multiple L1 backgrounds for L2 speakers are of great interest to researchers who aim to explore the potential influences of various L1 backgrounds on L2 production. By using such a spoken L2 corpus, researchers are able to investigate whether a particular linguistic feature is unique to L2 speakers with specific L1 backgrounds or common to all L2 speakers of the target language (Johansson, 2009). It is essential to bear in mind that the prerequisite for a reliable corpus analysis of the influences of L1 backgrounds on L2 production is to make other variables as similar as possible, such as that the L2 speakers should be at the same proficiency level and have the same tasks. Given this consideration, the LINDSEI is a good source for L2 corpus studies, as each subcorpus of the LINDSEI is designed with the same criteria which represents the speech of L2 speakers with a specific L1 background. Drawing on this design, the LINDSEI has benefited numerous L2 studies. For instance, Buysse (2015) compares the use of the discourse marker *well* by L2 speakers of English with different L1 backgrounds by using the Dutch, Spanish, German, Chinese, and French components of the LINDSEI. Although the LINDSEI contributes greatly to the academic community, it is worth noting that the diversity of linguistic backgrounds for L2 speakers does not have any bearing on the value of an L2 corpus. It is not the case that a spoken L2 corpus must cover multiple L1 backgrounds for L2 speakers, and it is always imperative to build a spoken L2 corpus that meets specific research purposes. According to this view, L2 corpora recording a single L1 background, such as the Country–Specific Corpus of Spoken Chinese Interlanguage (Korean Learners) (韩国学习者汉语中介语口语语料库), can be representative and make contributions to the research community as well (for more information on this corpus, see Table A-8 in Appendix A).

In addition to L1 backgrounds, spoken proficiency[14] is among the most significant features in spoken L2 corpus design. L2 corpus compilers have good reason to be attentive to

[14] In this chapter, the term 'proficiency' is used in a narrow sense in which it has come to refer to oral skills.

the matter of proficiency. In L2 corpus design, well-defined proficiency levels have the potential to fine-tune the selection criteria of the target L2 speakers of a language, thereby contributing to the representativeness of the corpus. Also, careful consideration of the assignment of L2 proficiency can lead to accurate conclusions about L2 language development and use (Carlsen, 2012; Gablasova et al., 2019b). Unfortunately, L2 proficiency is sometimes a poorly controlled factor in spoken L2 corpus design as well as in SLA research (Callies, 2015; Carlsen, 2012; Thomas, 1994, 2006). In practice, it is noticeable that the spoken L2 corpora in existence vary in the scope of their interest in the diverse skills subsumed under the term 'proficiency'. The LINDSEI and the two spoken L2 Chinese corpora do report that they contain L2 speakers at different levels of proficiency (see Table 2 above), but they do not provide enough details to explain the assessments of proficiency levels they performed. As a result, it is hard to make sufficient judgements about the pros and cons of their assessments and to evaluate to what extent these corpora are representative of their target L2 population.

In spoken L2 corpus design, proficiency levels of L2 participants have usually been operationalised by means of external criteria such as the institutional status of the L2 speakers or the time spent learning the L2 (see Callies, 2015; Carlsen, 2012; Granger, 1998a, 2004, 2008). Global measures for assessing L2 proficiency based on external criteria have been supported by Granger (1998a), where she has also claimed that after this initial assessment, "it will be the researcher's task to characterize learners' proficiency in terms of internal evidence" (p. 9). In her following works on L2 corpus design criteria, Granger sticks to her claims about the way to establish proficiency in L2 corpora (e.g., Granger, 2004, 2008); the external assessment in terms of proficiency is in line with the widely accepted principle, which was addressed earlier by Atkins et al. (1991), namely that a corpus should be built based primarily on external criteria. In contrast to Granger's statements, some researchers make the claim that the proxy measures based on external criteria are unreliable to assess L2 proficiency (e.g., Callies, 2015; Carlsen, 2012; Gablasova et al., 2019b; Thomas, 1994, 2006; Tono, 2003). As early as 1994, Thomas (1994) warns against assigning proficiency levels on the basis of institutional status, because institutions differ considerably "in the standards by which they assign a given status to L2 individuals, and in the rigidity with which those standards are maintained" (p. 317). Given that the International Corpus of Learner Chinese (全球汉语学习者语料库) contains material from different universities, without sufficient measures of proficiency assessment, the varying levels of proficiency in this L2 corpus may introduce a

source of errors and affect the validity of intergroup comparability with respect to proficiency (Tono, 2003).

Drawing on these debates, Granger (2012) claims that one possible way of improving the reliability of L2 proficiency assessment is to "complement the ethnographic data with additional data obtained, for example, by submitting students to standardized questionnaires (motivation test, aptitude test, general proficiency test, vocabulary test)" (p. 9). The means of using standardised questionnaires seems to fall in the category of in-house placement tests and research-internal tests summarised by Thomas (1994). As Thomas puts it, the advantage of using in-house placement/research-internal tests is that L2 participants can all be tested by using the same tests which at least establishes the internal consistency of proficiency. On the other hand, the validity and reliability of privately developed tests vary and the results of such tests cannot be readily generalised to a wider population outside of a project. Furthermore, there is the problem of criteria used in partitioning proficiency groups. That is, how to divide L2 proficiency levels on the basis of test scores. Failing to provide sufficient accounts of the above concerns in the descriptions of corpora, the measures of using institutional status, in-house placement tests, and research-internal tests to assess proficiency are dubious bases for gathering data to L2 corpora and generalising corpus results.

To establish a reliable estimate of L2 proficiency in L2 corpus design, the TLC employs standardised tests to assess L2 proficiency. Specifically, it bases the assessment of proficiency on L2 speakers' performance in the speaking tasks by using the proficiency ratings which are awarded by trained raters of the Graded Examinations in Spoken English (GESE)[15]. In order to make the ratings given in the GESE exam comparable outside of this corpus, the marks then are validated using the bands of the Common European Framework of Reference for Languages (CEFR) (Gablasova et al., 2019b). The TLC shows that a reliable proficiency assignment is viable "by applying insights and practice from the professional field of language testing and assessment" (Carlsen, 2012, p. 161). It should be noted that the TLC is different from the other three L2 corpora in that it is a collection of oral examinations. In the case of the TLC, the benefits of using standardised test scores are apparent: (i) the content of the GESE exam "is available for public scrutiny" (Thomas, 1994, p. 324), which enhances the reliability and transparency of L2 proficiency assignment, and (ii) the ratings as recognisable benchmarks enable the corpus compilers to select the ideal L2 participants for this corpus. Acknowledging

[15] This is an exam developed and administered by Trinity College London.

the advantages of this measure, it is not always the ideal for any spoken L2 corpora. As regards L2 proficiency of spoken Chinese, the two notable standardised oral tests are *Hanyu Shuiping Kouyu Kaoshi* (HSKK)–the oral speaking test of HSK (*Hanyu Shuiping Kaoshi*–the Chinese proficiency test) and the Test of Practical Chinese (C. Test). As we will see in Chapter 3, none of the L2 participants involved in the corpus built in this thesis took any standardised oral tests. Therefore, it was impossible to assess L2 proficiency by using scores of standardised tests in this thesis. It also is questionable to assess L2 proficiency levels on the basis of scores of certain standardised oral tests which were obtained a long time ago, for the reason that such scores may misrepresent L2 speakers' abilities at the time of data gathering. The measure used by Gráf (2017) to assess L2 proficiency for the Czech component of the second version of the LINDSEI, however, may provide some insights into spoken L2 Chinese corpus design. According to Gráf (2017), the corpus compilers invited three professional IELTs examiners who had been previously trained to rate proficiency in accordance with CEFR levels to assess L2 proficiency. Certainly, this method can increase the reliability of proficiency assignment; it still is worth considering carefully the assessment standards which professional examiners will use to award the ratings in order to meet specific research purposes.

So far, I have illuminated the advantages and the disadvantages of different measures for assessing L2 proficiency in relation to spoken L2 corpus design. As "a fuzzy variable" (Carlsen, 2012, p. 161), proficiency relates to some variables listed in Table 3 above, such as years of exposure to the L2 and whether study has taken place in an L2 environment, therefore such proficiency-related variables will not be discussed further in this chapter. Consideration with respect to L2 proficiency of L2 speakers of Chinese will be given in Chapters 3 and 7.

At the end of this subsection, it should be noted that gender has been a popular variable in corpus design. In his review of the construction of the BNC1994, Burnard (2002) proposes the notions of selection criteria and descriptive criteria. Selection criteria include variables that should be predefined, and a target proportion need to be identified for each. Descriptive criteria are not controlled during data collection but are recorded to "maximise variability" (Burnard, 2002, p. 6). It is evident that gender is typically regarded as a selection criterion in terms of participants in corpus design (e.g., Gablasova et al., 2019b; Leech, 1993; Love, 2020; Love et al., 2017; Svartvik, 1990). By using corpora that are encoded with such sociolinguistic metadata, researchers can conduct gender studies (e.g., Baker, 2014), or examine the potential influences of gender on language production to give a more accurate and systematic description of the usage of particular patterns (McEnery et al., 2006). Gender is often characterized as a

stable male/female binary in corpus projects, despite the existence of intersex and trans people (Baker, 2014, p. 209). An exception recently is the Spoken BNC2014 in which the compilers replaced the female or male prompt by a free-text box to "avoid presupposing that all participants would willingly describe their gender in this binary fashion" (Love, 2017, p. 46). It seems that gender studies have improved the corpus compilers' awareness of the complexity of gender categories; as a result, it is necessary to choose an appropriate way to gather gender information. And, the strategy of collecting gender adopted by the Spoken BNC2014 project can be extended in other corpus projects. Returning to the four spoken L2 corpora, it is unclear which methods they used to gather gender information.

In addition to the matter of the male/female binary division, it seems that, in corpus design, there is a desire to represent female and male participants in equal numbers in order to achieve the balance of a corpus. Ideally, in corpus studies, each gender group should have an equal number of participants and individuals should contribute the same amount of speech, therefore neither each group's nor individuals' speech idiosyncrasies will skew the results. In practice, it is not necessarily the case that a corpus must be gender-balanced nor always feasible to create a gender-balanced corpus. From a perspective of balance, it is problematic to underrepresent males or females in a corpus. However, in terms of corpus representativeness, a completely equal number of participants/word counts for both genders in a corpus may not represent the distribution of the population. Ritchie and Roser (2019) provide an overview of the gender ratio—the number of males relatives to females in a society—across the world, and the conclusions they come to include: (i) in almost every country births of males outnumber births of females, and (ii) the gender ratio tends to decrease from being male-biased to female-biased over the life span. One thing these conclusions tell us clearly is that males and females are not distributed evenly in any population, but it does not mean that a gender-balanced corpus or a corpus that is female-biased is poorly representative of the population under examination. Corpus representativeness is difficult to evaluate (see 2.2.3), and it is not uncommon that compromises have to be made between what can maximise representativeness and what is practicable in corpus practices (see Chapter 3). Also, if researchers are simply interested in female speech, for example, it seems reasonable to build a corpus featuring only female participants. To sum up, gender information should be recorded in corpus design and relevant decisions should be made consistent with the specific research purposes. Further discussions in terms of the matter of gender will be discussed in Chapters 5 and 7.

Apart from the above variables, there are other factors in terms of L2 speakers that should be considered in L2 corpus design as well. Given that the other variables listed in Table 3 are less questionable than the variables highlighted in this section, and while the discussion above covers the most challenging decision areas, a thorough discussion covering all the variables pertinent to L2 speakers in spoken corpus design is beyond the scope of this chapter. In this thesis, descriptions about the variables regarding L2 speakers of Chinese will be given in Chapter 3. Ensuring an appropriate selection of target L2 participants is the first step to corpus compilation, and the next step is to gather data. In what follows, I thus reflect on variables which relate to spoken L2 data and collection.

### 2.2.2 Spoken L2 Data and Collection

In principle, spoken corpora can contain different types of data for specific linguistic research. There are some variables that need to be considered prior to data collection in spoken L2 corpus design, such as data types, settings, monologues/dialogues, relationship between interlocutors, and the medium of data collection. In this subsection, I first give a brief discussion of the data types in the four spoken L2 corpora, then pay attention to the aspects in terms of L1–L2 interaction.

Granger (2012) points out that L2 corpus data falls within the L2 data types distinguished by Ellis (1994) in SLA research, namely natural language use and clinically elicited data. Natural language use data occurs in the course of using the target language in authentic communication that L2 speakers engage in "when they are not being studied" (Ellis, 1994, p. 270). The main strength of naturalistic data is that it provides information about what L2 speakers actually do with their L2 knowledge and skills when they engage in naturally occurring communication. It is, however, recognised that gathering fully L2 natural speech data is rather difficult and time-consuming and may encounter ethical issues, so studies of natural speech are normally carried out with limited data which cannot provide researchers with sufficient evidence to draw reliable inferences in SLA research. Following this view, Granger (2012) claims that difficulties in collecting L2 natural speech data to some extent lead SLA researchers to resort to elicited data, such as role plays or picture descriptions. For some SLA researchers, elicited data have certain advantages that natural speech data do not have. For example, elicited data can provide researchers with greater control over the data, and the

likelihood of capturing relevant types of linguistic output can be maximised (Ädel, 2015; Ellis, 1994).

Elicited data have been popular in spoken L2 corpora as well. It is noticeable that the four spoken L2 corpora given in Table 2 are primarily composed of elicited data. Nonetheless, L2 speakers of a language are required to produce speech in many more registers than oral tests, academic discussions, or picture descriptions. It is true that in many cases these forms of L2 speech can be easily recorded and analysed by language teachers or researchers at school or university. But there is good reason to believe that naturally occurring communication in daily life is among the forms of speaking that we suppose L2 speakers, especially those who study or live in an L2-speaking country, may inevitably engage in. Spontaneous spoken data has already received considerable attention in the area of spoken L1 corpora (e.g., more recently the Spoken BNC2014). With the development of technologies, the difficulty in recording L2 natural speech has largely been mitigated, though spoken data collection (both L1 and L2) is still much more time-consuming and labour-intensive than written data collection. Unfortunately, the preference for elicited data appears not to have changed very much until now. L2 natural speech data are still absent from existing spoken L2 corpora and the variety of these L2 corpora is rather limited. Admittedly, the type of data closely reflects specific research goals. It is worth being alerted to the concern about whether the dominant elicited data in L2 corpora and relevant corpus studies have resulted in a skewed perspective on L2 speakers' capabilities for communication, conceiving of them as deficient communicators who are striving to achieve a certain degree of nativelikeness (Firth & Wagner, 1997). Following this view, corpus compilers should be encouraged to collect L2 natural language use in communication to promote the diversity of L2 speech in corpora and to broaden the traditional database for L2 studies.

Another feature of the spoken L2 corpora that can be observed in Table 2 is that the LINDSEI, the TLC, and the GLCLC contain spoken L2 data in L1–L2 interactions. Some decades ago, communication between L1 and L2 speakers inside and outside classrooms became a popular focus in SLA studies (e.g., Firth & Wagner, 1997; Gass & Varonis, 1985, 1994; Long, 1983a, 1983b; Varonis & Gass, 1985). More recently, available spoken L2 corpora of L1–L2 conversational interaction have provided a bounty of evidence to expand the range of research interests in L2 performance (e.g., Aijmer, 2011; Chen & Xu, 2019; De Cock, 2004; Gablasova, Brezina, McEnery, et al., 2017). In corpus studies making use of L1–L2 interactional data, there is a tendency to exclude L1 contributions and focus entirely on L2

performance. It is not my intention to criticise this L2-centred method, but rather to stress the potential effects of L1 interlocutors on L2 performance. Using corpora of L1–L2 interaction in this way does not indicate that L1 interlocutors play no role in the process of L2 data collection. From the point of view of corpus design, variables with respect to the L1 interlocutors in L1–L2 interaction should be considered carefully and recorded in metadata. On the basis of the LINDSEI, the TLC, and the GLCLC, the role of the L1 interlocutor can be occupied by a college student, a teacher, or an examiner who is a native speaker of the target language. The roles of the L1 interlocutors largely indicate the roles of the L2 speakers in interaction. For example, the TLC contains spoken data produced by L1 and L2 speakers of English, but it is also appropriate to say that this corpus contains the language of interlocutors in the roles of examiners and examinees. This view may give rise to a question: in the case of the TLC and the like, to what extent are we confident to say that the L2 language is produced by the interlocutors in the role of being non-native speakers rather than of being examinees. An in-depth consideration of this question will enable researchers to provide more scientific explanations to corpus results. Based on this view, it is worth taking into consideration the relations between the L1 and L2 interlocutors in L2 corpus design. The same issue existed in this study when building the spoken Chinese corpus (see Section 3.5 in Chapter 3). A further discussion in terms of the relation between L1 and L2 speakers can be found in Chapter 5.

Aside from the above issue, there are other issues that need to be considered in corpus deign. Table 2 shows that each L1–L2 interaction can have different L1 and L2 speakers. Or, the same L1 speaker can engage in each L1–L2 interaction, e.g., the spoken L2 corpus built in this study. In the latter circumstance, the advantage is that the L1 speaker's vernacular style is largely repeated across the interactions, so it is feasible to evaluate the effect of the L1 speech on the L2 output. Another important concern is the medium through which L1–L2 interactions can be conducted. Traditionally, the qualitative data is mostly gathered face-to-face. With change in digital communication technologies over the last few decades, there are many feasible alternatives to face-to-face interaction, such as telephone and the internet techniques (e.g., Skype, and more recently Zoom). The advantages of using internet technologies and the telephone as research mediums to gather data have been discussed in the literature, such as saving travelling costs and offering greater flexibility in time and location of data collection than face-to-face data collection (e.g., Iacono et al., 2016; Irvine, 2011; Irvine et al., 2013; Janghorban et al., 2014; Lobe et al., 2020). To repeat a point made above, research mediums need to be documented in metadata too, especially in the case that different mediums are

adopted in a corpus project (see Chapter 3), so that corpus users can evaluate the potential influences of specific mediums on language production whenever necessary.

Until now, I have discussed some crucial variables relevant to L2 speakers as well as L1–L2 interaction in L2 corpus design. Other variables such as the length of and the number of interlocutors involved in each interaction which relate to sample size will be discussed in the following subsection.

### 2.2.3 Corpus Size and Representativeness

In his renowned book *Corpus, concordance, collocation*, Sinclair (1991) shows clearly his attitude to corpus size: "[t]he only guidance I would give is that a corpus should be as large as possible, and should keep on growing" (p. 18). The main virtue of a corpus being large is that "the underlying regularities have a better chance of showing through the superficial variations" (Sinclair, 2004, p. 189). If the regularities found in a corpus can be generalised to some entire language or language variety, this corpus then is thought to be representative of that language or language variety. In other words, the larger a corpus is, the more representative it will be. The optimal size of a corpus however is determined by many factors and there are no explicit criteria to decide the size for a large corpus. Clearly, the dimensions of a so-called large corpus vary with the date it was compiled: half-a-million-word spoken corpora (e.g., the LINDSEI and the GLCLC) tend to be regarded as small by today's standards. Moreover, compared to some large-scale spoken L1 corpora, such as the Spoken BNC2014 which contains 11.5 million words, the TLC and the International Corpus of Learner Chinese (全球汉语学习者语料库) which are large in comparison with other spoken L2 corpora are relatively small. Additionally, Biber (1993) contends that size is not the most significant consideration in achieving corpus representativeness. A well-designed corpus can be representative of a particular language but small in size. For example, the LINDSEI is well-designed according to some strict criteria, while its subcorpora are relatively small with approximately 100,000 words each (Gilquin et al., 2010). By using some components of the LINDSEI and its native speaker counterpart, the Louvain Corpus of Native English Conversation (LOCNEC), which is also a small corpus consisting of 170,533 words, a number of corpus studies on L2 use of English discourse markers (e.g., *well*) have been carried out and contributed greatly to our growing understanding of the similarities and differences between L1 and L2 use (e.g., Aijmer, 2011; Buysse, 2015,

2017). With the above discussions in mind, compilers are encouraged to make efforts to create large spoken L2 corpora to provide sufficient evidence to L2 studies, but it should be also borne in mind that an L2 corpus will only be useful "if it has been compiled on the basis of strict design criteria" (Granger, 2012, p. 9).

As regards corpus representativeness, there is agreement that representativeness is not simply affected by the overall size of a corpus, but is vague and difficult to measure (Leech, 2007; McEnery et al., 2006; Sinclair, 2005; Weisser, 2016). Because of this, Kilgarriff and Grefenstette (2003) hold the view that the matter of representativeness is "a pressing yet almost untouched practical and theoretical issue" in corpus linguistics (p. 340). The difficulty of achieving representativeness nevertheless does not detract from the actuality that it is worth aiming at (Leech, 2007; Sinclair, 2005). Corpus linguists have provided strategies to ensure representativeness. As has been discussed previously in this section, "a thorough definition of the target population and decisions concerning the method of sampling are prior considerations" (Biber, 1993, p. 243) in achieving representativeness. In addition, Sinclair (2005) claims that the most important step towards achieving representativeness is to fully document all steps concerning corpus design and building, so that users can "inspect not only the contents of a corpus but the reasons that the contents are as they are" (p. 9). These proposals for improving the representativeness of a corpus have been used when building the spoken Chinese corpus in this thesis (see Chapter 3 for detailed information).

Practically, to improve the representativeness of spoken L2 corpora, the number of L2 speakers that produce the data need to be considered as well in corpus design. Table 2 above shows that both the TLC and the GLCLC feature a large quantity of L2 informants (over 2,000 and 1,492 respectively), while each subcorpus of the LINDSEI includes about 50 L2 speakers. It seems appropriate to state that, other things being equal, the more L2 speakers a corpus features, the more representative and balanced the corpus will be. In addition to the number of L2 speakers in a corpus, it is equally important to take into consideration the size of the sample text produced by each L2 speaker. All else being equal, an impediment to representativeness is that very limited output of each individual is contained in a corpus featuring a large number of L2 speakers, for this corpus may lack the ability to reliably represent each speaker's language use. For instance, other things being equal, a spoken corpus that consists of 100 ten-minute recordings of interaction may be more representative than a corpus including 500 2-minute recordings of interaction, because the texts in the former corpus may be long enough to represent the distributions of linguistic features (Biber, 1993). It should be noted that it is not

uncommon in practice that some sample sizes are larger than others in a corpus. In this case, different strategies have been taken with reference to sample size. The first computer corpus, the Brown Corpus, which is a sample of American printed English of the year 1961, contains 500 2,000-word sample texts (Francis, 1980). This model has been followed by a string of successors, notably the Lancaster-Oslo/Bergen Corpus (LOB)[16]. Corpora of this type are typically fixed at a particular size and usually contain relatively short samples, so they are called *sample corpora* by Sinclair (1991) who provides a different perspective on the sample corpus approach. Although the 2,000-word sample of the Brown style has been perpetuated in later corpora with the purpose of carrying out contrastive research, Sinclair (2005) insists that there is no virtue from a linguistic point of view in selecting samples of all the same size (also see 5.2 in Chapter 5). Further, Sinclair (2005) states clearly that:

> Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size. (p. 7)

Returning to the four spoken L2 corpora in Table 2, it is unclear which strategies they employed in their design, but it is evident that most of the well-known spoken L1 corpora are very much consonant with the suggestion of Sinclair (2005). In this thesis, the issue of size was considered, and detailed discussions in terms of varied sample sizes will be given in Chapter 5.

In conclusion, both representativeness and balance are closely linked to corpus size. Theoretically, as discussed above, representativeness and balance can be achieved by adopting systematic and painstaking sampling strategies. In practice, as "many resources of bias are not so readily predictable" (Milroy & Gordon, 2003, p. 25), compromises might be made between the theoretically desirable and what is realistically possible (Crowdy, 1993). Having come this far, it is of great value to highlight Sinclair's claim (2005) regarding these paramount corpus features which is applicable to any corpora:

> The corpus builder should retain, as target notions, representativeness and balance. While there are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components. (p. 9)

[16] For detailed information, see http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html.

### 2.2.4 Transcription and Annotation

Transcription is an integral component in spoken corpus compilation (Crowdy, 1994; Kirk & Andersen, 2016), and is "effectively the first stage of analysis and interpretation" (Cameron, 2001, p. 43). Thus, transcription quality has direct implications for the reliability and usability of a spoken corpus (see Chapter 4). In the literature on English, the significance of transcription has been addressed by a number of scholars, not only in spoken corpus design and compilation (e.g., Crowdy, 1994; Edwards, 2014; Kirk, 1992; Leech et al., 2014; Love et al., 2017; Schmidt, 2016), but also in studies employing discourse analysis (e.g., Du Bois, 1991; Du Bois et al., 1992; Du Bois et al., 1993; Ochs, 1979). In contrast to the fruits of transcription in studies on spoken English, transcription has received disproportionately little attention in its own right and has seldom been discussed in great depth in the Chinese research literature[17]. It should be noted that transcription here refers to the representation of *Putonghua* in written forms. In the following review of work published in *China National Knowledge Infrastructure* (CNKI)[18] and selected papers from a conference focusing primarily on Chinese learner corpus research—*the International Symposium on the Construction and Application of Chinese Interlanguage corpora (汉语中介语语料库建设与应用国际学术讨论会)*[19], I will demonstrate some main issues of transcription in spoken Chinese corpus construction.

To date, a very limited number of articles have focused exclusively on transcription in Chinese. Some of these studies are devoted to introducing and discussing current widely used transcription schemes intended for research in English (e.g., Duan, 2010; He, 2020; Liu, 2016; Tao, 2004), such as the Santa Barbara school conventions (Du Bois et al., 1992) and Gail Jefferson's transcription notation (Jefferson, 1983). While the accounts of current transcription conventions offered in previous Chinese studies are too brief, features of current transcription schemes are rarely explained comprehensively to the Chinese research community. As a result, such studies on transcription in the Chinese research literature neither contribute greatly to Chinese researchers' knowledge of current English transcription schemes nor to sufficiently guide transcription for spoken Chinese corpus construction. Moreover, discussions in previous studies on transcription remain highly theoretical; nonetheless, crucial theoretical issues

[17] Here, the discussion is limited to the scope of spoken monomodal corpora. Transcription for speech corpora and multimodal corpora are beyond the scope of this study.
[18] CNKI is a key national research and information publishing institution in China: https://www.cnki.net/.
[19] It is the only conference on Chinese learner corpus research being held in mainland China to date that I am aware of. This conference has published five books from 2011 to 2021, containing 152 conference papers in total. While there are only two papers focus exclusively on transcription, and two of three papers introducing spoken L2 corpora give brief accounts of transcription.

surrounding transcription, such as the nature of transcription, the issue of standardisation (e.g., the categories of spoken features to be recorded and codes for these categories), and the consistency of transcripts, have so far not receive enough attention in the Chinese research community. In the absence of meaningful and exhaustive theoretical discussions on these issues, as we will see later in this subsection, transcriptions are conducted rather crudely in practice.

On the other hand, some researchers have begun to look at transcription in a relatively methodical way. To my best knowledge, the paper written by Lu et al. (2014) is the only work offering detailed conventions for orthographic transcription in the Chinese literature at the time of writing, although their focus is solely on features such as utterance-final articles and minimal response tokens (e.g., 啊 *a* 'ah', 哦 *o* 'oh') in Chinese conversations. Lu et al. notice a variety of pronunciations of such features in their data and exemplify their observations by following the Santa Barbara school conventions. However, Lu et al. (2014) also argue that it is not always feasible for transcribers to identify all variations of these features in real conversations, for it may require a high level of expertise and considerable time to characterise these features. This being the case, to maximise transcription consistency in a spoken corpus, it is necessary to limit the set of such Chinese features into "a categorially justified minimum" (Andersen, 2016, p. 343). An account of transcribing minimal respoken tokens for the spoken Chinese corpus can be found in Chapter 4.

Unfortunately, such considerations when transcribing specific features in spoken Chinese have not received much attention in the Chinese academic community. Gail Jefferson's transcription notation remains an influential system which is adopted by researchers to guide transcriptions in spoken L1 Chinese corpora (e.g., Quan, 2017; Y. Wang, 2016). There are good reasons for researchers to follow such well-known transcription systems to transcribe speech in written forms. Transcription is not a straightforward task; the first issue that researchers encounter is how to identify and represent features appropriately. Following current transcription schemes, researchers can directly use the standards to identify Chinese features and mitigate the burden of finding symbols to represent those features. At the same time, this has resulted in some issues, in that studies in the Chinese research literature often provide only a cursory description of the transcription process, claiming to record features including pauses, repeats, and paralinguistic features in the transcripts. Clearly, these features are ubiquitous and important in speech, but it is unclear why these features receive more attention than other features, or whether other features are captured in the transcripts as well.

Similarly, with reference to transcription in spoken L2 Chinese corpora, it is noticeable that the great majority of studies have mentioned transcription in passing. There is a sentiment that compilers of spoken L2 Chinese corpora prefer to create their own transcription schemes (e.g., Hu & Xu, 2020; Xu et al., 2021; Zhang, 2019). However, transcribing is still not a prime focus of concern for most researchers working on spoken L2 Chinese corpora; therefore, it is not surprising that the transcription process itself is seldom described in great depth or detail in studies on spoken L2 Chinese corpora. For example, the transcription principles of the two spoken L2 Chinese corpora listed in Table 2 are not transparent. Having come this far, a very important issue that needs to be addressed then is the consistency of transcribing. As a reminder, the International Corpus of Learner Chinese (全球汉语学习者语料库) contains material gathered by a number of scholars from different universities (see 2.2.1 and Appendix A), which indicates that the data were prepared by a number of transcribers. This being the case, the issue of the consistency of transcripts manifests itself in the transcription process. As Andersen (2016) puts it, "[t]he key word in spoken corpus transcription is consistency" (p. 324). In the literature on English, this issue has received sufficient attention and been discussed both theoretically and practically (e.g., Cucchiarini, 1996; Edwards, 2014; Gablasova et al., 2019b; Garrard et al., 2011; Love et al., 2017), as high agreement between transcribers guarantees great transcription accuracy and accordingly the high quality of transcripts provide more reliable data to corpus studies. To my best knowledge, there is no study that pays attention to the issue of consistency to date in the Chinese scholarly publications. In consequence, without documented transcription schemes and transparent transcription processes, it is unlikely that users can evaluate whether the results of a corpus study on L2 production reflect the real features of L2 speakers of Chinese or whether they are affected by the transcription decisions in some degree.

Acknowledging the issues in L2 transcription, it is admitted that some researchers have made efforts to pay more attention to spoken L2 corpus transcription (e.g., Hu & Wang, 2011; Liu, 2020; Zhang, 2016). Such discussions to some extent promote researchers' awareness of the importance of transcription, and also provide some useful suggestions to the transcription for the spoken Chinese corpus built in this thesis. In Chapter 4, I give a detailed account of how the transcription was conducted, seeking to encourage Chinese researchers to look at transcription in a more systematic way.

In contrast to the absence of detailed discussions of transcription in L2 Chinese corpus design, annotation has had considerable attention in the Chinese literature on L2 production

(e.g., Xiao & Zhou, 2014; Zhang, 2013, 2019; Zhao & Lin, 2019). It is noticeable that the two spoken L2 Chinese corpora in Table 2 (as well as other L2 corpora in Appendix A) are error-tagged. Annotation, e.g., error tags and part-of-speech (POS) tags, is concerned with a "more abstract relationship" (Edwards, 2014, p. 20) between linguistic units, which is an important step taken to make the transcribed data machine-readable. Nowadays, computer technologies have enabled linguists and researchers to annotate data automatically. For instance, McEnery et al. (2019) point out that annotation tools are now well embedded in stand-alone packages such as Sketch Engine (Kilgarriff et al., 2014) or #LancsBox (Brezina et al., 2020). Also, there are many useful tools developed for Chinese, e.g., the International Corpus of Learner Chinese (全球汉语学习者语料库) includes an annotation tool[20] which enables users to separate Chinese characters into meaningful units[21] and annotate the segmented units with POS tags. In the field of corpus linguistics, it is widely accepted that annotation can enrich a corpus with all sorts of mark-ups, because annotated data allow for more meaningful searches for corpus data motivated by linguistic categories (e.g. all nouns in a corpus) rather than looking for a single word; then the future use and analysis of the corpus will be considerably facilitated (McEnery et al., 2019; McEnery & Hardie, 2012; Meyer, 2002; Rayson, 2015).

Conversely, Sinclair (2004) cautions against the indiscriminate use of tagging in studies and claims that "[t]he interspersing of tags in a language text is a perilous activity, because the text thereby loses its integrity, and no matter how careful one is the original text cannot be reliably retrieved" (p. 191). He points out that it is often the case that as long as a text is marked up with tags, the computer tends to work with the tags and ignore the language. Annotation is fundamentally interpretative in nature, and the tagging rules underlying varied annotation tools represent different understandings of the language; in this regard, Sinclair maintains that the use of annotated data in a corpus study is an alternative way to study the language. In the case of Chinese annotation, many annotation tools are developed mainly on the basis of written L1 data, so such tools are not likely to be sensitive to certain linguistic features in spoken Chinese. For example, the discourse marker 就是 *jiushi* 'be exactly' that will be studied in Chapter 6, tends to be tagged as an adverb[22] in the data I gathered. If researchers observe the data directly

[20] http://qqk.blcu.edu.cn/#/search/strTools.
[21] The fact that there are no spaces between Chinese characters makes the data impossible to be read and analysed by computer without word segmentation, so dividing characters into meaningful units is an inevitable step taken to process both spoken and written Chinese automatically.
[22] There is a Chinese website which holds many annotation tools (http://clr.ccnu.edu.cn/software.action?page=1). I have used one of the tools to process my data, and it has turned out that all the occurrences of 就是 *jiushi* were tagged as adverbs.

through the tags and pay no attention to the language per se, they may fail to find out that 就是 *jiushi* 'be exactly' in some situations has pragmatic functions in speech. Therefore, it is essential to bear in mind that annotated corpora should not be overestimated and plain texts are important in corpus research, and that annotated spoken texts should always be approached with caution.

Drawing on the debates over annotation, whether an L2 corpus should be used in raw format or needs to be enriched with annotation very much depends on the object of study (Granger, 2012). It should be borne in mind that transcription and annotation are two important steps in computerising data when building a spoken corpus, thus the relevant procedures need to be transparent and documented in detail. Consideration of transcription with respect to the spoken Chinese corpus built in this thesis will be presented in Chapter 4. As we will see later in this thesis, annotation has not been included in the procedure of the design and compilation of the spoken Chinese corpus, but some discussion of this decision will be given in Chapter 8.

## 2.3 The L1–L2 Comparative Approach

In this section the attention shifts from accessing spoken L2 data to making use of the data. Certainly, an L2 corpus can be used alone to investigate L2 production (e.g., Chen & Xu, 2019; Gablasova & Brezina, 2015; Gablasova, Brezina, McEnery, et al., 2017). It can also be compared to other L2 corpora to examine differences between L2 groups. In the latter case, an L1 corpus is often taken as the source of the predictors to show what is considered acceptable in that target language (e.g., Buysse, 2017; Gries & Wulff, 2013; Martinez-Garcia & Wulff, 2012). This L1–L2 comparative approach is the focus of this section, but the discussion is restricted to the situation in which one spoken L2 corpus is compared to another spoken L1 corpus. To begin with, I critically review the debates towards L1–L2 comparisons in L2 studies, which primarily point to L1 controls. Then I turn to consider the demands for the L1 corpora that are used in L1–L2 contrastive studies from the perspective of corpus linguistics.

### 2.3.1 Debates on Comparisons with L1 Speakers

The L1–L2 comparative approach has been very influential in corpus studies on L2 production in both written and spoken language (e.g., Aijmer, 2011; Callies, 2013; De Cock, 2004; Durrant

& Schmitt, 2009; Gilquin & Paquot, 2008; Gilquin, 2008; Gries & Wulff, 2013). Typically, the L1 corpus used in such L1–L2 comparisons is to be the source of the standards of language use, showing what are considered acceptable in the target language. The underlying assumption in terms of using L1 controls is that "success in acquisition research is typically defined in relation to degree of nativelikeness" (Lardiere, 2013, p. 675). From the point of view of SLA research, Birdsong (2005, 2006) claims that the most basic motivation to continue employing L1 controls is descriptive: by observing L1 performance, researchers can bypass presumed norms and draw inferences with respect to L2 acquisition based on empirical data. Thus, researchers "will have an increasingly firm empirical foundation for developing theory" (Birdsong, 2005, p. 325). Moreover, Granger (2015) holds the view that L1–L2 comparisons are very powerful heuristic tools to circumscribe the differences between advanced L2 speakers and L1 speakers of a language. Therefore, from the perspective of corpus linguistics, the use of L1–L2 comparisons can make the findings more informative than if an L2 corpus is considered alone (Gablasova, Brezina, & McEnery, 2017; Granger, 2015).

In contrast to the proponents of L1–L2 comparisons, many researchers, in agreement with Bley-Vroman (1983) who warns of the comparative fallacy, have stipulated that L2 language must be described and studied in its own right, and they have cautioned against analysing L2 production in comparison with L1 controls (e.g., Jones et al., 2018; Lakshmanan & Selinker, 2001; Larsen-Freeman, 2014; Selinker, 2014). The argument of the comparative fallacy is this: if L2 performance is analysed relative to the target L1 scheme, then SLA researchers will fail to appreciate the autonomy and integrity of L2 language and make it difficult to avoid seeing L2 language "as nothing but deficient" (Larsen-Freeman, 2014, p. 217). Other impediments to L1–L2 comparisons have been elucidated by Birdsong and Gertken (2013), such as problems of limited applicability of evidence of nativelikeness and the matter of null results. The matter of null results needs some comments here. In an L1–L2 contrastive study, the finding of null results means the absence of statistical differences between L1 and L2 use. For most L1–L2 contrastive studies, if not all, the hypothesis usually is that there are differences between L1 and L2 production. Therefore, the null results not only challenge the hypothesis, but also provide an impetus for researchers to scrutinise the overall design of the study in order to provide appropriate explanations of the null finding. In this sense, as Birdsong and Gertken (2013) put it, "recognition of what can go wrong with comparing non-natives to natives has had the salutary effect of focusing attention on how to operate within a comparative

paradigm without running afoul of the shortcomings of the monolingual native control design" (p. 115).

The danger, then, is that the research will tend to evaluate L2 production against the 'norm' of L1 production rather than in its own right, whether this be the result of explicit data comparison or simple L1 speaker intuition (Gablasova, Brezina, & McEnery, 2017; Granger, 2012, 2015; Lardiere, 2013). As Granger (2015) puts it, "the whole debate surrounding the comparative fallacy is a healthy reminder that interlanguages can—and indeed should—also be studied in their own right" (p. 14). With these theoretical considerations in mind, in what follows I will consider spoken L1 corpora in L1–L2 comparisons from the perspective of corpus linguistics.

### 2.3.2 Concerns about Spoken L1 Corpora in L1–L2 Contrastive Studies

In L1–L2 contrastive corpus studies, an important issue to concern the analyst is whether an L1 corpus can be a representative sample of the population of interest. That is to say, whether the investigation of an L1 corpus can stand as a proxy for the study of some entire language or variety of a language (Leech, 2007). As Leech (2007) rightly notes, "without representativeness, whatever is found to be true of a corpus, is simply true of that corpus—and cannot be extended to anything else" (p. 135). The importance of the representativeness of an L1 corpus lies in the fact that, in L1–L2 contrastive studies, the representativeness of the L1 corpus limits the kinds of research questions that can be addressed and the generalizability of the results of the research (Biber & Jones, 2008). Moreover, L1 representativeness also has direct implications for corpus comparability. Leech (2007) states that the requirement of comparability depends at least partly on that of representativity: "comparable corpora[23] permit precise comparisons between two varieties or states of a language, but only if the corpora are reasonably representative of their respective varieties" (p. 142). Nonetheless, representativeness and comparability are ultimately

[23] Leech (2007, p.141) defines comparable corpora as a set of two or more corpora whose design differs, as far as possible, in terms of only one parameter: the temporal or regional provenance of the textual universe from which the corpus is sampled. According to this view, the Brown and the LOB are comparable corpora. However, McEnery and Xiao (2007) argue that corpora containing components of varieties of the same language are not comparable corpora (e.g., the Brown and the LOB), because all corpora as a source for linguistic research have always been pre-eminently suited for comparative studies. In McEnery and Xiao (2007, p.20), a comparable corpus is a corpus containing components that are collected using the same sampling frame and similar balance and representativeness, e.g., the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period. Following Leech (2007), this thesis treats corpora that are designed for comparing varieties as comparable corpora.

incompatible ways of looking at corpus design: an attempt to achieve greater comparability of two corpora may impede the representativity of each corpus, and vice versa (Leech, 2007). To put it differently, the comparability of two corpora may be achieved at the expense of the representativeness of each corpus. Therefore, when building comparable L1 and L2 corpora, it is of paramount importance to balance the representativity and comparability of the corpora.

To date, in corpus design and compilation, the above concern pertinent to comparable L1 and L2 corpora has not yet been investigated in-depth. It can, however, be argued that corpora that are designed for comparing L1 and L2 language, such as the LINDSEI and its L1 counterpart–the LOCNEC, appear to prioritise corpus comparability over representativeness. In other words, ensuring comparability with the L2 corpus is a major consideration when building the L1 corpus. This is exactly the case of the LINDSEI and the LOCNEC (De Cock, 2004), which were designed with the same criteria, e.g., the L1 participants were required to conduct the same tasks as the L2 participants. Admittedly, this kind of design guarantees that the two corpora are comparable in the sense the data were gathered in comparable situations. However, Gries and Deshors (2014) argue that "most previous research has so far adopted a very lax interpretation of 'in a comparable situation'—namely, the interpretation that the corpora are comparable because the NSs and NNSs were in a similar language-production setting" (p. 113). It is problematic to justify the comparability on the basis of the single factor that the L1 and L2 data were produced in the same situation, because this method ignores other features that may affect the use of a language pattern, such as the syntactic structures the pattern occurs (Gries & Deshors, 2014). From this perspective, Gries and Deshors (2014) then contend that a comparable situation needs to be defined in a more fine-grained and comprehensive way by taking into consideration both the linguistic/contextual features and the traditional decontextualised features. The focus of Gries and Deshors (2014) is to introduce more appropriate statistical methods to investigate variation between corpus data, but their comments on comparable situation provide impetus to reconsider the assessment of corpus comparability of L1 and L2 corpora.

Aside from the dilemma in terms of representativeness and comparability, another issue is the possibility of achieving comparability. When comparing a spoken L1 and L2 corpus, it is understandable that they are expected to differ in terms of only one variable (i.e., being native vs non-native speakers), but be similar in other respects (e.g., age, topics) (Gablasova, Brezina, & McEnery, 2017; Leech, 2007). In these circumstances, an observed contrast in linguistic frequency between the two corpora is likely to be due to the variability between the two corpora

rather than variability within one corpus or within the other (Leech, 2007). However, this seemingly straightforward assumption is problematic in corpus design and compilation practice. For instance, the original example of comparability is that of the Brown Corpus and the LOB: the LOB was intended to match the Brown Corpus in all respects apart from the country of origin (the UK versus the USA); however, the two written corpora are not identical in composition due to practical constraints, e.g., the Western and Adventure Fiction category (N) in the Brown Corpus contained many more Western Fiction texts than the LOB (Leech, 2007). In practice, it may be even harder to build or find two spoken corpora that are identical in all respects apart from only the target variable under examination. Still in the case of the LINDSEI and the LOCNEC, De Cock (2004) claims that there are differences between them other than the native/non-native distinction, such as age of speakers and the degree of interactivity between interlocutors (also see Gablasova, Brezina, & McEnery, 2017). Accordingly, in L1–L2 contrastive studies, it is not necessarily the case that the same design criteria can guarantee an L1 and L2 corpus only differ in native/non-native distinction. Therefore, when using an L1 and L2 corpus which are claimed to be comparable, it is good practice to examine whether there are considerable differences between the corpora. By doing this, researchers are able to claim that the findings reflect the realities of language use rather than are caused by the artefact of the corpus design or the data collection methods. The issue of corpus comparability will be discussed in some detail in 2.4.1.

## 2.4 Conducting L1–L2 Contrastive Studies to Investigate L2 Production

The findings of L1–L2 comparative studies are subject to corpus comparability (Callies, 2015), so ignoring comparability between L1 and L2 corpora may obviate more valuable insights into the characteristics of L2 production. As has been noted previously, evaluating to what extent an L1 and L2 corpus used in a study are comparable is a complex issue which cannot be reduced to how similar/different the language production situation is (Gablasova, Brezina, & McEnery, 2017). However, it seems that paying attention to similarities/differences in corpus design and creation between comparable L1 and L2 corpora can be a good starting point. After this, the second step that can be taken is to investigate variation within each corpus. These steps then can lead researchers to develop or adopt appropriate statistical methods to study L2 production in a more comprehensive and scientific way. Finally, when interpreting the results, there are two crucial issues which should be illuminated carefully: the stereotypical views toward L1

speakers as expert users and L2 speakers as deficient communicators. In this section I will consider these issues in turn.

### 2.4.1 Comparing L1 and L2 Corpora

In much research comparing L2 use to L1 use, the effect of being native or non-native speakers of a language is typically treated as the vital variable which can substantially affect L2 speakers' use. However, it has been noted that variables, such as gender, age and speaker roles, have influences on both L1 and L2 production (e.g. Baker, 2014; Brezina & Meyerhoff, 2014; Fuller, 2003; Gass & Varonis, 1986; Müller, 2005; Schweinberger, 2018). Accordingly, if the spoken L1 and L2 corpus involved in a contrastive study differ in the distribution of gender groups, for example, the L2 corpus is imbalanced in favour of male speakers while the L1 corpus is relatively gender-balanced, it would be a disservice to the inferences in relation to the differences between L1 and L2 production if the differences in gender distributions were not taken into consideration. So, in such L1–L2 contrastive studies, observed characteristics of L2 production are not necessarily due to their status of being non-native speakers, but can in fact result from differences in corpus design and compilation between L1 and L2 corpora on which the research is based (Ädel, 2008). Moreover, as has been noted in 2.3.2 above, even with the same design criteria, the comparable L1 and L2 corpus can differ considerably in some respects other than the native/non-native distinction. To facilitate the validity and reliability of the corpus results of L1–L2 contrastive studies, researchers have responsibilities to be familiar with the employed L1 and L2 corpus and be aware of any factors that may affect the corpus results. This being the case, it is necessary to explore and manifest the divergences between L1 and L2 corpora that are used in L1–L2 contrastive studies. However, comparing L1 and L2 corpora has long been disregarded in previous studies using the L1–L2 comparative approach. With this background, Chapter 5 to some extent has taken the first step forward to investigate the differences between the spoken L1 and L2 corpus built in this thesis which stem from corpus design and practical compromises, seeking to encourage researchers to take into consideration the differences between corpora when interpreting the observed differences between L1 and L2 production.

In L1–L2 contrastive research, researchers typically analyse aggregate data to abstract away from individual speakers to identify specific language patterns often defined on the basis of the L1. Inter-speaker variability has been noticed, but in a substantial part of L2 corpus

studies to date such individual differences still go unnoticed or tend to be disregarded (Callies, 2015). For example, Aijmer (2011) examines the differences/similarities between the L1 and L2 speakers in the usage of *well* by using data from the Swedish component of the LINDSEI and its native counterpart—the LOCNEC. In this study, Aijmer (2011) compares the two corpora and concludes that Swedish learners overuse *well* in comparison to L1 speakers of English. This comparison which focuses largely on group differences is the so-called aggregate data methodology (Brezina, 2018; Brezina & Meyerhoff, 2014). It should be mentioned that aggregating data is a normal procedure in every corpus design and comparative research, which enables researchers to have general understandings of linguistic phenomena; on the other hand, this approach highlights the differences between the L1 and L2 groups but mostly ignores the differences within each group (Brezina & Meyerhoff, 2014; Gablasova, Brezina, & McEnery, 2017; Gablasova, Brezina, McEnery, et al., 2017). In L1–L2 contrastive studies, it is arguably necessary to aggregate data to reveal the differences between L1 and L2 production, but the study of inter-group differences should be treated as the starting point for our exploration of the characteristics of L2 production, further investigations of intra-group variations also need to be carried out to deepen the understanding of L2 acquisition. The issue of within group variation will be scrutinized in Chapters 5 and 7.

In L1–L2 contrastive studies, it should be noted that different statistical methods can lead to somewhat divergent results, and divergent results adduce different interpretations and conclusions. Therefore, one crucial issue that researchers using corpus data must grapple with is to choose the best statistical method for a specific analysis. In this thesis, the statistical methods used will be demonstrated in Chapter 6 when instigating the L2 use of the marker 就是 *jiushi*.

### 2.4.2 Interpretation: Overuse/Underuse in L2 Production

At the end of this section, my discussion turns to interpretations of the corpus results in L1–L2 contrastive studies. Gries and Deshors (2014) rightly point out that many of the traditional L1–L2 contrastive studies base their analyses of differences between L1 and L2 production on the different frequencies with which the phenomenon in question occurs in the L1 and L2 corpora. This method leads researchers to interpret that L2 speakers overuse or underuse the patterns under examination (e.g., Aijmer, 2011; Buysse, 2012, 2015; Müller, 2005). Two decades ago,

Leech (1998) warns against the danger of using the terms 'overuse' and 'underuse' in a judgemental spirit in L2 corpus research, insisting that they should be interpreted descriptively. Likewise, Gilquin and Paquot (2008) claim that the two terms "are descriptive, not prescriptive, terms: they merely refer to the fact that a linguistic form is found significantly more or less in the learner corpus than in the reference corpus" (p. 58). A different perspective on the terms is offered by Aston (2008) who holds the view that the terms 'overuse' and 'underuse' should be avoided, for the terms indicate that "the learner should at all times attempt to conform to native-speaker norms" (p. 343). Clearly, Aston's criticism is not merely technical but also takes into consideration the limits of using L1 models. In L1–L2 studies, L2 speakers are typically treated as deficient communicators in most L2 corpus studies and interpretations of L2 speakers' choices of language use tend to be limited to that of insufficient proficiency, while adult L1 speakers of the target language "are often thought to process language in ways that are efficient and accurate, and with little inter-individual variation" (Birdsong & Gertken, 2013, p. 122). As such, L1 use is often seen as a benchmark which enables researchers to analyse whether the L2 speakers approach frequency and use that are typical of the target language in a social community (Magliacane & Howard, 2019, p. 5). Nevertheless, McCarthy and Carter (2001) claim that L1 speakers are not necessarily expert users of the target language. Similarly, when talking about the significance of the concept of the native speaker for English language testing and teaching, Davies (2011) argues that the idea that "all these native speakers are at C2, the highest level on the Council of Europe Reference Scale, makes no sense. Some perhaps are, but they are unusual" (p. 306). These claims indicate that L1 speakers are a heterogenous group, and it is not necessarily the case that any L1 speakers can provide models that L2 speakers would want to imitate. As a result, caution about the mindset that elevates idealised L1 speakers over stereotypicalized L2 speakers who are seen as deficient communicators is required.

In addition, SLA researchers have provided a revealing view of L2 speakers which can affect how we interpret L2 performance in corpus studies: an L2 speaker as a bilingual is not two languages in one person, rather, "he or she has a unique and specific linguistic configuration" (Grosjean, 1989, p. 6). Thus, because of the coexistence and interpenetration of the two languages in a bilingual, neither the L1 nor the L2 of a bilingual can be expected to be identical in all respects to the language of a monolingual (Birdsong, 2005; Birdsong & Gertken, 2013). Based on this view, when comparing L1 and L2 performance, it is not appropriate to limit interpretation of the use of L2 speakers to that of insufficient proficiency, rather, it may be worthwhile to consider the effects of the two language systems on L2 production. Moreover,

Gries and Deshors (2014) from a methodological perspective argue that "over-/underuse counts per se do not even speak to the issue because any over-/underuse by a learner may be due to the learner being more/less often in linguistic/contextual situations requiring the supposedly over-/underused choice" (p. 113).

## 2.5 Summary

This chapter has discussed aspects with respect to spoken L2 corpus design and considered issues to which attention should be paid when using spoken L2 corpora in L1–L2 contrastive studies. In the first part, this chapter has discussed theoretical and methodological concerns in terms of spoken L2 corpus design by reviewing four documented corpus resources of spoken L2 English and Chinese. Discussions on corpus criteria in terms of L2 speakers and L2 data collection will be used as guidelines for the compilation of the spoken Chinese corpus. The second part has turned to consider the widely used L1–L2 comparative approach. In this part, I have shown the debates over this approach and clarified the requirements for the comparable L1 corpora involved in L1–L2 contrastive studies. Although features of L1 corpus design and compilation have not been considered in this part, discussion of the representativeness and comparability of L1 corpora will facilitate the process of the development of the spoken L1 corpus. At the end of this chapter, I have also considered some issues involved in L1–L2 contrastive studies. In the following chapter, I will give a detailed description of the design of the spoken Chinese corpus.

# Chapter 3 Corpus Design and Data Collection

## 3.1 Introduction

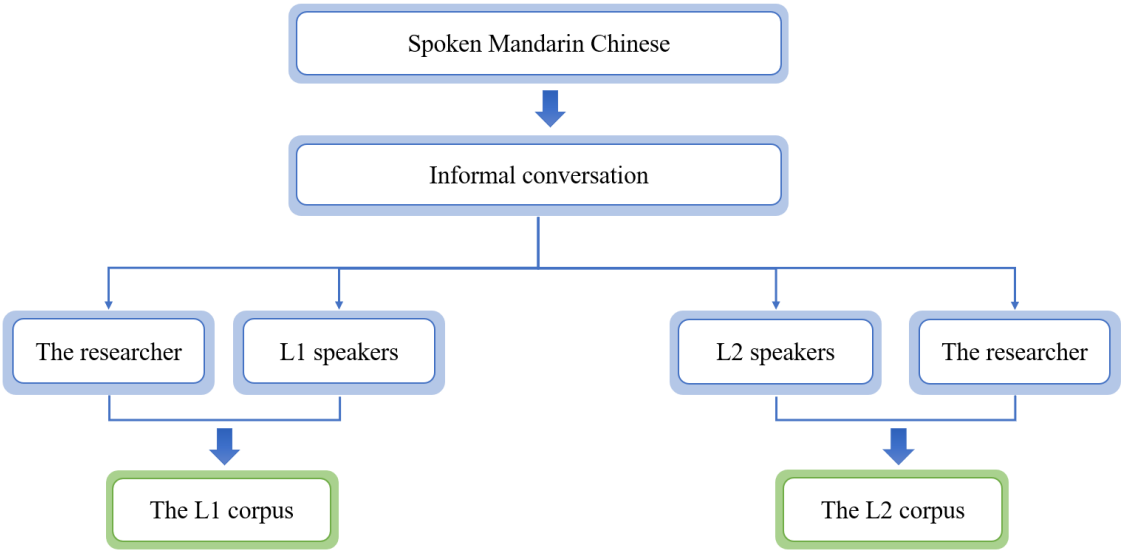
The preceding chapter has reviewed some key considerations on spoken corpus design. In this chapter, the focus is primarily on the methodological decisions taken in the spoken Chinese corpus design and data collection. To begin, in Section 3.2 I outline the overall design of the spoken Chinese corpus. I then clarify the notion of spoken Chinese and explain the L1 speakers of Chinese who are one of the target population in this study. Particular attention is also paid to the target L2 speakers of Chinese, and the size this corpus attempted to achieve. In this section I explicate the reasons for employing the unstructured interviewing method as well. Following this is an explanation of ethical considerations with respect to spoken data collection in Section 3.3. Then I move on to discuss aspects of participant recruitment and the procedure of conducting interviews in terms of the L1 and L2 group respectively in Sections 3.4 and 3.5. After this, Section 3.6 considers the collected audio recordings. This chapter finally closes with a brief summary in Section 3.7.

## 3.2 Spoken Corpus Design

In this section I primarily demonstrate my consideration of the overall design of the spoken Chinese corpus which consists of an L1 corpus and an L2 corpus (see Figure 2 overleaf). The spoken L2 corpus was designed to sample L1–L2 informal interaction between the researcher and L2 speakers of Chinese. To ensure comparability, its counterpart, the L1 corpus, contained L1–L1 informal interaction between the researcher and L1 speakers of Chinese. I have discussed the dilemma of corpus representativeness and comparability in Chapter 2. Bearing this in mind, in this section I will show how the representativeness and comparability of the two corpora were considered in corpus design by discussing the target language and speakers, and the sampling size associated with the data collection method.

**Figure 2**

*The Overall Design of the Spoken Chinese Corpus*




### 3.2.1 Spoken Chinese and the Target L1 Participants

Chinese is the most widely used language in China (see Figure 3). According to Huang and Liao (2017), there are seven major dialect groups of modern Chinese: Northern dialects, Min, Kejia (Hakka), Wu, Xiang, Yue (Cantonese), and Gan (see Figure 4). The use of a commonly spoken language has been promoted since the establish of the People's Republic of China in 1949, which established *Putonghua* (common language), as the official language of education widely learned and spoken over a large geographical area (for details, see Lin, 1998; D. Wang, 2016; Wang, 1999). *Putonghua* is also referred to as 'Mandarin', 'Mandarin Chinese', 'Modern Chinese' or 'Standard Chinese' in previous studies written in English (for further discussions, see Ding et al., 2000; Lu, 2009; Xu & Fang, 2019). It has used the pronunciation of Beijing dialect as its standard pronunciation, the words of Northern dialects as its basic vocabulary, and the modal writing of the modern vernacular prose as the norm for grammar (Huang & Liao, 2017, p. 1). *Putonghua* can be written in *Pinyin* (a Romanised writing system) and simplified Chinese characters (examples can be found in Chapter 4).

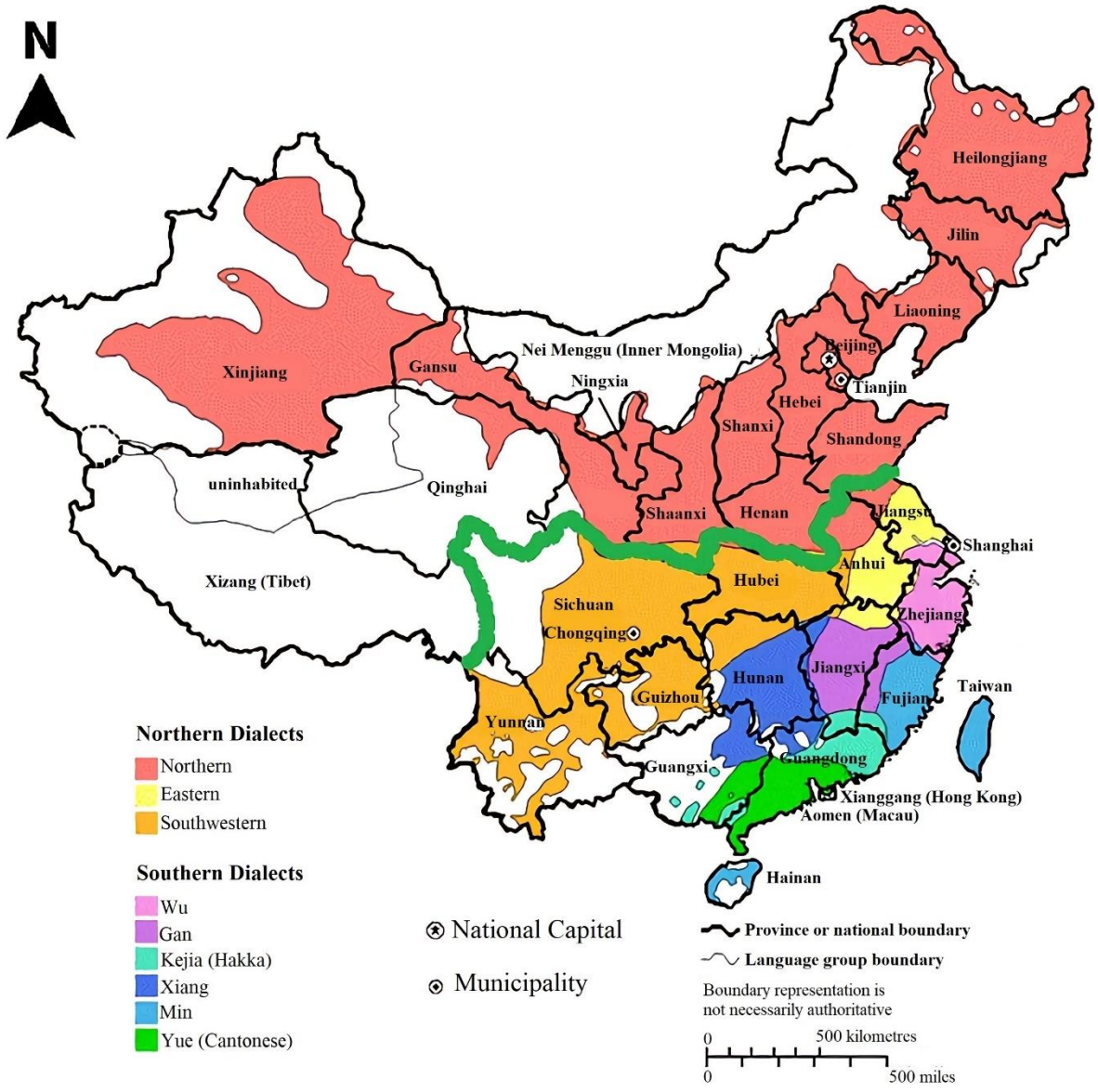
**Figure 3**

*Chinese and Non-Chinese Language Groups in China*


*Note.* There are 56 official ethnic groups in China. According to the national statistics of the 2010 Population Census (http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm), the Han Chinese account for 91.51% of the population in mainland China. The language of Han people is Chinese, which is shared by the Hui minority.

**Figure 4**

Seven Major Dialect Groups of Chinese[24]




It should be noted that *Putonghua* has diverged from Mandarin Chinese used in Taiwan which is called *Guoyu* (national language), due to Taiwan being separated politically from mainland China for decades (differences between *Putonghua* and Taiwan Mandarin are

[24] The green line (which is added by the author) is a simplification of the Qing Mountains–Huai River line which is a geographical and climatic dividing line between northern China and southern China. It should be noted that 'Northern dialects' is defined from a linguistic perspective which does not necessarily relate to the regional split. Figure 3 and Figure 4 were originally retrieved from http://www.dartmouth.edu/~chinese/maps/maps.html. Since this study follows the definition of modern Chinese provided by Huang and Liao (2017), the original term 'Mandarin' used in Figure 4, was replaced by the revised term 'Northern Dialects'.

discussed from various perpectives in Deng et al., 2006; Dong & Huang, 2020; Fang, 2013; Tseng, 2004). However, some researchers claim that there is growing convergence between *Putonghua* and Taiwan Mandarin in vocabulary (e.g., Dang et al., 2017; Diao, 2015), as the interaction between the mainland and Taiwan becomes more frequent. With this background in mind, it seems appropriate to consider *Putonghua*, and Taiwan Mandarin separately. Therefore, the spoken Chinese discussed in this thesis refers to *Putonghua* used in the mainland, and all L1 and L2 participants were selected on the basis that they were speakers or learners of *Putonghua* rather than Chinese dialects and Taiwan Mandarin.

To delimit the boundary of the target L1 population, there are two issues that should be clarified in terms of the selection criteria of native Chinese speakers. Figure 3 shows that China is a multilingual country, but due to *the Law of the People's Republic of China on the Standard Spoken and Written Chinese Language*[25] effective in 2001, *Putonghua* is used as a national lingua franca among all language groups in the mainland. However, in this study, people of ethnic minorities whose native language is not Chinese but who can speak *Putonghua* were not considered eligible for inclusion in order to control the variables that may affect language use. Therefore, native Chinese speakers in this study were limited to the Han people whose native language is Chinese. It should be noted that the language situation of the Han population can be broadly characterised as diglossia with dialect bilingualism (Li, 2006): although *Putonghua* is the official language in the mainland, both Chinese dialects and *Putonghua* are used in daily life. Given that *Putonghua* was developed from selected linguistic features of Chinese dialects, in this study I prioritised collecting Chinese spoken by educated adult native Chinese speakers who came from the north of China (see 3.4.1 for more information).

In addition, with reference to the notion of native speakers, monolingual native speakers have been preferred by many researchers in SLA research (e.g., Birdsong & Gertken, 2013; Mack, 1997; Ortega, 2012). According to Mack (1997), a monolingual native speaker refers to "an individual who has been exposed to a specific language from infancy and who can function effectively in ONLY one language" (p. 115). It is recognised that in a world that is increasingly globalised, a growing number of people have some knowledge of more than one language. In China, English has become one of the key strands of Chinese education, and numerous kindergarten schools have already taught English to Chinese toddlers (for detailed information,

[25] Law of the People's Republic of China on the Standard Spoken and Written Chinese Language: http://en.moe.gov.cn/Resources/Laws_and_Policies/201506/t20150626_191388.html.

see Bolton & Graddol, 2012). Also, the Chinese government has imposed a policy of having English in the National Curriculum since year 3 of primary school in 2003, with some areas starting from even Primary One and Two (Qi, 2016). However, it is not assumed that an individual is bilingual if he or she possesses even a small amount of knowledge of a second or foreign language (Mack, 1997). Given the fact that a great many Chinese graduates who have had many years of English education at university rarely use English in their daily life, it is appropriate for the purpose of this study to regard them as monolingual native speakers of Chinese. It is beyond the scope of this research to engage in the debate over whether monolingual or bilingual native speakers can be used in L2 studies. Following the definition of native speakers suggested by Hyltenstama and Abrahamsson (2012), native Chinese speakers in this study are defined as Chinese citizens who have been exposed to Chinese from infancy and who can function effectively in Chinese and continue to use it regularly throughout the life span, no matter whether they are seen as monolingual or bilingual.

### 3.2.2 The Target L2 Participants

At the very beginning of this corpus design, it was proposed that L2 speakers of Chinese could range from active learners to non-native speakers who used Chinese in their professional life or for leisure. However, many issues emerged in the procedure of corpus design, such as L2 participants' nationalities and the number of participants that needed. To investigate the possibility of recruiting enough participants and of collecting informal speech for the project, a small pilot survey was carried out[26]. In this small pilot survey, four L2 participants (one female, three males), who came from four countries: Thailand, UK, Russia, and Congo, were recruited prior to the data collection of the current project. The L2 participants and I were not previously known to one another. They had studied and lived in China for at least two years, but none of them had taken any Chinese proficiency tests. Permission was obtained from all participants and they were aware that their speech was being recorded. Various topics were provided for each participant. Each recording of L2 speech was approximately 20 minutes long, and the average yield per minute of recordings was around 170 words[27]. One participant, who

[26] In this small pilot study, five L1 participants (three females and two males) who were my friends and acquaintances were also recruited. Each L1 participant had a 30-minute long conversation with me and on average they spoke at least 200 words per minute.
[27] In this thesis, when talking about Chinese, the term 'words' means Chinese characters.

was studying in New Zealand at the time of interviewing, had very limited vocabulary in Chinese. As a result, the whole conversation was conducted unavoidably in English.

According to the L2 conversations gathered in the small pilot survey, it was advisable to confine the research focus to L2 speakers of Chinese who were at intermediate to advanced proficiency level, considering the likelihood of obtaining enough material from them. However, as Gass and Selinker (2008) put it, there is no absolute accepted cut-off point for beginner, intermediate, and advanced: one researcher's advanced category may correspond to another's intermediate category. In this study, the qualification of HSK test and the number of years of learning Chinese were considered as two important factors for identifying their proficiency level theoretically (the problems of using these measures to assign L2 proficiency will be given in 3.5.1). Additionally, as has been discussed in Chapter 2, the language backgrounds of L2 participants could have an impact on their production. To reduce the complexity of language production, another key decision made in the L2 corpus design was to focus exclusively on speech produced by native English speakers. Since I was studying in New Zealand, which is an English-speaking country, it ought not be difficult to recruit enough New Zealanders who were native English speakers to take part in this project. Therefore, New Zealanders who were at the intermediate to advanced Chinese proficiency levels were the target recruits in this study. However, New Zealanders of Chinese ethnicity who were characterised as heritage speakers of Chinese were not considered eligible for inclusion in the present study. This decision was taken because "the nature of language learning for heritage language learners differs from language learning involving non-heritage language learners" (Gass & Selinker, 2008, p. 24). Accordingly, L2 speakers of Chinese were restricted to New Zealanders of non-Chinese ethnicity to ensure the uniformity of the language background for the present study. What should be mentioned is that I did not attempt to completely adhere to this predetermined sampling frame, as it was uncertain whether enough ideal New Zealand participants could be recruited or not. Thus, the above criteria were extended to include any Australians of non-Chinese ethnicity who were at the intermediate to advanced Chinese proficiency levels, since Australia is also an English-speaking country and shares many similar linguistic characteristics with New Zealand.

The criteria for participant recruitment are given in Table 4. A key decision made early on in the creation of the corpus was to gather speech contributed by adult L2 speakers of Chinese: people whose ages range from 18 to 50 years old. Moreover, the male/female binary division was used to gather gender information. In terms of gender, the primary intention was,

as far as possible, to recruit equal numbers of women and men in order to make the corpus as representative and balanced as possible. Likewise, L1 recruits' ages should range from 18-50 years old as well to ensure that the populations represented in the two corpora were comparable. Additionally, it was decided to collect metadata categories such as occupation, educational level, and the places where both the L1 and L2 participants studied or worked[28].

**Table 4**

*Criteria of Participant Recruitment*

| Criteria | L2 participants | L1 participants |
|---|---|---|
| **Selection criteria** | age<br>gender<br>Chinese proficiency level<br>L1 background | age<br>gender<br>region |
| **Descriptive criteria** | educational level<br>occupation<br>cities have lived in in China<br>current residential city<br>relationship with the researcher | educational level<br>occupation<br>current residential city<br>relationship with the researcher |

In this pilot study, a built-in recorder on my laptop was used to record all the interviews, which proved later that the recordings were of sufficient quality for transcription. The decision of selecting the built-in recorder on my laptop as the recording equipment was made due to the following reasons. First of all, this recorder made sure that all data would be recorded directly in digital format and I could transcribe all the recordings directly on my computer. Secondly, given the quality of the recordings gathered in the small pilot survey, I was confident that this approach to spoken data collection would be successful. Therefore, I decided to use this built-in recorder to make audio recordings. Further discussions regarding this recording approach can be found in 3.4.2.

To sum up, it confirmed that once participants were recruited they were able to carry out the task successfully. This small pilot study was a preparation for the corpus project, consequently, the recordings gathered in this pilot study were excluded from the construction of the spoken Chinese corpus.

[28] China is a huge country geographically, which results in the noticeable existence and use in everyday life of various Chinese dialects. Consequently, it is inevitable that dialects impact on the usages of spoken Chinese in daily communication.

### 3.2.3 Corpus and Sample Size

In the initial proposal, the spoken Chinese corpus was designed to gather 400,000 words of informal conversational interaction: 200,000 words of L1–L1 interaction for the L1 corpus and 200,000 words of L1–L2 interaction for the L2 corpus. As has been discussed in Chapter 2, a spoken corpus of 200,000 words was reasonably small in comparison to some large-scale spoken corpora in existence, such as the Spoken BNC2014 and the TLC. Although a corpus, no matter how large it is, cannot capture all patterns of the language being investigated nor represent them in precisely "the correct proportions" (Sinclair, 2005, p. 3), there was no doubt that such a large-scale spoken corpus as the Spoken BNC 2014 and the TLC would give greater statistical validity. It is easy to see in recent times the considerable advantage in putting quantity first in spoken language collection, as there is plenty of spoken language around (Sinclair, 2014). In addition, spoken corpus building has benefited greatly from the development of technology. Various available software, such as video/audio recorders and transcription tools, have been widely adopted by corpus compilers to make corpus creation less labour-intensive. However, the optimum size of a corpus depends largely on the research questions being addressed and the practicalities (e.g. available time and costs). Additionally, the intrinsic complexity of spoken language keeps the building of a spoken corpus a time-consuming and laborious task. Given that I was not attempting to achieve quantity at the cost of quality, 200,000 words for each subcorpus were considered to be adequate in the present study. Smaller corpora, such as the spoken Chinese corpus built in this study, can be very useful to the research community, particularly in light of the lack of publicly accessible spoken Chinese corpora. In contrast to the claim of Sinclair (2004) that being small in a corpus is "simply a limitation" (p. 189), some researchers argue that small corpora have advantages in research (Carter & McCarthy, 1995; Ghadessy et al., 2001; Koester, 2010; McCarthy & Carter, 2001). As Koester (2010) puts it, a small corpus allows a much closer link between the corpus and the contexts in which the linguistic patterns under examination in the corpus were produced. For instance, the spoken Chinese corpus created in this study, enabled me to examine all the occurrences of 就是 *jiushi* in context (see Chapter 6). Acknowledging the benefits of small corpora, the spoken Chinese corpus can be more valuable if it will continue to grow beyond the present project.

To assemble enough material for the spoken Chinese corpus, there were three factors in terms of the sample size which had to be taken into consideration: the number of speakers, the length of conversations, and the number of conversations. Based on the small pilot study in

3.2.2, assuming that the number of words spoken by an adult L1 speaker of Chinese per minute on average was approximately 200 words, the corpus then would need 12 L1 participants, if each could participate in three 30-minute conversations. Given that L2 speakers of Chinese might not speak Chinese as fluently as native speakers, a 15- to 20-minute conversation was appropriate. Considering the possibility and difficulty of recruiting enough L2 participants, all L2 participants would be asked to contribute three conversations. As such, the corpus would need about 30 L2 participants if an adult L2 participant could speak 150 words per minute (see Table 5).

**Table 5**

*The Initial Targets of Data Collection*

| The spoken Chinese corpus | Size (words) | No. of participants | Length of each interview (minutes) | No. of interviews per person |
|---|---|---|---|---|
| The L1 corpus | 200,000 | 12 | 30 | 3 |
| The L2 corpus | 200,000 | 30 | 15-20 | 3 |

Assembling equally sized samples is theoretically an efficient way to achieve balance and representativeness (Baker, 2006), while this goal was not necessarily to be achieved by strictly controlling the lengths of conversations. In practice, the expected number of words included in a conversation was not in line with a fixed length of conversation, as the speaking rates may vary. It seemed to me that the length of each conversation could be rather flexible and follow the natural course of the interaction. I therefore decided to make no effort to control the length of conversations in this study.

### 3.2.4 The Unstructured Interviewing Method

Naturalness is highly valued in assembling unplanned, spontaneous spoken data (Chafe et al., 2013; Crystal & Davy, 1975; McCarthy, 1998). One common practice to gather spontaneous conversations in spoken corpus building is to recruit contributors and then ask them to record their everyday conversations without researchers getting involved (e.g., Crystal & Davy, 1975; Love et al., 2017). This method enables the data to be gathered as naturally as possible. In this study, I employed an unstructured interviewing method to collect spoken data. As has been mentioned in Chapter 1, this method was selected as the means of data collection due to some practical restraints (e.g., time and costs), as well as my consideration of facilitating the

comparability of the two groups. The unstructured interview, according to Dörnyei (2007), allows maximum flexibility to follow the interviewee in unpredictable directions, with only minimal interference from the research agenda. It attempts to create a relaxed atmosphere in which the respondent may reveal more than he or she would in informal contexts, with the interviewer assuming a listening role. By using this method, it was hoped to mitigate the influence of any observers' involvement as well as to maximise the overall quantity of data in the present study.

This study was designed so that all interviews were conducted between me (as the interviewer) and one participant (as the interviewee). To conduct the unstructured interview, neither interview guides nor specific topics were necessarily provided in advance, although I prepared a few opening questions and optional topics, such as travelling, language learning, culture and so on. Secondly, during the interview, I as the interviewer felt free to ask occasional questions for clarification or to give reinforcement feedback to keep the interview moving, but interjections were kept to a minimum. To eliminate the pressure as far as possible and ensure enough L2 participants could take part in the project, it was planned to conduct interviews at participants' convenience, improving the rapport between the L2 participants and me. Furthermore, all participants had various options in terms of channels for conducting interviews open to them: for example, online interviews via Skype or WeChat (the most widely used social application in China), telephone interviews or face-to-face interviews. Every participant would be informed before I started to record the conversation. Crowdy (1993) asserts that when participants are aware they are being recorded, an initial period of unease or unnaturalness occurs but vanishes quickly. So, the potential influence of the presence of the recorder on language output could be largely ignored.

## 3.3 Ethical Considerations

Before collecting data through audio recording, it is essential that appropriate ethical procedures are addressed (Adolphs & Knight, 2010; Thompson, 2010; Weisser, 2016). To represent spoken language at its most natural, some well-known spoken corpora, such as the LLC (Svartvik, 1990) and the BNC (Crowdy, 1993) recorded speech surreptitiously, although permission to use the material was obtained from speakers afterwards. It has become common practice in recent times to obtain permission from participants prior to the recording in spoken corpus building. Accordingly, it is important to thoroughly consider all relevant ethical issues

and fully discuss them with participants before collecting spoken data. If recordings are to be made public in any form, informed consent should be obtained from participants prior to the recording to ensure that they understand the nature of the recording and the format of distribution and access. In addition, participants' identities should be protected. Names, along with other identifiers, can be modified to avoid identification. If the recordings are not made publicly accessible, they should be stored on a password-protected computer with access limited to the research team.

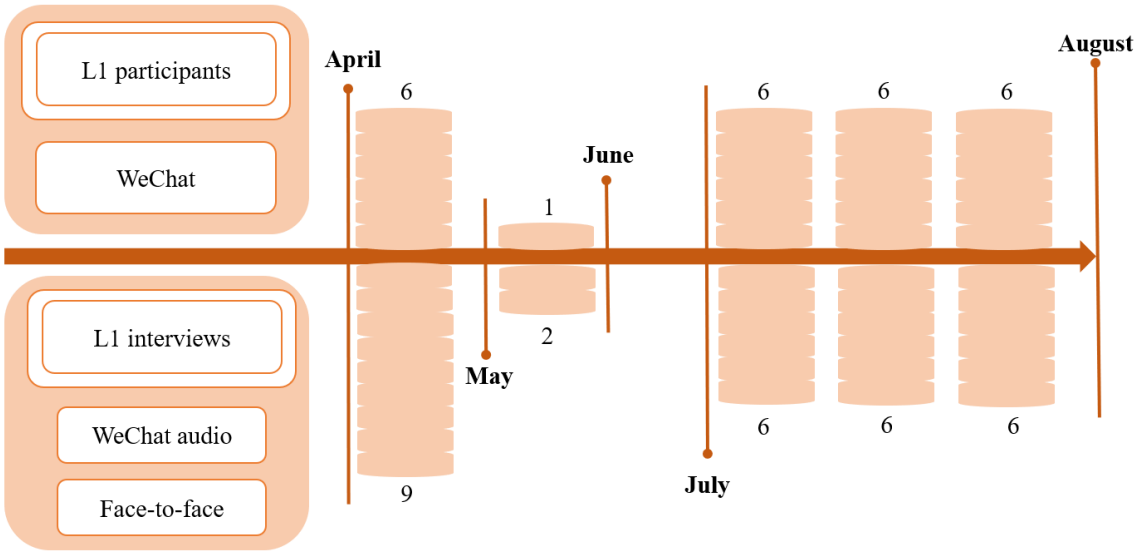
My Human Ethics Application was approved by Massey University Human Ethics Committee (Ethics Notification: 4000018814) in 2018 before I started collecting my data (see Appendix B). I considered all relevant ethical issues and fully discussed them with participants before collecting spoken data for the spoken Chinese corpus creation. To ensure participants had a good understanding of what participation involved, each was provided with an Information Sheet, which contained detailed information of my research. As two groups were involved in the present study, I prepared two Information Sheets with slight differences for L1 participants and L2 participants (see Appendices C and D). The two Information Sheets showed participants all the information related to their decision to participate. A copy of the Participant Consent Form (see Appendix E) was then provided which was signed by the participants and returned to me. All participants voluntarily took part in my study and they had the right to withdraw from the study at any time if they did not want to continue. Participants were aware that all conversations would be recorded and all the transcripts would be publicly available. However, all references to full names and places or other identifiable information would be replaced with pseudonyms (see Chapter 4 for further information), and recordings would not be made publicly available, as audio data exists as an "audio fingerprint" (Adolphs & Knight, 2010, p. 43), which makes it relatively easy to identity participants involved. All audio recordings would be stored on a password-protected computer with access limited to the researcher.

## 3.4 The L1 Group: Participant Recruitment and Data Collection

As Figure 5 shows, the vast majority of the L1 participants in this study were invited to attend the interviews between April and July in 2018. In this section I will give a detailed account of the L1 participant recruitment and the procedure of conducting interviews.

**Figure 5**

*Timeline of L1 Participant Recruitment and Interview Conducting*




*Note*. The upper side indicates the number of L1 participants recruited from April to August in 2018; the lower part represents the number of L1 interviews conducted via WeChat or face-to-face during that period of time in 2018.

### 3.4.1 The L1 Participants

For the recruitment of the L1 participants, a total of 25 L1 speakers of Chinese were recruited via WeChat. As Table 6 shows, there was a conspicuous region imbalance between the number of northerners and that of southerners, for I prioritized collecting Chinese spoken by the northerners (see 3.2.1). For practical reasons, the sampling frame for the L1 participants were conditionally flexible: people who were born in southern China but have lived in northern cities for at least three years would also be included in the project. Specifically, all the L1 participants were Han Chinese and had lived in Beijing for at least three years at the time of interviewing. Beijing is a *Putonghua*-speaking city; therefore, although this study covers L1 speakers of Chinese with different dialect backgrounds[29] (see Appendix F), the L1 participants shared a similar language environment. However, it should be pointed out that not all L1 participants

[29] Although L1 speakers of Chinese who come from different provinces can be broadly characterised as the Northern Dialect speakers (including the researcher), they actually speak different local dialects (see Figure 4).

were in China at the time of interviewing. There are five participants who took part in my research while studying overseas in 2018 (for detailed information see Appendix F).

**Table 6**

*The Demographic Features of the L1 Participants*

| Selection criteria | Demographic group | No. of speakers |
|---|---|---|
| Region | northern China | 18 |
| | southern China | 7 |
| Age | 20-25 | 3 |
| | 26-30 | 14 |
| | 31-35 | 6 |
| | 36-40 | 2 |
| Gender | female | 15 |
| | male | 10 |
| **Descriptive criteria** | **Demographic group** | **No. of speakers** |
| Highest qualification | graduate | 2 |
| | postgraduate | 22 |
| | doctorate | 1 |
| Relationship with the researcher | friends and acquaintances | 22 |
| | strangers | 3 |

In the original proposal, I decided to invite my friends and acquaintances who were native Chinese speakers to join this project in order to gather enough data and to save time. Previous studies have argued that relationships between speakers in conversation have impacts on language use (Farr et al., 2004; Myers, 2008). However, the relationships between myself and participants were not taken into consideration when recruiting participants. In retrospect, this decision has a potential impact on the subsequent analysis of the L2 use which I will discuss in some detail in Chapter 5.

Additionally, gender information was gathered based on a binary division of the sexes. When recruiting participants, since most of the L1 participants were my friends and acquaintances, I presupposed that all of them would willingly describe their gender in this binary division, so none of the respondents were identified as non-binary. A retrospective look at this strategy of gathering gender information, it was careless and rather subjective. In future corpus projects, appropriate ways should be adopted to gather such information. Further discussion on this issue will be provided in Chapter 7.

Table 6 also shows that the L1 recruits were categorized with respect to four age groups, rather than the exact age. The age information was not self-reported by the L1 participants, rather, I made an educated guess regarding participant age. One reason was that it is to some extent a cultural taboo to ask women's age in China. Another one was that the guesses were possible since I knew most of the L1 participants. As for the three strangers, since they were introduced by my friends, it also was possible to give appropriate guesses according to the relationship between them and my friends.

To conclude, all this metadata information was recorded for users who are interested in exploring relations between language use and these metadata categories. In the following, I will give a detailed description of gathering the L1–L1 interaction for the spoken L1 corpus.

### 3.4.2 The L1 Interviews

As all L1 participants gave informed consent prior to recording, interviews were conducted under the awareness of the conditions established for the study. Participants were asked not to prepare for the interview to ensure they could talk as naturally as possible. In this study, a total of 29 interviews were conducted between the L1 participants and me. Table 7 shows the number of audio recordings produced by each demographic group. For age, the relatively high number of recordings in the 26-30 group is a corollary of the recruitment and data collection method: at the beginning of the L1 data collection, two participants (one female and one male) took part in the project twice and one female participant contributed three interviews.

**Table 7**

*The Number of Speakers in and Recordings of the L1 Interviews for Each Demographic Group*

| Selection criteria | Demographic group | No. of speakers | No. of recordings |
|---|---|---|---|
| Region | northern China | 18 | 20 |
| | southern China | 7 | 9 |
| Age | 20-25 | 3 | 3 |
| | 26-30 | 14 | 18 |
| | 31-35 | 6 | 6 |
| | 36-40 | 2 | 2 |
| Gender | female | 15 | 18 |
| | male | 10 | 11 |

Since the majority of the L1 participants were in China and I was in New Zealand at the time of data collection, most of the L1–L1 conversations had to be conducted online. As Table 8 shows, the L1–L1 conversations were mainly conducted via WeChat. WeChat is a multifunctional messaging and calling application which enables users to make free video or audio calls. It has a phone and desktop version. In this study I used my smartphone to make WeChat audio calls with each L1 participant and then used the built-in recorder on my laptop to record all the L1–L1 conversations. Aside from the WeChat interviews, there were four face-to-face interviews which were conducted in New Zealand. Even though the computer with the built-in recorder was placed in front of me rather than the L1 participants, the quality of most of the WeChat audio recordings was good enough to be applicable to orthographic transcription. However, it was unavoidable that one or two recordings were affected by the unsteady Internet connection; as a result, the turns of certain L1 speakers on the audio recordings were not audible which affected the transcription procedure (see 4.3.1 in Chapter 4). This is an important issue that I did not anticipate. In retrospect, it would have been helpful to require the participants to make recordings as well to guarantee the good quality of recordings of online conversation.

**Table 8**

*Interviewing Channels for Gathering the L1 Data*

| Interviewing channel | No. of speakers | No. of interviews |
|---|---|---|
| Face-to-face | 3 | 4 |
| WeChat audio | 22 | 25 |
| **Total** | **25** | **29** |

In addition to the interviewing channels, there are other features of the L1–L1 informal interaction. According to the Information Sheet, an interview duration of 30 minutes was initially suggested to each L1 participant and each of them would talk with me three times. When I started collecting spoken L1 data, however, it seemed inappropriate to set a 30-minute countdown timer for each interview and ask the L1 participants to end the conversation at the end of this period. In consequence, as Table 9 shows, the length of these L1 interviews varies considerably. At the beginning of the spoken L1 data collection, by design, I talked with three participants two or three times; however, I was aware of the possibility that fewer than 10 L1 participants would be needed if each of them was required to engage in two or three 30-minute conversations with me. Acknowledging the negative effect of this possibility on the

representativeness of the L1 corpus, I then decided to have only one interview with each L1 participant who later joined the project.

**Table 9**

*Metadata of the L1-L1 Interactions*

| Metadata | Category | No. of speakers | No. of interviews |
|---|---|---|---|
| Length of each interview (mins) | 20-30 | 9 | 9 |
| | 30-40 | 13 | 14 |
| | 40-50 | 4 | 5 |
| | over 60 | 1 | 1 |
| Number of interviews per person (times) | one | 22 | 22 |
| | two | 2 | 4 |
| | three | 1 | 3 |

*Note*. The numbers in the third column (i.e., *No. of speakers*) were given according to the length of each interview rather than the total number of the L1 participants involved in this project.

During the interviews, as the interviewer, my tasks were mainly to clarify questions and elicit responses as I consciously encouraged the L1 participants to speak as far as possible. Consequently, I did not contribute equally to each conversation (see Section 5.3 in Chapter 5 for further discussion). The degree of the interactivity between each L1 participant and me varied due to the relationships between us, the topics, and the willingness to communicate of the interviewees. As regards topics, the range of the topics was relatively broad, reflecting the L1 participants' interests, with no topic bias assumed. Some popular topics covered in the L1 data include travelling, food, culture, and language teaching experiences.

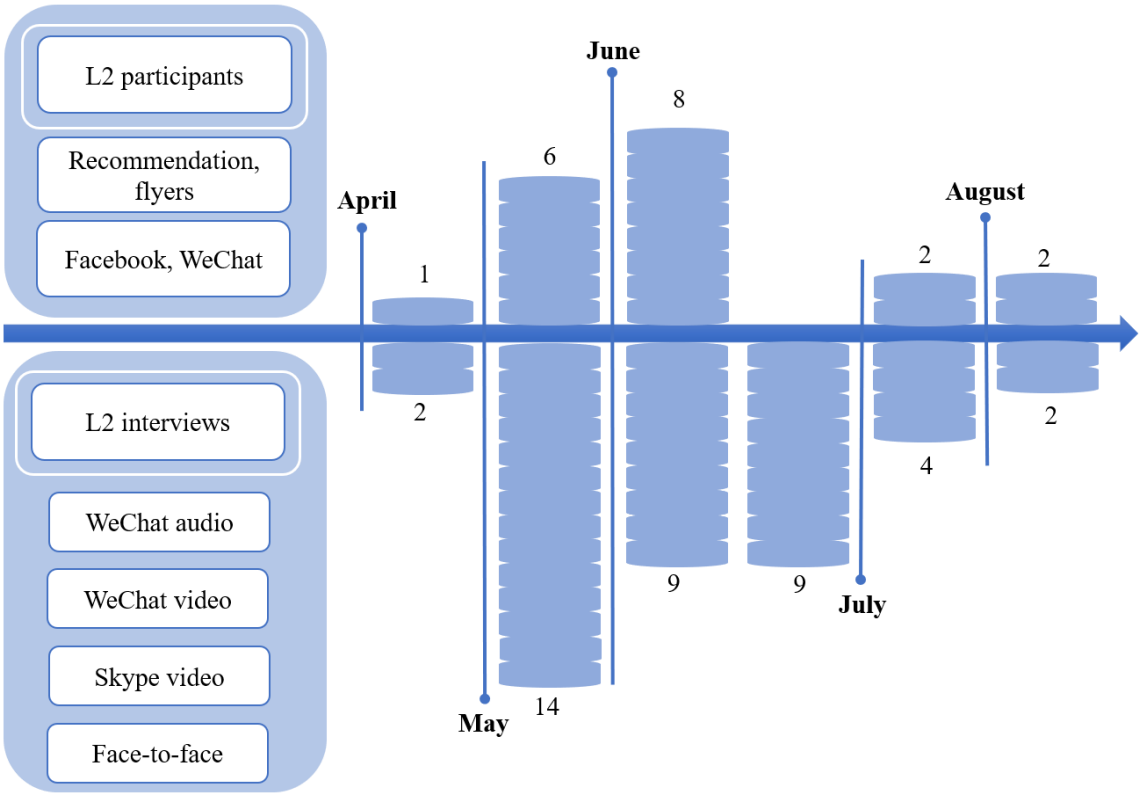
## 3.5 The L2 Group: Participant Recruitment and Data Collection

In this study, all L2 participants were volunteers who gave up their own time to attend the interview. They were recruited personally rather than via any institutions. Figure 6 shows that the majority of participants were recruited between April and July in 2018. However, the recruitment period was extended to August as two more L2 participants contacted me and showed their interest in the project. Participants were required to complete informed consent forms prior to recording; interviews hence were conducted under the awareness of the established conditions. Moreover, L2 participants were not encouraged to prepare for each

interview to ensure they could talk as naturally as possible. In this section I will introduce the process of the L2 participant recruitment and data collection.

**Figure 6**

*Timeline of L2 Participant Recruitment and Interviews*




*Note*. The upper side in terms of the L2 participants shows the number of participants recruited from April to August 2018; the lower part indicates the number of interviews the L2 participants attended from April to August 2018.

### 3.5.1 The L2 Participants

As Figure 6 shows, I recruited L2 participants mainly in May and June 2018. Making contact with L2 participants initially proved difficult. Documentation regarding L2 participant recruitment reached some Chinese teachers and Chinese volunteers who worked at Confucius Institutes in New Zealand directly at the very beginning of data collection in 2018. With the help of several Chinese teachers/volunteers and my supervisors, I obtained opportunities to

explain my research request to some New Zealanders who could speak fluent Chinese through emails and WeChat. All of them kindly agreed to help me with my data collection and some of them then introduced their friends to me. I also tried using social media platforms (Facebook, WeChat and Twitter) as well as flyers to spread my request for volunteers widely. Finally, 19 L2 participants agreed to take part in the project. All L2 participants were strangers to me until they were contacted for the interviews.

Seventeen New Zealanders and two Australians participated in the project. They were all native English speakers and none of them were Chinese heritage learners/speakers. There was one female participant with a Pacific family background. All the other L2 participants were Pākehā (New Zealand and Australian Europeans). Table 10 shows that the majority of the L2 participants were aged between 18 and 25. Unpredictably, female L2 speakers of Chinese are underrepresented due to the relatively small number of participants. The resulting sampling to some extent ran the risk of being biased in favour of males (for further discussion, see Chapter 5).

**Table 10**

*The Demographic Features of the L2 Participants*

| Selection criteria | Demographic group | No. of speakers |
|---|---|---|
| Nationality | New Zealand | 17 |
| | Australian | 2 |
| Age | 18-25 | 9 |
| | 26-30 | 3 |
| | 31-35 | 4 |
| | 36-40 | 3 |
| Gender | female | 5 |
| | male | 14 |

The likelihood that enough material would be obtained from these L2 participants was largely dependent on their Chinese proficiency level. It was intended that qualification of the HSK test and the number of the years of learning Chinese would be the criteria used to identify their proficiency level (see 3.2.2). However, there was no requirement for L2 participants to actually show their HSK test results to the researcher prior to the interview taking place. Thus, the measure of institutional status was the main method I used to assess the proficiency levels of L2 participants. This method was commonly employed by some Chinese teachers when they recommended participants to me. Three L2 participants were studying Chinese in New Zealand

at university and were identified by their Chinese teachers as intermediate or advanced speakers of Chinese since they were at Year 2 or Year 3. However, two of them actually could not communicate sufficiently with me in Chinese. So, it seemed to me that it was imprecise to identify L2 proficiency of participants on the basis of the years of learning Chinese, for it obviously ignored many important variables in terms of participants and learning environment. Additionally, several participants were recommended as their Chinese teachers assigned their proficiency on the basis of the results of their HSK tests. Given that the HSK test is a measure of written Chinese proficiency for L2 speakers of Chinese, it was not an efficient criterion to identify L2 participants' oral abilities. To sum up, there was no available objective criterion that could be employed to evaluate L2 participants' proficiency in spoken Chinese. However, all of them had read the information sheet (so they were aware that the present research required higher Chinese proficiency level), and the majority of the L2 participants self-identified as intermediate to advanced speakers of Chinese. Further discussion on L2 proficiency will be given in Chapter 7.

### 3.5.2 The L2 Interviews

In this study 40 interviews were gathered for the creation of the spoken L2 corpus[30]. Table 11 shows the number of audio recordings produced by each demographic group.

**Table 11**

*The Number of Recordings of L2 Interviews Produced by Each Demographic Group*

| Selection criteria | Demographic group | No. of speakers | No. of recordings |
|---|---|---|---|
| Nationality | New Zealand | 17 | 36 |
| | Australian | 2 | 4 |
| Age | 18-25 | 9 | 18 |
| | 26-30 | 3 | 5 |
| | 31-35 | 4 | 10 |
| | 36-40 | 3 | 7 |
| Gender | female | 5 | 11 |
| | male | 14 | 29 |

[30] Three interviews were conducted after I finished transcribing the recordings and had achieved the goal of gathering 200,000 words of L2 data. Since the main task at that stage was to transcribe the gathered L1 recordings, I left these three L2 recordings aside for later inclusion in the corpus.

It is noted that the 18-25 group contributed more interviews than other age groups. As regards gender, there are almost three times as many recordings produced by male speakers as female speakers of Chinese. It seems clear that these features of the L2 interviews differ from the L1 interviews. Further discussion on the differences between the two groups will be given in Chapter 5.

In this study, L2 participants had various options in terms of channels for conducting interviews open to them (see Table 12). Some participants who were in New Zealand preferred talking over Skype; participants who were in China preferred WeChat interviews. There were also three face-to-face interviews that were conducted in New Zealand. Various interviewing channels were used to assemble spoken L2 Chinese, while all the interviews were recorded in audio format only.

**Table 12**

*Interviewing Channels for Gathering the L2 Data*

| Interviewing channel | No. of speakers | No. of interviews |
|---|---|---|
| Face-to-face | 2 | 3 |
| Skype video | 3 | 7 |
| WeChat video | 4 | 9 |
| WeChat audio | 12 | 21 |
| **Total** | **19** | **40** |

As for topics, there was no specific topic designed prior to each interview. Each L2 interview started with their Chinese learning experience, which was designed to make the L2 participants feel at ease. Topics for L2 participants needed to be easily accessible to all and to include matters they were likely to have talked about in Chinese before, including Chinese learning experience, life in China (e.g., study, food, culture), and travelling. What should be mentioned is that some of the L2 interviews were rather monologic in nature with the interviewer's contribution consisting mostly of back-channelling (see Chapter 5).

Although the suggested interview duration for L2 participants was 10- to 20 minutes, I did not end the conversation if the L2 participant was willing to continue it. Consequently, the length of conversations varies considerably. Table 13 reveals that there were six L2 recruits who contributed one interview per person. Some of them were not able to talk with me again during the data collection period as they were busy; some contacted me after I had gathered

over 200,000 words of speech, at which time I was in the process of revising my transcripts rather than focusing on collecting data, which made it impossible to gather more recordings.

**Table 13**

*Metadata of the L2 Interviews*

| Metadata | Category | No. of speakers | No. of interviews |
|---|---|---|---|
| Length (mins) | 10-20 | 6 | 8 |
| | 20-30 | 6 | 7 |
| | 30-40 | 10 | 12 |
| | 40-50 | 8 | 11 |
| | over 60 | 2 | 2 |
| Number of interviews per person (times) | one | 6 | 6 |
| | two | 6 | 12 |
| | three | 6 | 18 |
| | four | 1 | 4 |

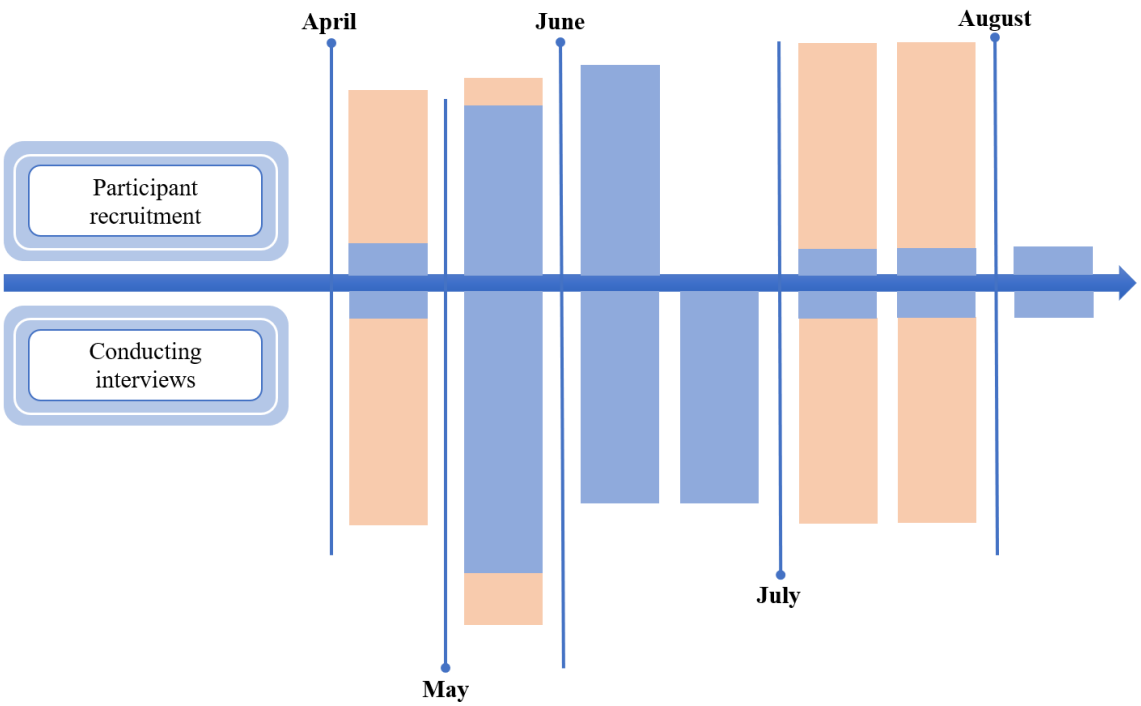
*Note*. The numbers in the third column (i.e., *No. of speakers*) were given according to the length of each interview and the interviewing channels rather than the total number of the L2 participants involved in this project.

As indicated above, it seemed to me that it was not feasible to make sure the L2 participants had mastered a certain level of Chinese sufficient to conduct a conversation. In consequence, two participants could reasonably be characterized as beginners, as they shifted to English quite often in order to communicate more efficiently. Therefore, their interviews were not included in the spoken L2 corpus (see Appendix J). Detailed speaker and interview metadata information is given in Appendix G.

To conclude, all interviews were conducted at the participants' convenience. Figure 7 below presents the timeline of the whole procedure with respect to participant recruitment and data collection. Recruiting participants, in particular L2 participants, was a time-consuming task. To save time as well as to gather enough spoken data as soon as possible, the tasks of participant recruitment and data collection were carried out at the same time. The comparability of the two groups will be discussed in some detail in Chapter 5.

**Figure 7**

*Timeline of Participant Recruitment and Data Collection*




*Note*. The blue parts represent the timeline of L2 participant recruitment and L2 interviews conducted from April to August 2018; the orange parts indicate the timeline of L1 participant recruitment and L1 interviews conducted from April to July 2018.

## 3.6 The Audio Recordings of Interviews

As a reminder, all interviews were recorded using a built-in recorder on my password-protected laptop, which directly made all audio recordings electronically. This section therefore reflects on two decisions made with respect to the collected recordings for the creation of the spoken Chinese corpus: the decision of collecting audio-only recordings and the role of the collected recordings in this study.

When designing the spoken Chinese corpus, it was decided that recordings would be made in audio format only. In the field of corpus linguistics, there is an increasing desire for spoken corpora that move beyond the textual dimension of communication (e.g., Adolphs &

Carter, 2013; Adolphs et al., 2015). Accordingly, there have been a number of attempts to build multimodal corpora which indeed include access to "an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language" (Knight, 2011, p. 392). Undoubtedly, multimodal corpora are invaluable resources for studying and analysing human communication (Adolphs & Carter, 2013; Allwood, 2010; Knight, 2011; Thompson, 2010). However, in this study I made no attempt to make video recordings, since lexical patterns of language use were the main focus of the current study.

The second decision I made in terms of the recordings was to use the built-in recorder on my laptop to record all the interviews. As has been discussed previously, the disadvantage of this approach is that, since the majority of interviews took place online, the quality of the turns of participants could be sometimes largely affected by unstable Internet connections. For good audio recordings of online communication, it seems necessary to place more than one recording device to record each participant individually. For future corpus projects aiming to gather online communication, this alternative approach can be expected to gather a high quality of recordings. Due to the outbreak of COVID-19, many researchers have been forced to transition from face-to-face communication to Internet-based communications (e.g., online conferences). Within this background, researchers have provided some quite useful discussions on recordings of online communication. For example, Busse and Kleiber (2020) share their experience of using *OBS Studio*[31] to record all live events for an international online conference. Lobe et al. (2020) describe and compare some software tools for researchers who need to carry out online data collection, including *Zoom*, *Skype*, and so on. Whether the methods demonstrated in these studies are feasible to be employed to spoken Chinese corpus construction and facilitate high quality of audio/video recordings need to be discussed further, but undoubtedly, they provide a new direction for data collection.

Once obtained the recordings, to take full advantage of the audio recordings collected for a spoken corpus, the usual first choice is transcription. Ttranscription of speech tends to be treated as the primary data in its own right in practice (Wichmann, 2008). With advanced technology, now there is the potential evidence of a video and audio stream tied to the transcript offering invaluable contextual and paralinguistic support to spoken corpus studies (McCarthy & O'Keeffe, 2010). Admittedly, spoken corpora can be considerably enhanced by the alignment of transcriptions to audio and, increasingly, video files through the insertion of

[31] OBS Studio: https://obsproject.com/.

appropriate ‘time’. In this study, all audio recordings however would not be publicly accessible, due to (i) the promise made to participants that their identities would not be revealed, and (ii) the fact that their permission was sought only for publication of the transcriptions and not the recordings themselves. It should be noted that the original recordings are retained, and it is always possible to return to the originals to re-transcribe the interaction if required.

## 3.7 Summary

In this chapter I stated that the spoken Chinese corpus was designed to represent present-day spoken Chinese used in mainland China. Bearing this aim in mind, 25 L1 speakers and 19 L2 speakers of Chinese contributed 60 unstructured interviews for the creation of the spoken Chinese corpus. Compared to the existing spoken L2 Chinese corpora listed in Chapter 1, the L2 group offered a much narrower range of L1 backgrounds: New Zealanders of non-Chinese ethnicity were included in this study. The focus on New Zealanders (and two Australians) of non-Chinese ethnicity led the researcher to pay attention to variables that have so far been underrepresented in L2 learner research, such as gender, age, and speaker roles. In addition, considering the difficulties of recruiting L2 participants, I invited some L2 participants to attend interviews three or four times, thus potentially increasing the presence of their communicative style in the data. The same issue could also be found in the L1 group. As pointed out previously, three L1 participants were encouraged to take part in two or three interviews. This decision may have influenced the inferences drawn from both the L1 and L2 data, and the possible influences therefore will be discussed and addressed in Chapter 5.

It is conventional to give a description of the L2 speakers in the building of an L2 corpus (e.g., Granger, 1998b, 2002, 2003); however, considerably less discussion seems to be available about how the L2 data were collected and how reliability of data collection methods was ensured (Gablasova et al., 2019b). Following the call for a transparent corpus design and detailed data collection methods, this chapter also provided a transparent procedure and a detailed description of the participant recruitment and data collection. Although the two groups were designed with the similar criteria, compromises had to be made due to practical constraints when collecting data. Variables that may impact the comparability of the two groups, and issues about how the L2 group resembles or differs from the L1 group will be considered carefully in Chapter 5. At the end of this chapter, I explained briefly the use of the

gathered audio recordings. Based on these audio recordings, Chapter 4 will account for the transcription process.

# Chapter 4 Transcription

## 4.1 Introduction

In this chapter I address the transcription of spoken data collected for the spoken Chinese corpus. To begin with, Section 4.2 considers the transcription design, highlighting the essential role that research purpose plays in the design of the transcription for the spoken Chinese corpus. Section 4.3 outlines the transcribing process associated with several tools employed in the transcription process. Section 4.4 provides a detailed account of the methodological decisions reached regarding the transcription of spoken Chinese, seeking to raise awareness of the nature of informal spoken Chinese to avoid transcription biases as far as possible. In Section 4.5, there is a particular focus on several important features captured in the transcription, offering a systematic way to address these features in the development of the spoken Chinese corpus. The spoken Chinese transcription scheme is provided in full in Appendix H accompanying the thesis. Given the above concerns, Section 4.6 further outlines strategies adopted to achieve a faithful transcription and to maximise the reliability and replicability of the transcription. Finally, this chapter concludes with a summary in Section 4.7.

## 4.2 Transcription Design

As the written representation of speech, transcriptions rather than audio/video recordings tend to be treated as the primary data and form the basis for the systematic analysis of speech (Breiteneder et al., 2006; Wichmann, 2008). Transcription quality therefore has direct implications for the reliability and usability of a spoken corpus (Gablasova et al., 2019b). Accordingly, it is necessary to reflect on choices made concerning what features of speech to preserve, what categories to use, and how to organize and represent these features in transcription before conducting all subsequent analyses of speech. In this section I discuss two main decisions made in the design of the transcription for the spoken Chinese corpus. The first decision is to carry out an orthographic transcription, and the second is to choose appropriate transcription guidelines.

### 4.2.1 Orthographic Transcription

Given current technology, corpus studies of spoken language still rely heavily on "a written medium" (Hunston, 2008, p. 160), which gives rise to the essential issue of how to best represent multi-layered spoken language with written records. Generally, the speech used in spoken corpora is recorded and represented "at least to the level of an orthographic transcription" (Wichmann, 2008, p. 195). Orthographic transcripts of these spoken corpora enable users who are interested in lexical, grammatical or pragmatic phenomena to carry out their research. For users whose interest lies primarily in the sounds of speech, it is of great value to capture information concerning prosodic features. Alongside the standard orthography, for instance, the LLC also includes detailed prosodic information which has remained a valuable resource for research communities over the years (Peppé, 2014; Svartvik, 1982, 1990; Svartvik & Quirk, 1980).

There are several concerns involved in this filtering process. First of all, transcription is a notoriously laborious and time-consuming task. For example, the TLC project involved "more than 3,500 hours of transcription time with many more hours spent on quality checking and post-processing of the data" (Brezina et al., 2019, p. 119). Also, transcription requires a considerable amount of skill and specific expertise (e.g., phoneticians are needed in order to produce a prosodic transcription). For spoken corpus projects, the cost and time associated with skills required for transcription should be taken into consideration. However, the development of a transcription for a spoken corpus should not be dominated by these factors. Rather, "the theoretical and analytical motivations of transcription" (Breiteneder et al., 2006, p. 172) are integral to and should guide the transcription design. With the research purposes in mind, corpus compilers therefore can "tailor the focus of the transcription accordingly" (Adolphs & Carter, 2013, p. 12), and then produce a transcript which is likely to "greatly help in finding regularities of interest free from the distraction of irrelevant detail" (Edwards, 2014, p. 19). In his account of transcription for the spoken component of the BNC, Crowdy (1994) proposes three essential questions, which foreground the significance of research purposes when designing a transcription scheme, i.e. "Who is the transcription for? How will it be used? What are the important features?" (p. 24).

In the present study, while the initial purpose of the corpus aimed to focus mainly on lexical applications, as noted in Chapter 1, the focus shifted to discourse markers. This change in focus accordingly affected decisions made about transcription. This corpus was also intended

to be used to facilitate the quantitative study of lexis, syntax, and pragmatics in terms of spoken Chinese for other users. Unlike those large-scale spoken corpora projects (e.g., the Spoken BNC2014 and the TLC), for the spoken Chinese corpus as a doctoral research project very limited transcription time was available. As Cameron (2001) suggests, for doctoral students who have a limited time to work on their spoken data, "it is important to keep expectations reasonable, and to develop the ability to judge when the transcript is good enough for the purpose at hand" (p. 39). Bearing these considerations in mind, an easily manipulated and consistently coded orthographic transcription would be sufficient to meet the demands likely to be placed upon the spoken Chinese corpus. To this end, a key decision made at an early stage was to adopt standard simplified Chinese characters to represent speech collected for the spoken Chinese corpus. The processing stages of the transcription are described below in Table 14. A detailed account of these transcription conventions is given in the following sections.

**Table 14**

*The Processing Stages of Transcription*

| Levels | | Features |
|---|---|---|
| **1** | Lexical vocalisation | words<br>speaker turns<br>numbers and dates<br>false starts and repairs<br>repetitions<br>truncated (unfinished) words<br>spelt-out words<br>the third person singular pronouns<br>acronyms and abbreviations<br>English units<br>unclear words<br>unfamiliar words<br>uncertain words |
| **2** | Semi/non-lexical vocalisation | backchannels<br>minimal response tokens |
| **3** | Learner language features | pronunciation errors |
| **4** | Anonymization | names, places, institutions |

### 4.2.2 Transcription Guidelines

Transcription effectively reflects our understanding of the way language functions and which features are seen as significant and worth capturing. However, transcription cannot be too

idiosyncratic and there is a need to follow certain transcription guidelines in order to make it reusable by the research community (Adolphs & Carter, 2013). To promote the validity and reliability of the transcripts and make the transcription maximally usable, I adopted two transcription devices which are summarized by Chafe (2014) to guide the spoken Chinese corpus transcription process: (i) take advantage of what users already know and (ii) representations should have an iconic value and be easy to manipulate, "so that the burden of learning arbitrary conventions is to that extent mitigated" (p. 55).

According to the discussions on transcription in Chapter 2, it is concluded that there is no systematic transcription scheme which is widely accepted and used in the compilation of spoken Chinese corpora. In addition, the credibility of the transcribed Chinese data in previous studies could be questioned on the grounds of methodological and theoretical deficiencies in transcription systems currently in use. Moreover, the spoken data on which these studies are based are not available, and the transcription procedure is neither transparent nor discussed in a systematic manner, which somewhat diminishes the reliability and replicability of these transcription conventions. As a result, it was decided not to directly reuse the transcription conventions discussed in the Chinese research literature and take these conventions for granted "as mundane and unproblematic" (Lapadat, 2000, p. 204). It then was worth taking spoken English transcription conventions into consideration, for many differing transcription standards exist both in corpus linguistics and discourse analysis (an overview can be found in Andersen, 2016) and have been explicated to a large extent systematically.

Having considered some existing transcription systems and their principles, I decided that the spoken Chinese corpus would broadly follow the transcription conventions observed in the Spoken BNC2014[32] (Love et al., 2017) and the TLC (Gablasova et al., 2019b, p. 154). As a well-designed spoken corpus, the Spoken BNC2014 is publicly available via the CQPweb interface[33] and its transcription procedure is transparent, which maximises the reliability and replicability of the transcription. As regards the TLC, which is a rather new spoken L2 corpus, the compilers have already discussed its transcription in some detail, and the availability of spoken data associated with the transcription guidelines makes this transcription scheme usable (e.g., Gablasova, Brezina, & McEnery, 2017; Gablasova et al., 2019a; Gablasova, Brezina, McEnery, et al., 2017). Since all the features normalised in these transcription systems relate

[32] The British National Corpus 2014 user manual and reference guide (version 1.1) can be found in http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf.
[33] The CQPweb: https://cqpweb.lancs.ac.uk/.

to spoken English research, it was problematic to simply adopt the transcription conventions to guide the spoken Chinese corpus transcription. Therefore, it was necessary and beneficial to listen to the audio recordings multiple times before making specific decisions in terms of the important features of spoken Chinese; it was only in this way that it would be possible to arrive at an accurate impression of the spoken Chinese data which could then be the basis of developing the transcription of both L1 and L2 speech.

## 4.3 Transcription Process

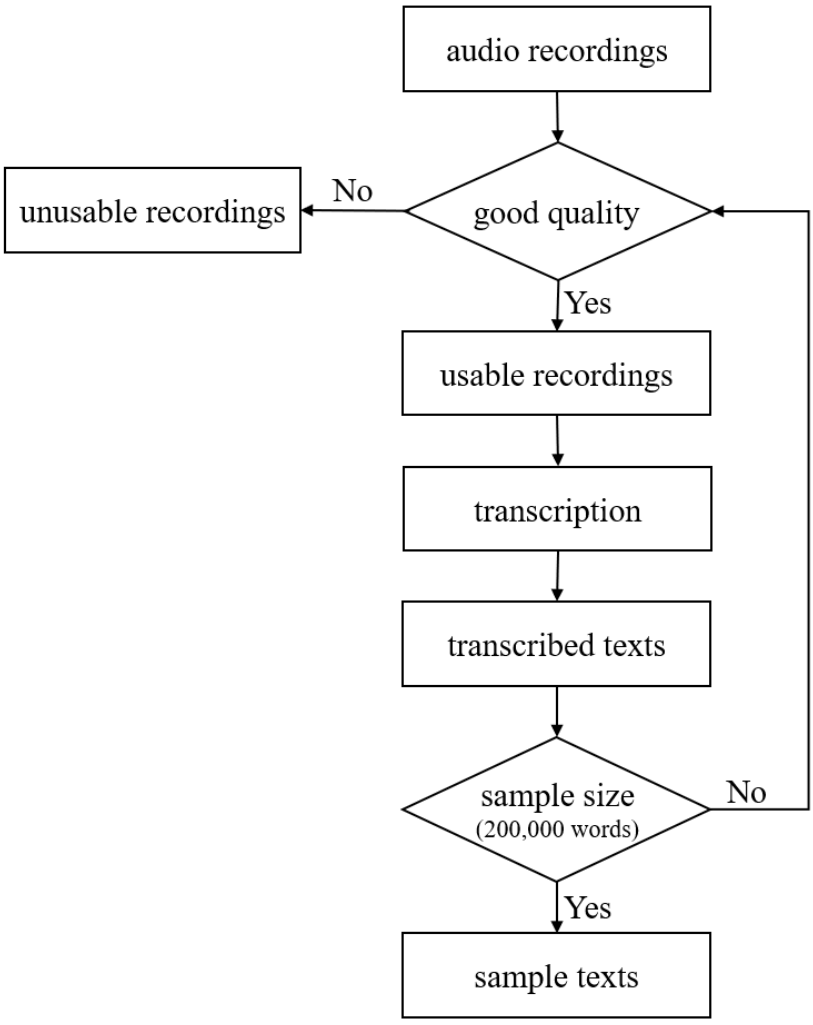
In this section I give a brief description of the transcription process associated with the tools that were employed in the process.

### 4.3.1 Getting Started

Since transcription is a laborious and time-consuming task, before carrying out the transcription, I checked the quality of all the audio recordings gathered for the spoken Chinese corpus (see Figure 8). It appeared that not all audio recordings were clear enough to facilitate accurate and reliable transcription. Several of them did not have a good audio signal, often with only one side of the conversation (i.e. my turns) audible. Aside from the poor-quality recordings, a few L2 recordings included quite lengthy periods of English conversations, which were of limited value to contribute to the spoken L2 corpus. Fortunately, as noted in Chapter 3, three more L2 participants took part in the project after I finished the data collection and transcription. Given that not all the L2 recordings would be usable for the compilation of the spoken Chinese corpus, these three recordings could have been substitutes for those unusable transcribed texts. However, due to time limitations, these three recordings were not transcribed in this study.

**Figure 8**

*Flowchart of Transcription Process*




*Note*. If the recording was of poor quality then it would be identified as an unusable recording which would not be transcribed; if it was of good quality then it would be transcribed. As each subcorpus aimed to include 200,000 words, the process of transcription was not completed until the goal was achieved; otherwise the same path was repeated again.

### 4.3.2 Transcription Tools and Process

Several tools were used to assist the transcription process (e.g., *iFlynote*, *InqScribe*, and *Praat*). Having investigated and evaluated some automatic transcription tools, I first used one of them,

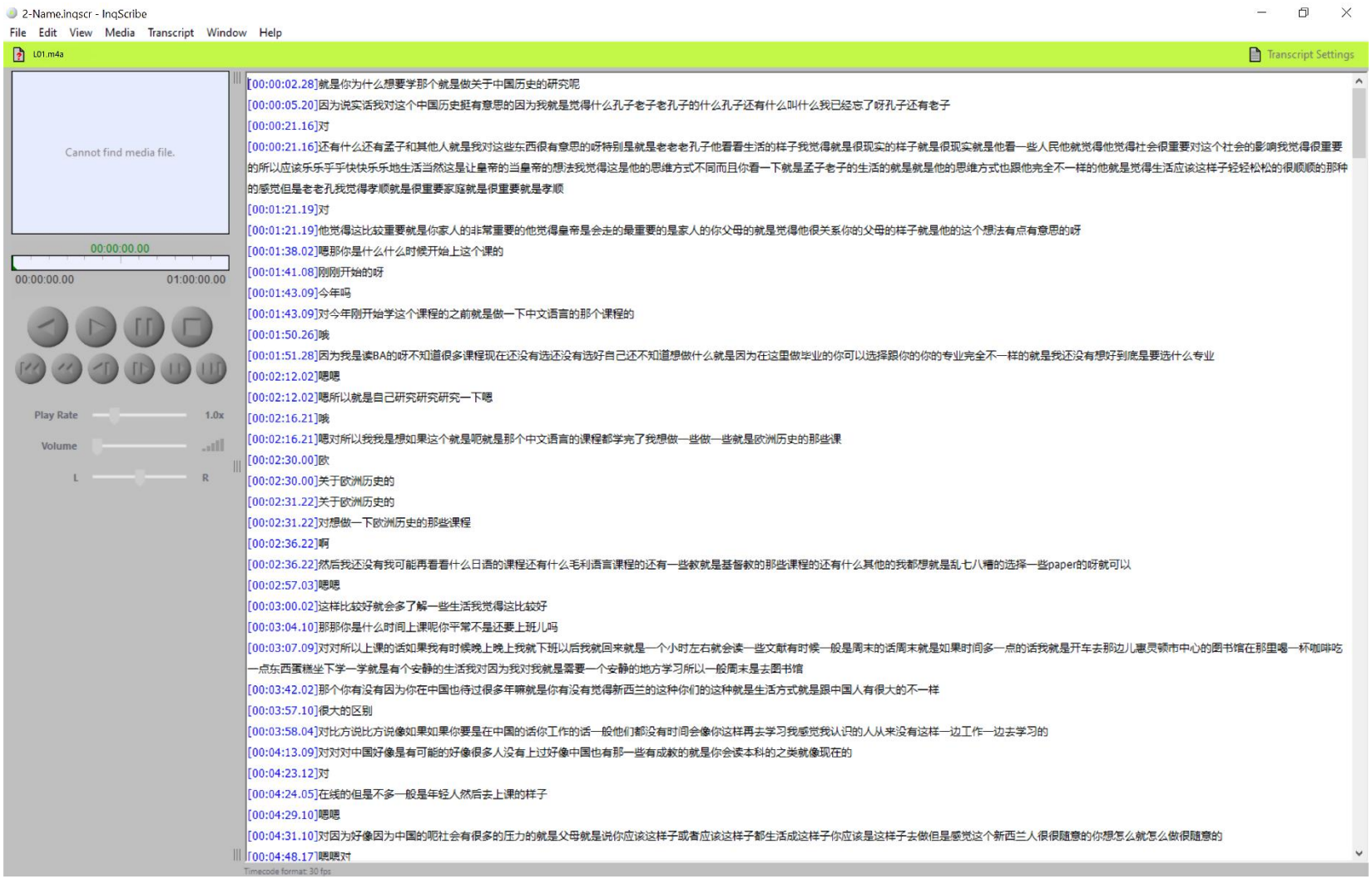
*iFlynote*[34] (a cloud note app which is developed by a Chinese information technology company), to support the process of transcription. However, it turned out that the transcription produced by this tool was insufficiently accurate for linguistic analysis. It was obvious to me that there was no getting around the fact that all the recordings must be manually transcribed for the creation of the spoken Chinese corpus. Considering the cost and training time as well as the ethical importance of anonymity, and, most importantly, to promote my familiarity with the data, I decided to transcribe all the recordings manually myself rather than recruit other transcribers.

In addition to *iFlynote*, a transcription software package called *InqScribe*[35] was also adopted in this study. While *InqScribe* is a transcription platform which supports manual transcription only, it enables users to play and transcribe audio/video recordings in the same window and to insert timecodes in the transcript (see Figure 9) which helped speed up the process of transcription considerably. Moreover, it also links the timecodes to the relevant points in the medial file so that the transcriber can easily locate specific sections of recordings. *InqScribe* is a particularly useful tool with audio recordings where there are no visual cues to help identify any given moment in the recording. By using this tool, at the outset of transcribing, I basically wrote down what I heard on the recordings first to arrive at as accurate an impression of the spoken data as possible, and then I went through all the recordings and sketched the general outlines. In this phase, there were some difficulties in transcribing the L2 speech. Due to inaccuracies in articulation and tone patterns, it was not always possible in the first instance to make a confident interpretation of what was being said by some participants. However, when such utterances were observed in their broader contexts of occurrence, it was sometimes possible to guess with a reasonable degree of confidence the speaker's likely intended meaning. In such cases, the intended meaning was recorded in the transcript rather than the actual sounds articulated. For example in (3), the L2 speaker of Chinese mentioned a Chinese movie—Wolf Warrior II—which he saw with his Chinese roommates, but I had trouble to follow him because of his nonstandard pronunciation of the name of the movie. However, the name of the movie became quite clear when I listened to the recording; therefore, in this case, I decided to record the correct name of the movie that the L2 speaker tried to communicate rather than the actual sounds articulated.

[34] *iFlynote*: http://www.iyuji.cn/iyuji/home.
[35] *InqScribe*: https://www.inqscribe.com/.

**Figure 9**

*Screenshot from InqScribe*



(3) An extract from the speech of a male L2 speaker of Chinese

<L11> 我偶尔跟他们看电影是吧看那个叫什么我们看什么兰战狼二

I occasionally with them see movies right see that what is the name we saw what lan Zhan Lang er

*I saw movies with them occasionally yeah (we) saw that (movie) it is called you know we saw you know lan Wolf Warrior II*

<S00> 战狼二是什么

Zhan Lang er is what

*What is Wolf Warrior II*

<L11> 战狼二这个电影

Zhang lang er this movie

*Wolf Warrior II the movie*

<S00> 我不知道

I do not know

*I do not know that movie*

<L11> 啊我可能说错了战狼

uh I may say it wrong Zhan Lang

*uh maybe my pronunciation was wrong Zhan Lang*

Carrying out the preliminary transcription, it was evident that there were a large number of features (e.g., pauses) which could not be captured effectively by using *InqScribe*. Thus, a computer program which is widely used for analysing, synthesizing and manipulating speech: *Praat*, was adopted at the later stage of transcription (see Figure 10). *Praat* was of great help in post-checking errors and filling in details (e.g., timing pauses in a relatively accurate way) (for a detailed user guide see Boersma, 2014). By using these tools, the first version of transcripts of a total of 36 recordings of L2 speech and 26 recordings of L1 speech were produced in 2018. Table 15 shows the timeline of the transcribing process. All the transcripts were saved as UTF-8 coded text files (for more details on Unicode Transformation Format, see McEnery & Xiao, 2005). Detailed information in terms of the L1 and L2 transcribed texts can be found in Appendices I and J. These plain text files would be released as one version of the spoken Chinese corpus for users interested in this corpus in the near future (for information of the release of the spoken Chinese corpus, see Chapter 5).

**Figure 10**

*Screenshot from Praat*

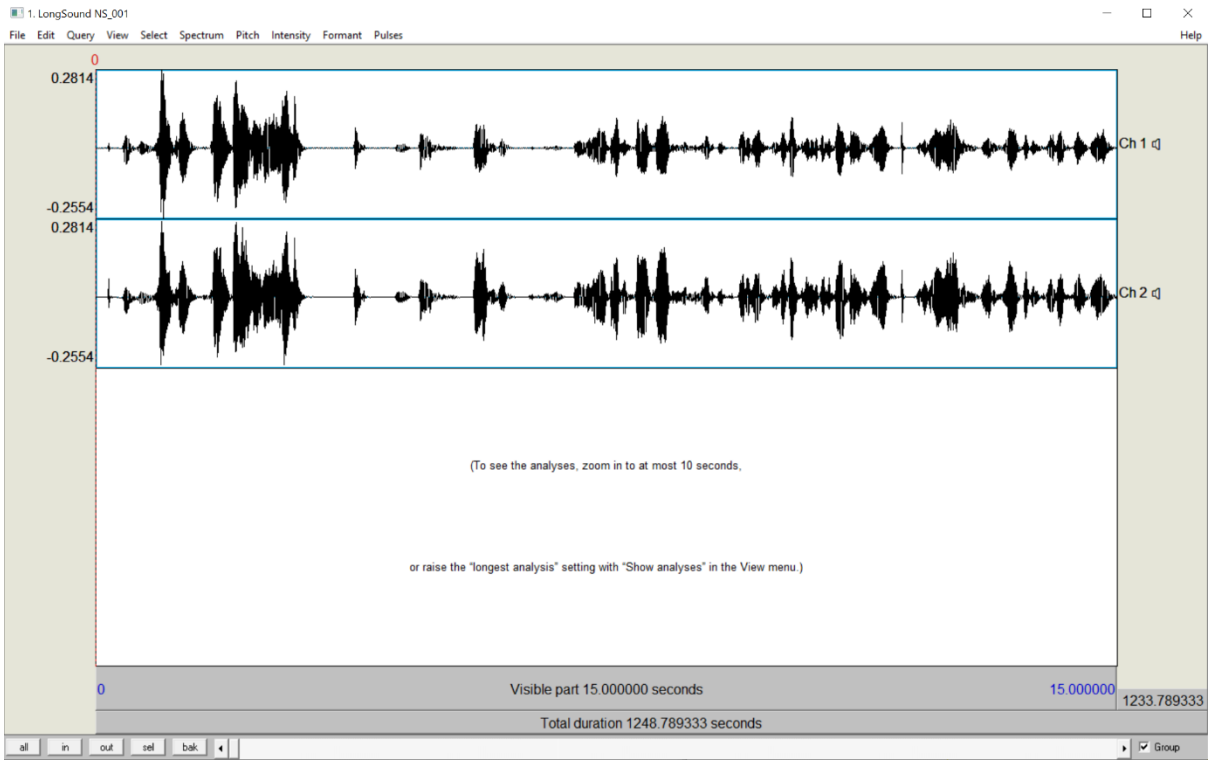



**Table 15**

*Transcripts for the Spoken Chinese Corpus*

| Data | Gathered recording | Untranscribed recording | Start date | End date |
|---|---|---|---|---|
| L1 speech | 29 | 3 | 23 July 2018 | 05 August 2018 |
| L2 speech | 40 | 4 | 25 June 2018 | 20 July 2018 |

*Note*. The final edition of the L2 corpus contains only 34 transcripts. Two transcripts that included many English expressions were not used for the building of the spoken L2 corpus (see Appendix J).

## 4.4 Considerations of Features

When transcribing the recordings, a number of features were noticeably identified, such as non-verbal sounds (e.g., cough), repetitions, and overlapping speech. As discussed previously, a transcription should be made in accord with the purposes of the research. With the research

purposes in mind, this section seeks to discuss which features are important and should be recorded in the transcripts.

### 4.4.1 Lexical/Non-Lexical Vocalisations

The general consensus in the body of research concerning English conversational interaction is that spoken language, especially informal speech, typically consists of a sizeable portion of features, such as false starts, repetitions, repairs, discourse markers (e.g., *well*), backchannels, and minimal response tokens (e.g., *erm*) (Atkins et al., 1991; Gablasova et al., 2019a; Garrard et al., 2011; Jenks, 2011; Kirk & Andersen, 2016; Lindsay & O'Connell, 1995). Anyone who works with talk needs to bear in mind that these features are not redundant and inarticulate; rather, they can also have very real interactional functions in signalling continued listenership, adding emphasis, and so forth. Failure to attend to these features as constitutive features of interaction may represent a significant omission and directly affect the subsequent analysis, so efforts need to be made to avoid hearing "spoken language in terms of the written model" (Cameron, 2001, p. 33). As has been noted previously, at least implicitly, the baseline of the simple orthographic transcription for the spoken Chinese corpus was to capture all the lexical vocalisations. That is, false starts, repairs, repetitions, discourse markers, and any other lexical vocalisations which could be represented in simplified Chinese characters would be recorded in the transcripts exactly as they occurred. Those features which were not part of the standard writing system, such as backchannels and minimal response tokens, but which could be signalled by the Romanization system *Pinyin* or English (as speakers used some English in the conversations), were captured in the transcription as well (see Section 4.5 for more information).

### 4.4.2 Non-Verbal Sounds

Spoken interaction is a complex and dynamic type of social communication which relies on verbal as well as non-verbal cues (e.g., laughter, cough, and inhalation) (Cameron, 2001; Jenks, 2011; Wichmann, 2008). A great many analysts who work with spoken language have pointed out that the non-verbal sounds carry valuable information in communication (Glenn, 2010; Holt, 2010; Jefferson, 1979; Jenks, 2009). As Table 16 shows clearly, non-verbal sounds are conventionally encoded in conversation analytic research as well as spoken corpus projects.

**Table 16**

*Features in Some Existing Transcription Systems*

| References | Transcription systems | Features | | |
|---|---|---|---|---|
| | | non-verbal sounds | tone/intonation units | overlaps |
| Atkinson and Heritage (1999); Lindsay and O'Connell (1995); Psathas and Anderson (1990); Sacks et al. (1974) | Gail Jefferson's transcription system | yes | – | yes |
| Du Bois (1991); Du Bois et al. (1992) | the Santa Barbara school conventions | yes | yes | yes |
| Nelson (2014) | the International Corpus of English (ICE) standard | yes | – | yes |
| MacWhinney (2000) | CHAT (Codes for the human analysis of transcripts) | yes | yes | yes |
| Payne (2014) | the Cambridge and Nottingham Corpus of Discourse in English (COBUILD) conventions | yes | – | yes |
| Crystal and Quirk (1964); Peppé (2014); Svartvik (1990); Svartvik and Quirk (1980) | the LLC conventions | yes | yes | yes |
| Crowdy (1994, 2014); Love (2017); Love et al. (2017) | the Spoken BNC conventions | yes | – | yes |
| Haslerud and Stenström (2014) | the Bergen Corpus of London Teenager Language (COLT) conventions | yes | yes | yes |
| Gilquin et al. (2010) | the LINDSEI guidelines | yes | – | yes |
| Gablasova et al. (2019b) | the Lancaster Spoken Language Transcription Guidelines (TLC) | yes | – | – |

Having briefly reviewed corpus-based studies in terms of language use, for instance studies of discourse markers, it seemed that while non-verbal sounds were included in the literature, they were not included in the analysis. In addition, Garrard et al. (2011) take the view that non-verbal sounds "should be removed in the interests of consistency" (p. 400), because the representation of various types of non-verbal sounds in transcription is somewhat idiosyncratic and unreliable. In this study, according to the aims of the spoken Chinese corpus, it appeared that the precise representation of these features would neither greatly enhance the corpus's value nor contribute to the subsequent analysis of Chinese discourse markers. In the end, it was decided to exclude all non-verbal sounds in the transcription. Ideally, the interpretation and analysis of such sounds should be left to researchers with access to the sound files who wish to "analyse the text in more detail" (Crowdy, 1994, p. 25). Admittedly, such additional analysis would not be possible in the initial release of the spoken Chinese corpus, because the sound files on which the transcripts were based would not be available as part of the spoken Chinese corpus due to ethical considerations.

### 4.4.3 Prosodic Features

As Chafe (1988) puts it, speech occurs in "a series of relatively brief spurts of vocalization" (p. 1). These spurts are called tone units (Crystal & Quirk, 1964) or intonation units (Chafe, 1988). In this study, instead of segmenting the flow of speech into tone units, I adopted the convention developed by the Spoken BNC2014, that is, taking speaker turns as the basic units (Love, 2017). What should be mentioned is that, in spoken corpora which consist of tone units, punctuation markers are usually used to signal the boundaries of these segments, e.g., the COBUILD corpus (Payne, 2014) and the BNC1994 (Crowdy, 1994, 2014). Since tone units were omitted in the present study, it was appropriate to exclude punctuation as well. In the literature on English, tone units associated with various levels of prosodic annotation (e.g., tones, pitch, pauses, and voice quality) were initially considered for inclusion in the transcripts (see Table 16). In this study, however, the idea of normalizing detailed prosodic features was ruled out: only the inclusion of pauses in the transcripts was considered, since the spoken Chinese corpus was chiefly concerned with lexis, syntax and pragmatics, not with prosody. Secondly, this decision was made partly on the grounds of economy. Even minimal prosodic transcription would be so costly in terms of time that it would be impractical for the present project. Moreover, a prosodic

transcription typically requires prosodic/phonetic expertise which was not possible given that no phonetician was involved in the project.

Despite the omnipresence of pauses in spoken communication, it is not uncommon to see pauses transcribed with subjective approximations in corpora (for examples, see Leech et al., 2014). As with turn-taking, pauses are not devoid of social meaning. Studies have shown that interactants use pauses to deal with overlapping talk in online settings where people can hear but not see each other (Jenks, 2009), display a preference or orientation to a previous turn, for example when declining an offer or invitation, and manage gaze and turn-taking practices (Goodwin, 1980). The length of pauses has also been shown to play an important role in the prosodic quality of surrounding talk (Krivokapić, 2007). Nowadays, many freely available media playback programs allow users to quickly and precisely time pauses (e.g., *Praat*), which has made the task of timing pauses easier than before. However, the task of marking pauses in the corpus transcripts is still time-consuming. This being the case, pauses were not recorded in the first version of the spoken Chinese corpus.

### 4.4.4 Interactional Features

Spoken communication entails a series of reciprocal, sequentially unfolding utterances and actions between speaker and listener (Jenks, 2011). When one or more speakers speak during another turn at talk, overlapping speech occurs. It is not uncommon to capture overlapping speech in informal conversations. Table 16 above shows that only the transcription scheme of the TLC does not mark overlaps, for (i) all conversations contain only two speakers in this L2 corpus, and (ii) L2 corpus research tends to be interested in lexis and grammar rather than conversational discourse (Love, 2017). Overlaps undoubtedly have discourse functions in talk, however, it seemed that they might have limited impacts on the analysis of discourse markers. In addition, making precise positions of overlaps was costly and did not seem necessary for the majority of analyses that use spoken corpora. As a result, overlapping speech was not captured in the transcripts of the spoken Chinese corpus.

In short, all the features discussed previously characterised spoken Chinese. With the research purposes in mind, a large number of features however were discarded to produce a simple orthographic transcription. Du Bois (1991) recommends that general transcription systems "must be adaptable", because "spoken discourse is complex enough in many layers

that it virtually demands to be approached from a variety of viewpoints" (p. 94). Following the principle of adaptability, additional layers of transcription could be added systematically later if the spoken Chinese corpus is to be employed for new research questions (Gablasova et al., 2019b).

## 4.5 Representing Informal Speech: Main Features

This section presents the main features identified and transcribed in the recordings for the creation of the spoken Chinese corpus, aiming to create a bespoke transcription scheme for this corpus. More information in terms of this bespoke transcription scheme can be found in Appendix H.

### 4.5.1 Lexical Words

The first step to develop a faithful orthographic transcription was to use standard orthography, i.e. simplified Chinese characters, to write down all the lexical vocalisations in linear forms exactly as they occurred in conversation. Several features with regard to words are emphasized in this subsection, including unclear words, uncertain words, and the third person pronouns in Chinese.

As has been noted in 4.3.1, some sound files were of poor quality which made certain parts of speech hard to be transcribed accurately. In consequence, it was necessary to make guesses during the transcription. Where a speaker was unclear but I was able to make a guess at what the speaker was saying, I wrote down the words and checked back over the guesses later by using *Praat*. If I was not sure my guesses were correct or not, a tag <uncertain=a guess> was added. Also, if I was unable to identify certain stretches of speech, these stretches then were marked as unclear words with timestamps, e.g., a tag <unclear=1.0> was used.

Unlike their English counterparts, *he* and *she*, the Chinese third person pronouns in the singular have exactly the same pronunciation *ta* (while distinctive Chinese characters are used in the writing system). In consequence, it was hard to decide whether 他 *ta* (he) or 她 *ta* (she) should be used in some situations when more context was not provided by participants. To save time and to represent speech as accurately as possible, where I was not able to make a guess at

the gender of the person mentioned by the participant, I used the capital letters ‘TA’ to signal the person (see the example below).

(4) An extract of the speech of a female L1 speaker of Chinese

<N01> 我也得见一下房东我不可能说房东不知道有 **TA** 这个房子里有我这号儿人对吧

I need to see the landlord I not possible to say the landlord does not know there is TA (his/her) house this house has me right

*I have to meet the landlord it is ridiculous that the landlord is not aware that I am living in his/her house right*

### 4.5.2 Anonymization

Given the ethical significance of anonymity, it was decided to omit any reference that would allow an individual to be identified from the transcription (Crowdy, 1994), including names, addresses, institutions, and so forth. English corpus compilers have made various approaches available to preserve anonymity. In corpus practice, some researchers use changed names and addresses to substitute original names and places, such as the COBUILD (Payne, 2014) and the ICE (Nelson, 2014). Likewise, the compilers of the COLT take the number of original syllables into consideration; as a result, they replace names and places with fake names and addresses consisting of the same number of syllables as the originals (Haslerud & Stenström, 2014). In the transcribed version of the COLT, last names, addresses, and telephone numbers have been deleted, while first names are real, i.e. they have not been replaced by fictitious names. The Spoken BNC2014 takes a slightly different approach to deal with the matter of anonymization. That is, it includes the gender of the name, where interpretable. The inclusion of gender was a crude attempt to acknowledge that “names …carry a certain amount of social and ethnic information” (Hasund, 1998, p. 13), which could be retained without compromising anonymity. Regarding spoken L2 corpora, tags such as <first name of interviewee>, <first name and full name of interviewer> or <name of professor> are used to replace names in the Taiwanese component of the LINDSEI (Huang, 2014). In Gablasova, Brezina, McEnery, et al. (2017), the names of non-famous people are replaced completely with the tag <name>.

Drawing on the existing transcription conventions, in this study, any personally identifiable information, including names, addresses, and locations or institutions that seem

unique to the speaker in some way were marked with the tags <name>, <address>, <city>, <university> and the like. The label 'TA' was also used to preserve anonymity of gender. In the recordings, truncated city names, such as 北北京 *Bei Beijing*, were also replaced by the tag <city> if they enabled the individuals to be identified. Two examples are given as follows.

(5) An extract of the speech of a male L2 speaker of Chinese

<L15> 我这个来自那个在新西兰的北岛的西部 er 然后我在那边儿长大然后长大以后就去那个 er **<university>**在**<city>**的那个大学

I zheige come from neige the western side of New Zealand's North Island I grew up there then when grew up went to neige er <university> in <city> that university

*I well come from like the western side of New Zealand's North Island er then I grew up there then when I grew up went to like er <university> in <city> that university*

(6) An extract from the speech of a female L2 speaker of Chinese

<L04> 明白了然后你你怎么认识**<name><name>**老师的

Understood then you you how know <name> <name> teacher

*I got it and you how did you know <name> teacher <name>*

<S00> 哦其实我不认识 **TA**

Oh in fact I do not know TA

*Oh actually I do not know TA*

### 4.5.3 Backchannels

In addition to the above features, another feature deserving of attention is backchannels in interaction. Backchannels are illustrated by White (1989) as follows:

> The term implies that there are two channels in conversation that operate simultaneously. The "main" channel is that through which the speaker (the person holding the floor) sends messages, whereas the "back" channel is that over which the listener (the addressed recipient of talk) gives useful information without claiming the floor. (p. 59)

Backchannels include verbal forms (e.g., *uh huh, yeah*) and nonverbal forms (e.g., head nods). In the case of this study, the term 'backchannels' was limited to verbal forms solely due

to the spoken Chinese corpus being developed as a monomodal corpus. Considering the research purposes outlined previously, the first question regarding backchannels was whether it was necessary to transform them into written records. On one hand, it seemed that there was no need to include the backchannels in the transcripts, because most of them were produced by me, while the primary aim of the present project was to investigate language use of participants, especially that of L2 speakers of Chinese. However, on the other hand, it was problematic to eradicate them, as backchannels have been identified as playing multifunctional roles in spoken discourse (Jefferson, 1984; Kjellmer, 2009; McCarthy, 2002; Peters & Wong, 2015; Schegloff, 1981). Backchannel vocalisations were evident at various points in the interview data, their production perhaps rendered all the more necessary by the video-conferencing format in which the interviews took place. In addition, in the case of the spoken Chinese corpus, which was characterized by informal conversational interactions, the eradication of backchannels in the transcription would impact the naturalness and representativeness of the spoken data considerably. Consequently, it was appropriate to record all the backchannels with written forms.

The second decision made with respect to backchannels in this study was the approach to representing them. Some of them, such as 对 *dui* 'yes' and 行 *xing* 'ok, all right', which could be found in major published dictionaries were transcribed directly in standard Chinese characters. Backchannels then mentioned in the following discussion refer exclusively to those which could not be represented in standard orthography. For those, an alternative employed by a large number of scholars is alphabetic Roman characters (e.g., *Pinyin*) (e.g., Lu et al., 2014; Tseng & Gibbon, 2006; Zeng & Liu, 2002). The use of alphabetic Roman characters to some extent improves the accuracy of transcription while simultaneously introducing a new challenge for the transcribers, i.e. standardization. In this study, the issue of standardization of backchannels was twofold: (i) transcription conventions with reference to backchannels vary considerably in previous studies concerning spoken Chinese, so a better attempt to adopt the most suitable transcription convention for the spoken Chinese corpus had to be made (e.g., use capital or lowercase letters), and (ii) in practice it was difficult to assess which meaningful distinctions were maintained between the sounds of backchannels and their orthographic forms (Love, 2017). For example, in this study, Pinyin *eng* /*ɛŋ*/ was used to capture backchannels which could not be represented in standard orthography but had similar sounds to *eng*. The main function of *eng* in the conversations was to indicate that the hearer was listening to the speaker or agreed with the speaker, e.g., *eng* in (7). By standardising this kind of backchannel,

it thus was impossible to know the distinctions between the sounds of backchannels and the orthographic form *eng* without accessing the original audio recordings. Bearing the research purposes in mind, in this study, it was not necessary to show clearly the distinctions between the sounds and the form *eng*. More importantly, as Andersen (2016) puts it, "[t]he key word in spoken corpus transcription is consistency" (p. 324). The decision to use *eng* to capture the similar backchannel sounds secured a consistent way of handling backchannels which could not be transcribed in standard orthography throughout the spoken Chinese corpus. In other words, this decision to some extent avoided a variety of deviant forms entering the spoken Chinese corpus and maximised transcription consistency.

(7) An extract from the speech of a male L1 speaker of Chinese

<L19> 在国内呢其实这个大学教育 erm 新西兰的这个教育的理念就不一样了因为国家的这个导向不太一样

in China actually university education erm New Zealand's education ideas are different because national policies are different

*in China actually university education erm New Zealand's education is different (from China) because of (they have) different policies*

<S00> **eng eng**

*eng eng*

### 4.5.4 Minimal Response Tokens

In the present study, for convenience I provisionally used the term 'minimal response tokens' to represent features such as *er* and *erm* which were produced by the interviewees in the main channels of communication, due to the complexity of terminology in previous studies (for more information about terminology, see Andersen, 2016; Tottie, 2013). As all the L2 participants involved in this project were native English speakers, it was evident that they used these English tokens frequently. Classifying these sounds seems to require a high level of inference on the part of the transcriber (Atkins et al., 1991; Love et al., 2017). This being the case, to avoid variability in the spoken Chinese corpus, spellings for these vocalisations used in the Spoken BNC2014 (Love, 2017, 2020) were adopted as guidelines in the spoken Chinese corpus transcription (see Table 17). Admittedly, some tokens used by the L1 speakers of Chinese in the conversations sounded quite similar to their English counterparts; therefore, the spellings

presented in Table 17 were also applicable to Chinese minimal response tokens where appropriate. Examples (e.g., *er* and *erm*) can be found in (5) and (7) in 4.5.2 and 4.5.3 respectively.

**Table 17**

*Spellings for Minimal Response Tokens Used in the Spoken BNC2014*

| What it sounds like, and usual meaning | How to write it |
|---|---|
| has the vowel found in 'father' or a similar vowel; usually = realisation, frustration or pain | ah |
| has the vowel found in 'road' or a similar vowel; usually = mild surprise or upset | oh |
| has the vowel in 'bed' or the vowel in 'made' or something similar, without an 'R' or 'M' sound at the end; usually = uncertainty, or 'please say again?' | eh |
| a long or short 'er' or 'uh' vowel, as in 'bird'; there may or may not be an 'R' sound at the end; usually = uncertainty | er |
| as for 'er' but ends as a nasal sound | erm |
| has a nasal 'M' or 'N' sound from start to end; usually = agreement | mm |
| like an 'er' but with a clear 'H' sound at the start; usually = surprise | huh |
| two shortened 'uh' or 'er'-type vowels with an 'H' sound between them, usually = disagreement; OR, a sound like the word 'ahah!'; usually = success or realisation | uhu |

Note. This table was adapted from "The Spoken British National Corpus 2014: Design, compilation and analysis", by Love, R., 2017, unpublished Doctoral dissertation, Lancaster University, p. 95.

### 4.5.5 Non-Native Speaker Features

So far, the features which have been discussed are contained in both the L1–L1 and L1–L2 conversations. In this section, I focus on some features that occurred only in the L2 recordings.

First of all, features that have been discussed previously, such as repairs, repetitions, false starts, pauses, and discourse markers are more likely to be interpreted as 'disfluencies' or 'errors' if they are uttered by L2 speakers (Gilquin & De Cock, 2013; Hasselgren, 2002; Lennon, 1990; Stenström & Svartvik, 1994; Temple, 2000). In their exploration of the pedagogical applications of the TLC, Gablasova et al. (2019a) suggest that features mentioned above can be expected both in L1 and L2 production, and yet some of these features "can have

a different function in learner language" (p. 13). As these features were ubiquitous in both the L1 speech and the L2 speech, they thus were all transcribed orthographically and not treated as errors in the present study.

It should be noted that the decision taken above does not mean that there are no errors in L2 speech. The point here is, as Corder (1967) notes, that "in normal adult speech in our native language we are continually committing errors of one sort or another" (p. 166). Therefore, it was necessary to make a distinction between "acceptable deviance" and "unacceptable deviance" (Stenström & Svartvik, 1994, p. 242) when transcribing L2 speech. In practice, it was noticeable that the L2 participants used incorrect or non-standard tones quite often; however, this rarely impacted the real-time communication. For instance, some L2 participants had strong dialectal accents, in this situation, it was unlikely and unnecessary to identify their pronunciations as errors. Admittedly, sometimes I misunderstood them when L2 speakers produced inaccurate tones, because in Chinese different tones can represent absolutely different meanings[36]. For example in (8) below, the TV series that the L2 speaker of Chinese mentioned was called *Guó Mén Yīng Xióng* (国门英雄 'National Heroes'). Since I was unfamiliar with this TV series, I misunderstood the participant's pronunciation and incorrectly interpreted the information as a different meaningless expression. Since this misunderstanding caused communicative issues, it thus was essential to categorize features of this kind as pronunciation errors. Consequently, the term 'error' used in the present study referred only to those utterances which were misunderstood by me. This definition seemed quite subjective but was feasible to be identified on the basis of the context.

(8) An extract from the speech of a male L2 speaker of Chinese

<L01> er 因为这个这个名字是你你看过 er 中国的一个电视剧叫 er guò 门英雄

er because this this name is you have you watched er Chinese a TV series called er *Guò Men Heroes*

*er because this this name is you have you watched er a Chinese TV series called er Guò Men Heroes*

<S00> 什么英雄

what heroes

[36] In Chinese there are five tones: flat, rising, falling-rising, falling, and a neutral tone. Different tones indicate different Chinese characters which represent different meanings. For example, hànyǔ 汉语 'Chinese' and hányǔ 韩语 'Korean'. On the other hand, same tones can sometimes represent different meanings.

*what heroes*

<L01> 叫 guò 门英雄

called *Guò Men Heroes*

(it is) called *Guò Men Heroes*

<S00> 过门英雄没有看过

*Guò Men Heroes* have not watched

*I have not watched Guò Men Heroes*

In addition, there were possibilities that there were some lexical and grammar 'errors' (which were usually identified based on written language in previous studies) in the L2 speech. Identifying lexical or grammar errors however seemed not to be a possibility to me, because the L2 corpus developed in the present study was not expected to support error analysis. What should be noted is that the term 'error' did not indicate that the identification of pronunciation errors aimed to support error analysis as well. Rather, it only meant that a set of pronunciation features in the L2 speech were identified and categorized as errors.

## 4.6 Consistency and Reliability of Transcription

In terms of transcription, some researchers have been concerned with the difficulty in replicating the interlocutors' experience in spoken discourse (e.g., Cook, 2014; O'Connell & Kowal, 1999), and have argued that "repeated listening on the part of the transcriber cannot incrementally approximate the experience of the original participants" (O'Connell & Kowal, 1999, p. 111). In this study, I engaged in all the conversations and then transcribed all the recordings, thus any differences "between transcriber's and participants' perceptions" (Cook, 2014, p. 38) can be largely mitigated. Also, in order to maximise the consistency and reliability of the transcription, I made several successive passes through the sound files. As described in Section 4.3, the first transcriptions of L1 and L2 speech were carried out for me to reach an in-depth understanding of informal spoken Chinese. Throughout this process, I made notes relating to issues observed on the sound files. Then I revisited the recordings to fill in the details and revised mistakes in the transcripts. Therefore, these passes to a large extent minimised the variability of mistakes which may occur in the transcripts. It is noticeable that some studies contain an appendix with simple notations of the transcription conventions without transcribed examples, or some contain a limited number of transcribed examples taken from a corpus rather

than the entire corpus. As a result, it is difficult to evaluate the reliability of the transcripts used in those studies. Conversely, the spoken Chinese corpus associated with the transcription scheme will be open to the public, which enables corpus users to evaluate the credibility of the transcripts.

## 4.7 Summary

Following the transcription guidelines of some existing spoken corpora, this chapter provides a transparent description of the transcription of the spoken Chinese corpus and creates a bespoke transcription scheme. Given that transcription has long received very limited attention in the Chinese research literature, it is expected that this chapter can raise awareness of the important role of transcription in spoken corpus design and building.

In practice, it is not feasible to capture the full complexity of spoken language in the transcripts. Bearing the research purposes in mind, I outlined the features that should be preserved in the transcription, and discussed the representations of these features. There were many features in the recordings which were common both in the L1 and L2 conversations, this chapter thus explained these common features together. Also, these features in practice were transcribed based on the same conventions. This ensures that any observed differences in L1 and L2 use are not actually a result of different transcription conventions employed in each group. In addition, it also placed some emphasis on the features that occurred in the L2 conversations. At the end of this chapter, I gave a brief account of the actions taken to maximise the reliability of the transcription.

This chapter along with Chapter 3 provide transparent procedures in terms of spoken corpus compilation, including spoken corpus design, data collection, and transcription. They now have the ability to answer the first research question proposed in Chapter 1: what are the main considerations in building a spoken Chinese corpus of informal interaction? In the following chapter, I will move on to discuss the comparability of the L1 corpus and the L2 corpus.

# Chapter 5 Comparing the L1 and L2 Corpus

## 5.1 Introduction

Chapters 3 and 4 have clarified some important steps in terms of the compilation of the spoken Chinese corpus, including corpus design, data collection, and transcription. Based on the work conducted in the preceding chapters, both the L1 and L2 corpora represent informal interaction in *Putonghua* used in mainland China, and the data were gathered by the researcher adopting the unstructured interviewing method. Table 18 shows that the target size of the spoken Chinese corpus[37] has been successfully achieved, approximately 33.8 hours of informal conversations have been transcribed and contained in this corpus. It also is noticeable that the L1 corpus and the L2 corpus are similar in size. However, the two corpora are not identical in composition (e.g., the L1 corpus includes more speakers than the L2 corpus), although they were designed and created as comparable corpora. Given that no corpora are alike even with the same design criteria (see Chapter 2), this chapter aims to investigate to what extent the spoken L1 corpus is comparable to the L2 corpus. In doing this, it is expected that users can approach the data with greater caution, taking into consideration the differences that were caused by the decisions and compromises made during data collection when interpreting the observed differences between L1 and L2 production. This chapter also considers the interspeaker variation within each corpus, seeking to provide some useful insights into the appropriate use of the data in the L1–L2 comparative study on the L2 use of the marker 就是 *jiushi* in Chapter 6.

**Table 18**

*Components of the Spoken Chinese Corpus*

| The Spoken Chinese Corpus | Size (words) | No. of speakers | No. of recordings | Length (minutes) |
|---|---|---|---|---|
| The L1 corpus | 228,306 | 22 | 26 | 910 |
| The L2 corpus | 220,792 | 14 | 34 | 1,119 |
| Total | 449,098 | 36 | 60 | 2,029 |

[37] In fact, 1,293 minutes of L2 recordings and 1,012 minutes of L1 recordings have been gathered for the current study. However, some recordings, due to the reasons discussed in Chapter 3, have not been used in this version of the spoken Chinese corpus.

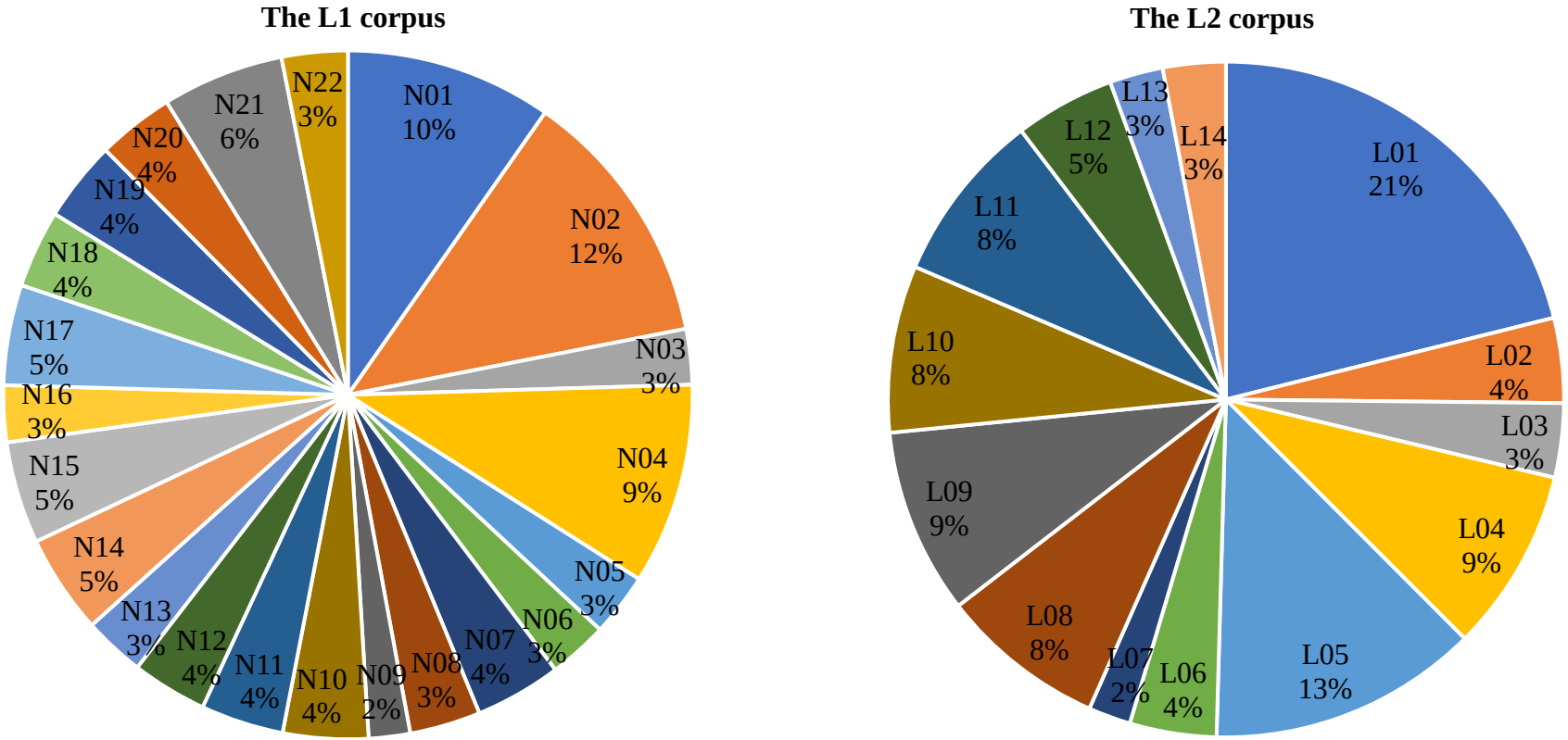
To begin with, Section 5.2 introduces the sample texts of each corpus, and gives an account of the similarities and differences between the L1 and L2 corpus. Then I consider the disparity in my contributions to the two corpora and address the interspeaker differences in each corpus in Section 5.3. Given that the L1 corpus primarily comprises informal interactions between friends and acquaintances, and the L2 corpus contains interactions between strangers, Section 5.4 is concerned with the potential effect of this design on subsequent analyses. This is followed in Section 5.5 by a consideration of the observed variation in terms of factors pertinent to the speakers, such as being native or non-native and gender. In Section 5.6, I introduce briefly how to access the first version of the spoken Chinese corpus. Finally, this chapter closes with a summary in Section 5.7. By doing this, it argues that care should be taken when generalising findings on the basis of the two corpora.

## 5.2 Sample Texts

Figure 11 presents visually each participant's contribution to the L1 corpus according to the total number of words they uttered in interaction (including the researcher's speech). Baker (2006) advises that if the participants have the same chance to attend the interview and produce equally sized samples, it is more likely to be able to claim that the corpus is representative of the linguistic production of the cohort of the group. In this study, it should be noted, however, that the proportion of each participant's speech in the L1 corpus is subject to variation. Three speakers, N01, N02, and N04, participated in two or three conversations which account for 32% of the L1 corpus. The imbalanced contributions run the risk that the corpus results would be biased if these speakers strongly prefer certain linguistic patterns (e.g., the item 就是 *jiushi*), because the results would more likely represent the language use of these speakers rather than the whole group (see Section 5.4 for further discussion). Likewise, there is considerable variation in the individual contributions in the L2 corpus, mainly caused by the exigencies of sampling and availability. However, the L2 corpus exhibits a greater degree of imbalance than the L1 corpus. Figure 11 shows clearly that speaker L01's output accounts for 21 % of the whole L2 corpus, which is about 10 times larger than the proportion of speaker L07's conversation. As a result, there is a risk that speaker L01's talk could have too great an influence on the whole group when using this L2 corpus to analyse L2 use.

**Figure 11**

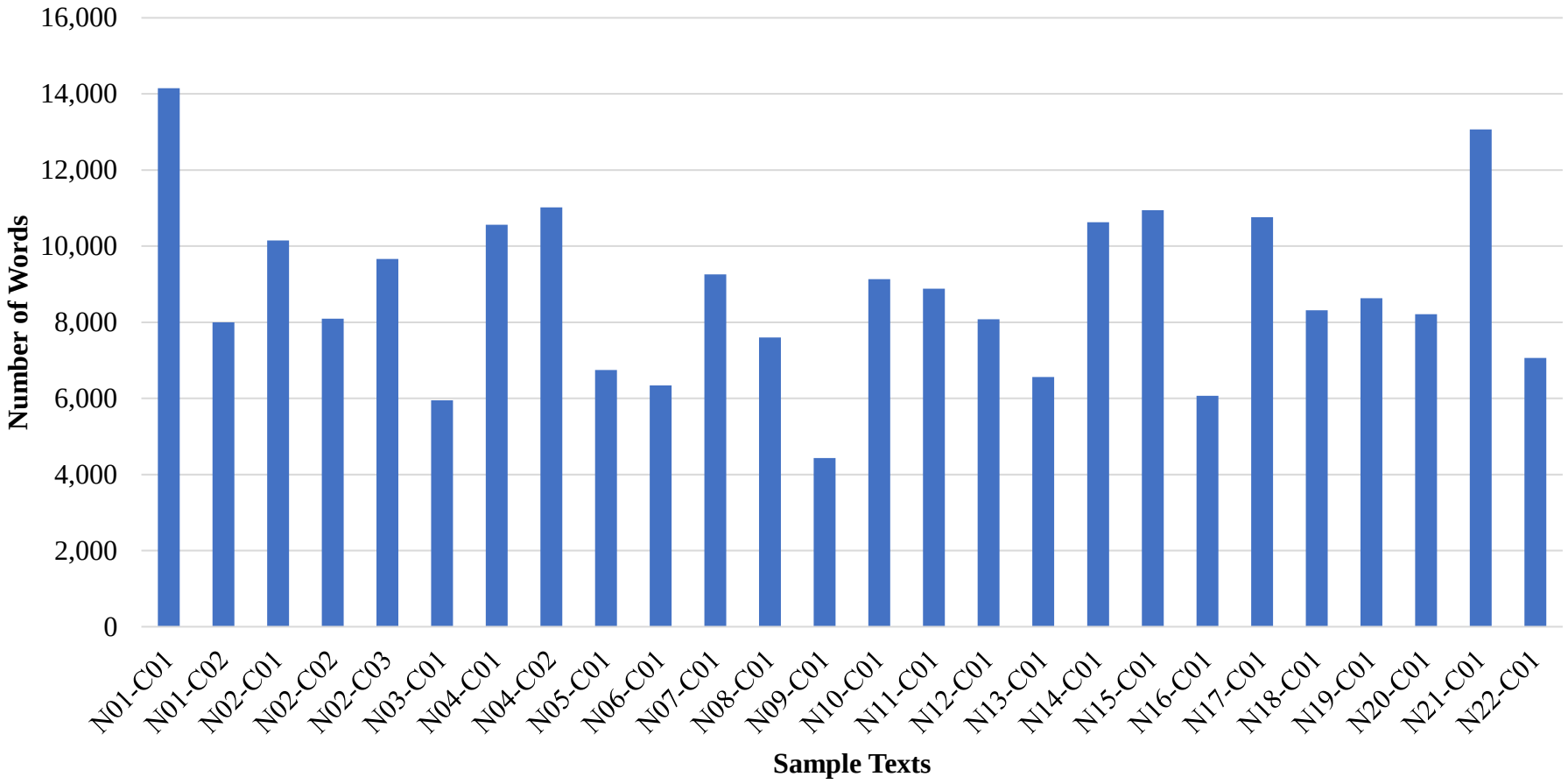
*Proportions of Speakers' Contributions in the L1 Corpus and the L2 Corpus*




*Note*. Each section represents the proportion of the interaction between the researcher and one participant.

**Figure 12**

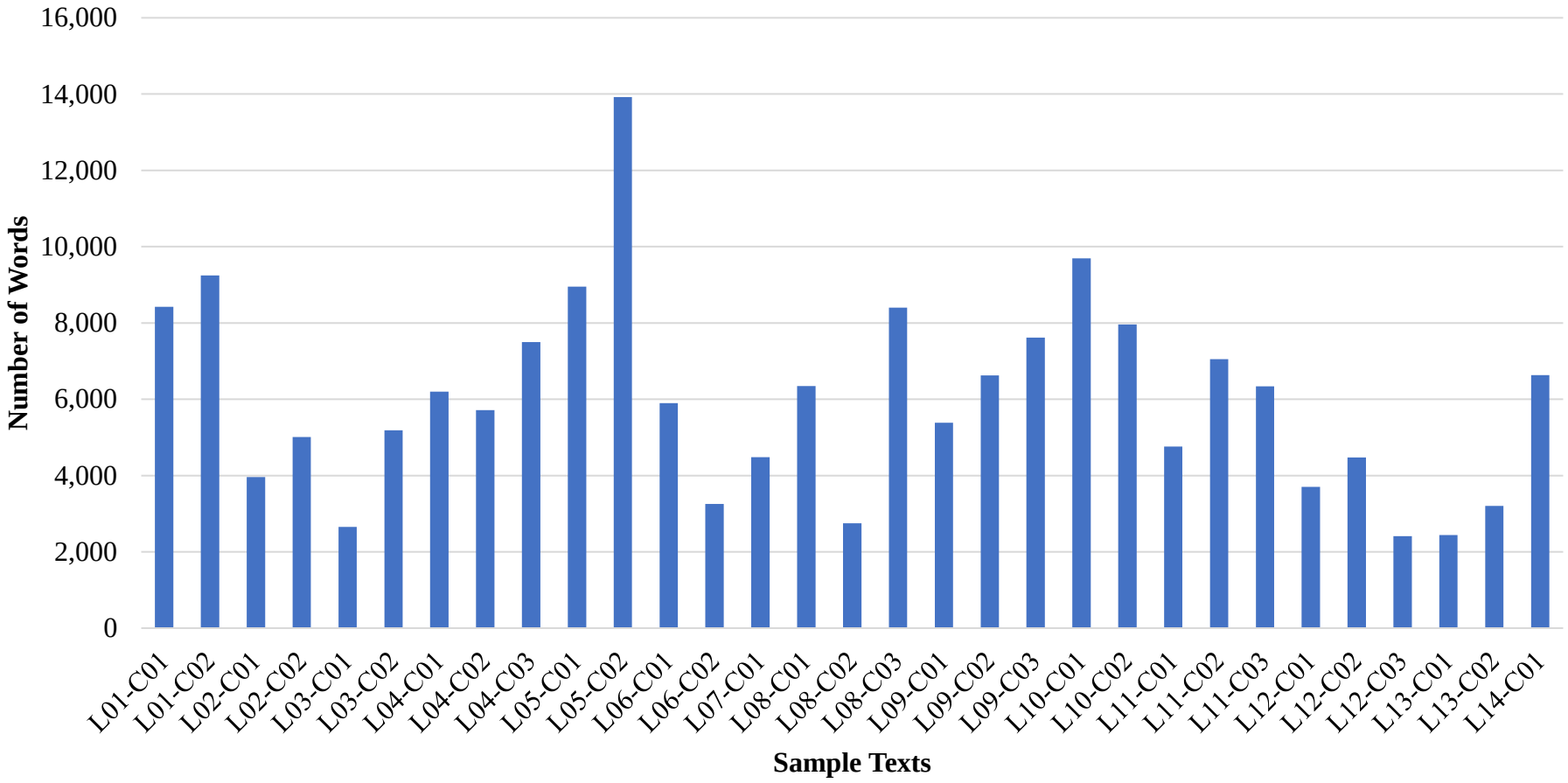
*Sample Sizes in the Spoken L1 Corpus*




*Note*. For detailed information of the number of tokens in each text, see Table I-2 in Appendix I.

**Figure 13**

*Sample Sizes in the Spoken L2 Corpus*




*Note*. For detailed information of the number of tokens in each text, see Table J-2 in Appendix J.

Just by visually inspecting Figures 12 and 13, we can easily see that the sample texts differ substantially in size in the two corpora, which is one primary reason for the diverse proportions of individual outputs. Including texts which contain significantly more words than the average in a small corpus could exert an undue influence on the results of queries (Sinclair, 2005). It thus is understandable that the first computer corpus, the Brown Corpus, is designed to contain 500 2,000-word sample texts (Francis, 1980). This model has been followed by a string of successors, notably the LOB, the Freiburg-LOB Corpus of British English (the F-LOB)[38], and the Freiburg-Brown Corpus of American English (the Frown)[39]. More recently, the London-Lund Corpus 2 (LLC-2, Pöldvere, 2019; Pöldvere et al., 2019), which was designed to be comparable to the LLC was created by employing a similar method to sample data: it comprises approximately 500,000 words, stored in 100 texts of 5,000 words each. Pöldvere et al. (2019) point out that "[o]ne text in the corpus is equivalent to either one single recording or multiple shorter recordings revolving around a similar subject matter and/or involving the same (one) speaker. Where possible, the recordings were transcribed in full; however, most of the texts in the corpus represent an excerpt from a recording" (p. 10). It has been argued that it is an unsafe assumption that any part of a conversation is representative of the whole, as "the result of research for decades of discourse and text analysis make it plain that position in a communicative event affects the local choices" (Sinclair, 2005, p. 7). Also, in corpus practices (e.g., the Spoken BNC2014), it is acceptable that conversations vary in length, for it reflects the nature of real conversations. Consistent with existing spoken corpora, both the L1 and L2 corpora consist of entire transcriptions of complete speech events.

In addition to the above features regarding the sample texts which are shared by the two corpora, it is noticeable that the L2 corpus contains more transcribed texts than the L1 corpus, while the L1 corpus features more speakers than the L2 corpus. These differences are the ramifications of the decisions and compromises made at the stages of corpus design and data collection which have been discussed in Chapter 3. To sum up, both the L1 and L2 corpora raise the matter of balance which bears on the representativeness of the two corpora due to the fact that limited data are gathered. It is therefore expected to enlarge both the L1 and L2 corpora in the future to achieve representativeness and balance as far as possible. As regards the matter of representativeness, it is also worth considering the researcher's production in the spoken

[38] The FLOB: http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/.
[39] The Frown: http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/.

Chinese corpus, to examine to what extent the researcher's presence affects both the L1 and L2 participants' outputs (see Sections 5.3 and 5.4).

## 5.3 Proportions of Participants' and the Researcher's Contributions

As has been noted in Chapter 3, I conducted all the interviews. Table 19 shows my contribution to the conversations in both the L1 and L2 corpora. It seems that in the L2 conversations I took a more active role, since the total number of my utterances is higher than that of my speech in the L1 corpus, although in each corpus the data were produced mostly by the participants.

**Table 19**

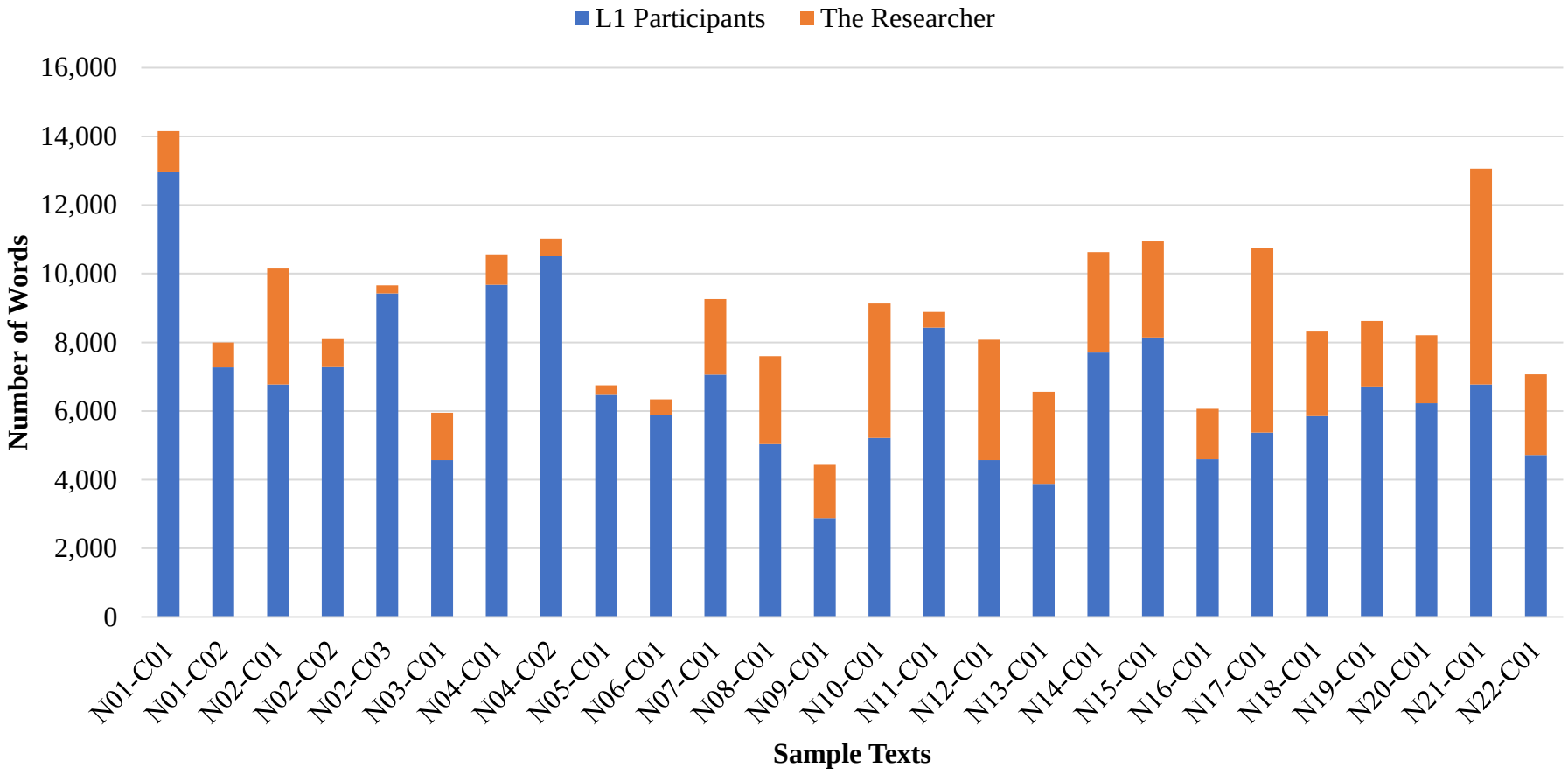
*The Number of Words in Each Corpus*

| The spoken Chinese corpus | Corpus size | Participant's speech | Researcher's speech |
|---|---|---|---|
| The L1 corpus | 228,306 words | 174,009 words | 54,297 words |
| The L2 corpus | 220,792 words | 146,598 words | 74,194 words |
| Total | 449,098 words | 320,607 words | 128,491 words |

Figures 14 and 15 show that the role of the researcher played out in different ways in different interviews. Correspondingly, the L1 and L2 participants enacted their roles differently in interactions in which the researcher made mainly brief responses compared with interviews in which the researcher spoke more freely. Thus, some interactions are highly monologic, while others are more dialogic in the corpus. In addition, according to Figures 14 and 15, it is evident that there are five conversations (one L1 and four L2 conversations) in which the researcher contributed more than the participants (see Appendices I and J for detailed information). The higher degree of the interactivity between the researcher and the participants presumably has something to do with the topics or the (both L1 and L2) participants' willingness to prompt the researcher to share her opinions during the conversations. When comparing the two corpora, it is difficult to measure or decide to what extent the L1 and L2 corpora differ in terms of the interactivity between the researcher and the participants. Nevertheless, it is worthwhile bearing in the mind that the researcher's role and the interactivity between interlocutors may affect the occurrences of certain linguistic patterns.

**Figure 14**

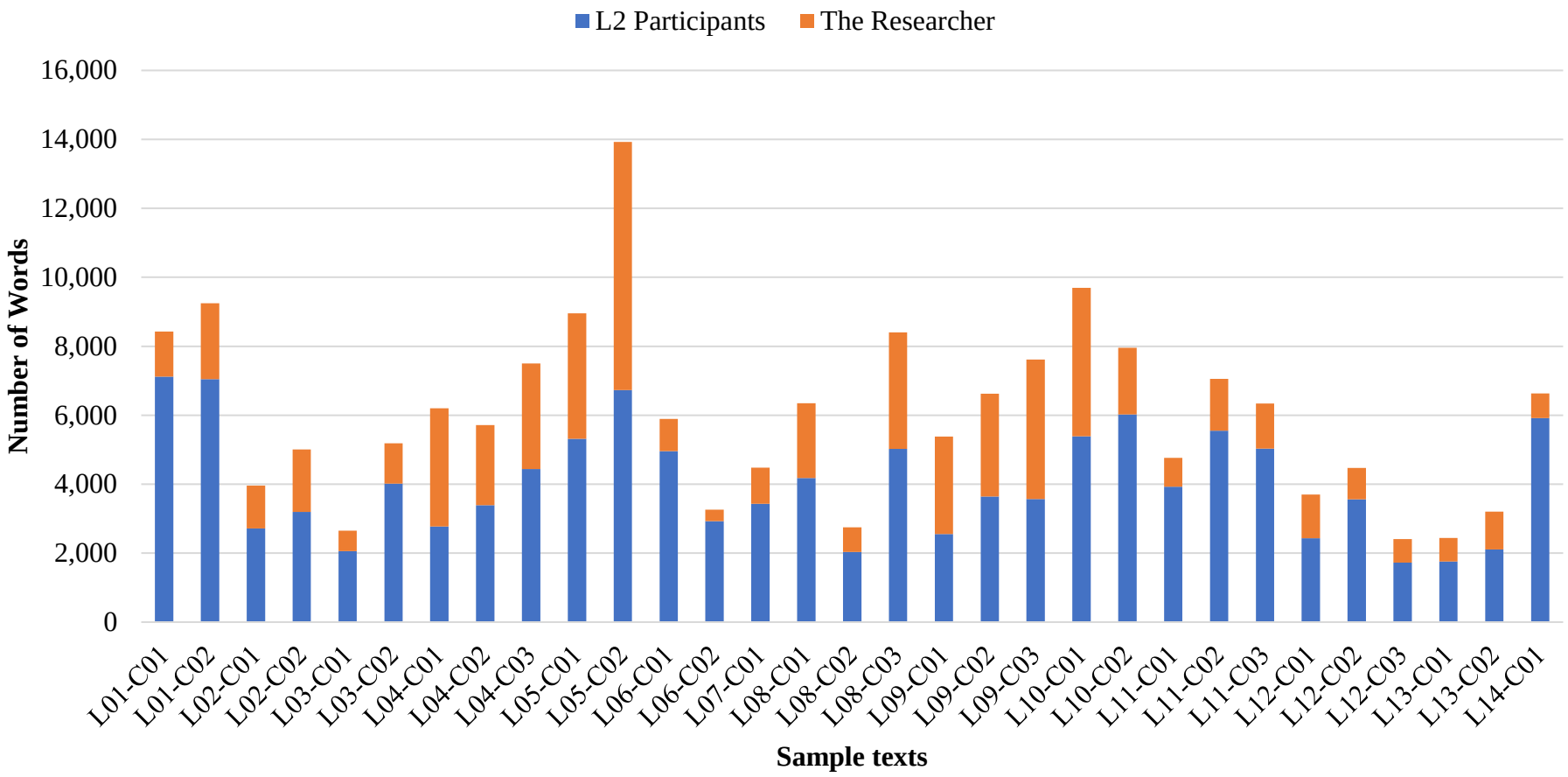
*Individual Differences in the Researcher's Contributions in the L1 Corpus*




*Note*. For detailed information of the number of tokens produced by the researcher and each L1 participant, see Table I-2 in Appendix I.

**Figure 15**

*Individual Differences in the Researcher's Contributions in the L2 Corpus*




*Note*. For detailed information of the number of tokens produced by the researcher and each L2 participant, see Table J-2 in Appendix J.

## 5.4 L1–L1 Interaction vs L1–L2 Interaction

One question that may be asked by SLA researchers or language teachers is whether or to what extent the researcher's speech impacts the L2 participants' use of certain patterns. The problem of this question is that it seems to ignore the possibility that the L1 speakers' usages may be influenced by the researcher, or that the L2 speakers' output may affect the researcher's speaking style or language use[40]. According to Giles and Smith (1979), when two people meet, there is a tendency for them to converge their speech style to become more alike that of those with whom they are interacting. This being the case, this section aims to investigate whether the design of L1–L1 interactions between friends and acquaintances for the L1 corpus and that of L1–L2 interactions between strangers for the L2 corpus affect the results of the differences between the L1 and L2 use of 就是 *jiushi* in Chapter 6. In what follows, I will discuss whether the frequency of the researcher's use of 就是 *jiushi* affected the L1 and L2 participants' use in each interaction.

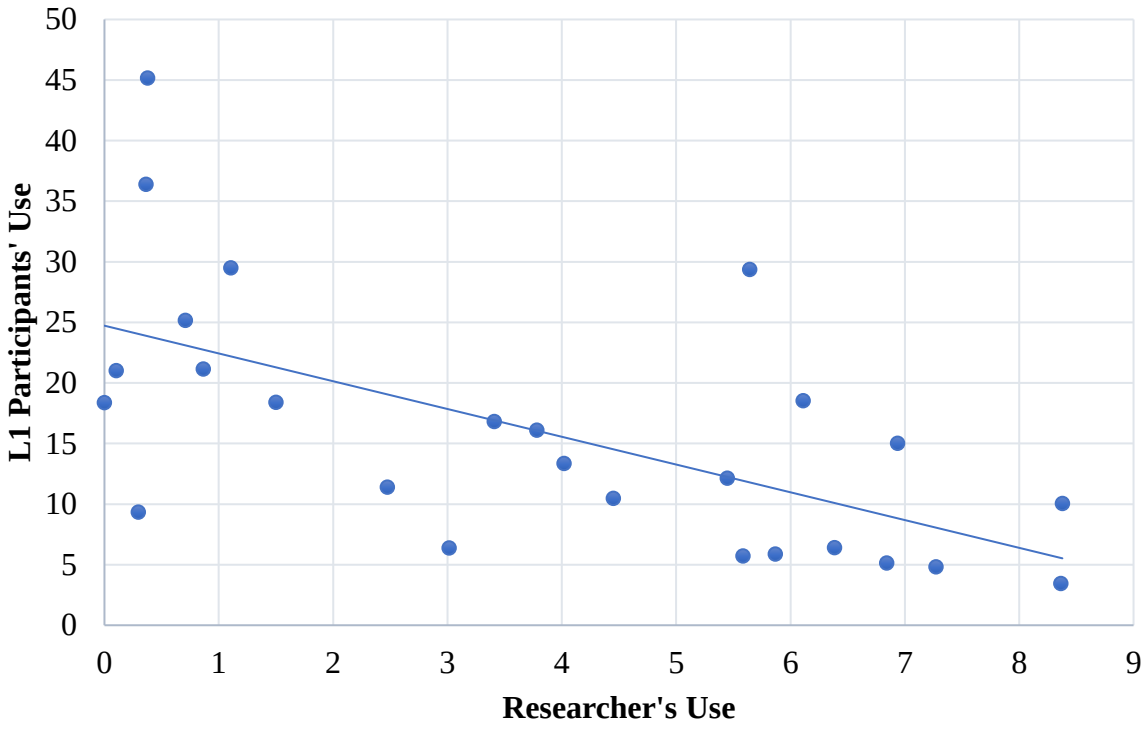
The relationship between the variable of the researcher's use and the variable of the participants' use of 就是 *jiushi* can be measured by adopting a technique called correlation. Correlation measures whether two variables "are related by looking at the extent to which they covary" (Brezina, 2018, p. 141): that is, we are looking at whether, if the value of one variable increases, the value of the other variable increases, decreases or stays the same. The strength of the relationship between two variables can be expressed numerically by using a correlation coefficient (Oakes, 1998). One basic kind of correlation is Pearson's correlation ($r$): the correlation coefficient always ranges from -1 to 1; in general, a negative number indicates a negative correlation (a value of -1 is obtained for a perfect negative correlation), a positive number shows a positive correlation (a value of 1 is obtained for a perfect positive correlation), and a value of 0 represents that there is no linear relationship between the two variables at all (Brezina, 2018; Oakes, 1998). In addition, the correlation coefficient should be complemented with a $p$-value to indicate whether there is enough evidence in the corpus to generalise the correlation to the population. Since there are many software tools (e.g., SPSS) that can be used to calculate correlation, mathematical details will not be introduced in this chapter (for detailed explanations on correlation, see Brezina, 2018).

[40] In my case, I tended to use simpler expressions to communicate with two L2 speakers when I realised that they had relatively limited vocabulary. Thus, this experience may support my opinion that on some occasions L1 speakers' speaking styles or language use may be influenced by L2 speakers in interactions.

To discuss whether the frequency of the use of 就是 *jiushi* by the researcher in each conversation affected the L1 participants' use, the relative frequencies of the two variables—the frequency of the L1 participants' use on the y-axis and the researcher's use on the x-axis are plotted in Figure 16 (for an explanation of relative frequency, see 5.5.1; detailed relative frequencies are given in Table K-2 in Appendix K). Each spot in the scatterplot represents an L1 sample text. It should be noted that the straight line in the graph is called the regression line or the line of the best fit (Brezina, 2018; Oakes, 1998), which marks the relationship between the researcher's use and the L1 participants' use of 就是 *jiushi*. The tighter the points cluster the regression line, the stronger the relationship between the two variables. When the regression line moves down from top left to bottom right, as Figure 16 shows, it means that there is a negative relationship between the researcher's use and the L1 participants' use ($r$ = -.615, p< 0.01): the researcher uses 就是 *jiushi* more frequently, the less 就是 *jiushi* occurs in the L1 participants' speech and vice versa.

**Figure 16**

*Correlation Between the Researcher's and the L1 Participants' Use of the Item 就是 Jiushi*[41]

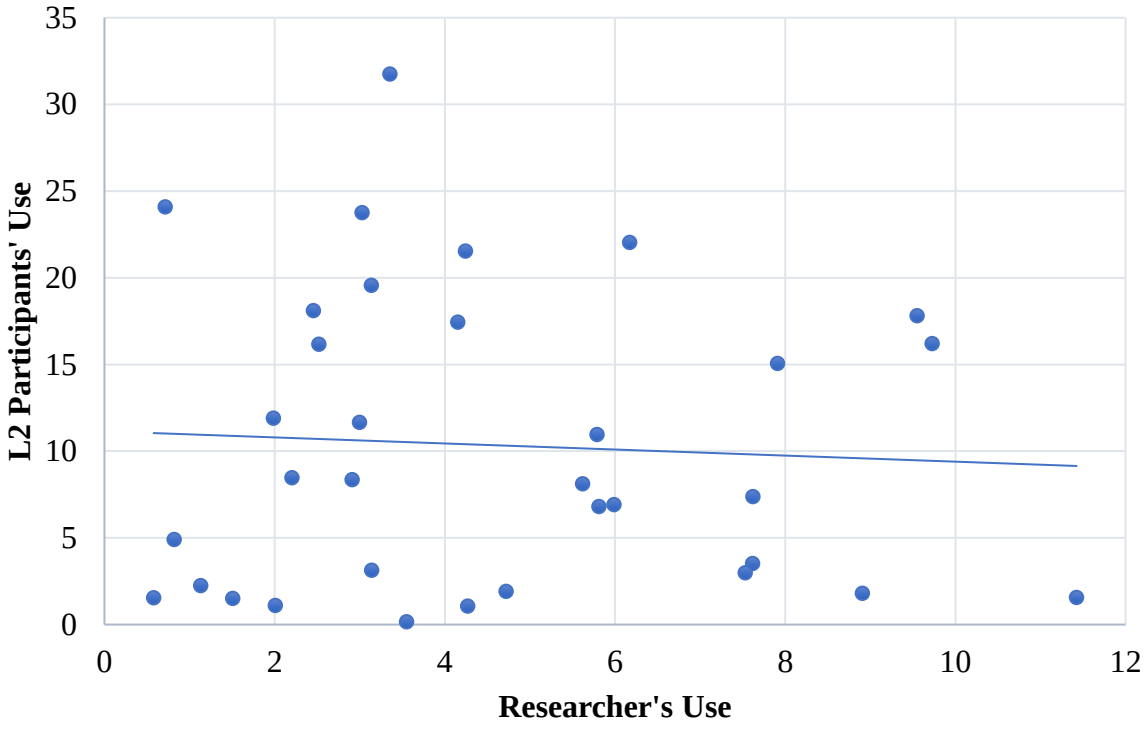



[41] This graph uses the relative frequencies of 就是 *jiushi* to measure correlation. Pearson's correlation ($r$) is used in this section and is calculated by using SPSS.

Similarly, the relative frequencies of two variables–the frequency of the L2 participants' use of 就是 *jiushi* on the y-axis and the researcher's use on the x-axis are plotted in Figure 17 to show the relationship between the researcher's use and the L2 participants' use of the item 就是 *jiushi* (see Table L-2 in Appendix L for the detailed relative frequency information). Each spot in the scatterplot represents an L2 sample text. We can see that the data points are scattered in an apparently random way around the horizontal regression line ($r$ = -.059, p> 0.5), which indicates that there is little or no relationship, or in statistical terms no correlation between the researcher's use and the L2 participants' use. In some L2 corpus studies on the use of individual discourse markers, L1 interviewers' speech is generally excluded from the analyses and all attention is paid to L2 interviewees' speech (e.g., Aijmer, 2011; Buysse, 2017). It however is unclear whether in those studies the interviewers' or researcher's influences have been carefully examined. According to the L2 data in this study, it seems that the researcher's use of 就是 *jiushi* does not affect the use of 就是 *jiushi* by the L2 participants, therefore, it is appropriate that the corpus analysis of the marker 就是 *jiushi* in Chapter 6 focuses only on the L2 participants' speech.

**Figure 17**

*Correlation Between the Researcher's and the L2 Participants' Use of the Item 就是 Jiushi*



In conclusion, the researcher's use of 就是 *jiushi* has different influences on the outputs of the L1 and L2 participants. However, due to the design of the two corpora, it is hard to know whether this difference between the L1–L1 and L1–L2 interactions is caused by the status of being native or non-native speakers or the different relationships between interlocutors. In the following section, I will give a brief discussion of the effect of being a native or non-native speaker.

## 5.5 Speakers in the Spoken Chinese Corpus

This section primarily discusses three factors in terms of the L1 and L2 participants that may have potential influences on language use: being native or non-native, gender, and age. For L2 speakers, there are many factors that may affect their use of certain linguistic patterns, such as their exposure to Chinese, and their proficiency level. Native or non-native status, as has been discussed in Chapter 2, is the fundamental concern in virtually any L1–L2 contrastive studies on L2 use.

### 5.5.1 L1 vs L2 Speakers

To discuss the effect of being a native or non-native speaker on language use, I will first give a brief account of the notion of relative frequency. Absolute/raw frequency is the most straightforward statistic used in corpus linguistics, which is simply the actual count of the occurrences of a particular word in a corpus (Brezina, 2018; Leech, 2011). Absolute frequency of words is a useful measure when we look at a single corpus. For example, I used absolute frequency to produce Table K-1 in Appendix K, in which the number of the occurrences of the item 就是 *jiushi* in each L1 text was given. According to Table K-1, it is easy to have a basic idea that some L1 speakers have a preference to using this item. However, absolute frequency is of little use when we want to compare two or more corpora. For instance, according to the data provided in Appendices K and L, 就是 *jiushi* occurred 3,837 times in the L1 corpus, while it only occurred 2,481 times in the L2 corpus. However, it is problematic to claim that L2 speakers appear to use 就是 *jiushi* less than L1 speakers in conversation, as the L1 corpus and the L2 corpus are of different sizes. As a general rule, relative/normalised frequency is essential

when we are to make comparisons between corpora (or texts) of different sizes. Relative frequency is calculated as follows:

$$relative\ frequency = \frac{\text{absolute frequency}}{\text{number of tokens in corpus}} \times basis\ for\ normalisation \qquad (1)$$

For example, the relative frequencies of 就是 *jiushi* in the L1 corpus and the L2 corpus (the L1 and L2 participants' uses only) can be calculated respectively as follows:

$$relative\ frequency\ (\text{就是}) = \frac{3{,}837}{174{,}009} \times 1{,}000 = 22.05 \qquad (2)$$

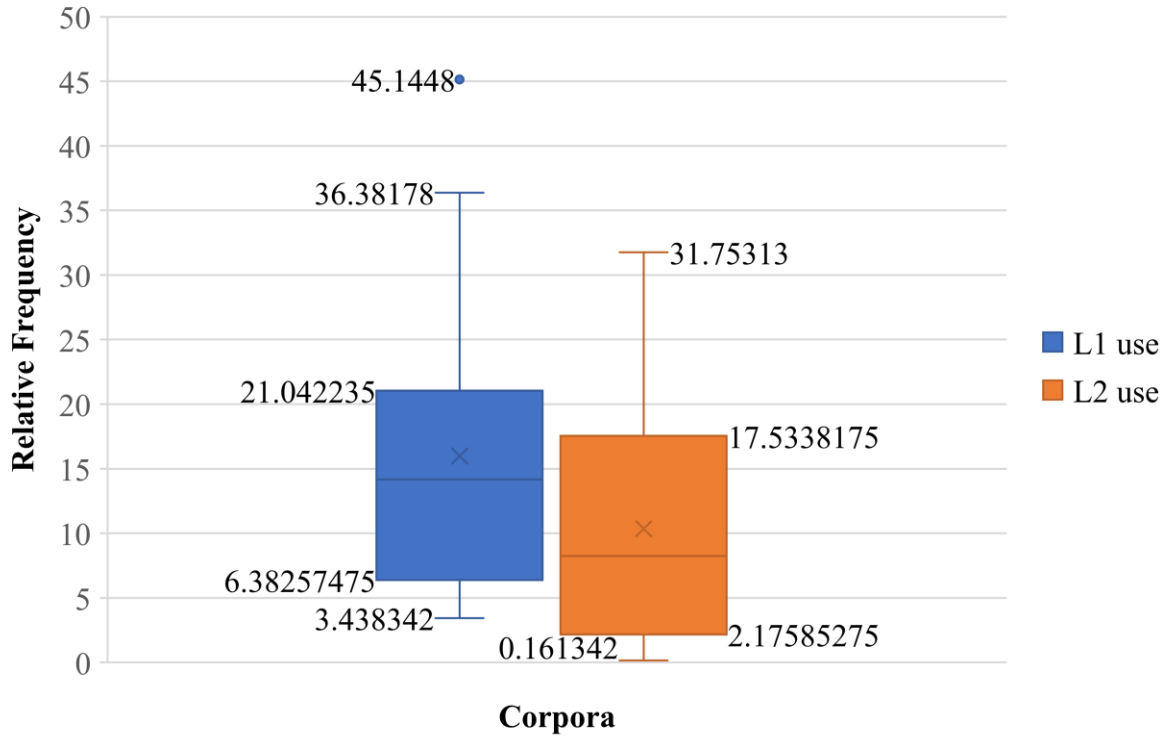
$$relative\ frequency\ (\text{就是}) = \frac{2{,}481}{146{,}598} \times 1{,}000 = 16.92 \qquad (3)$$

In this study I chose 1,000 words as the basis for normalisation. Equation (2) is the relative frequency of 就是 *jiushi* in the L1 corpus, which indicates that on average, there are about 22 instances of 就是 *jiushi* for every 1,000 tokens in the L1 corpus. Equation (3) shows that, on average, there are about 17 instances of 就是 *jiushi* for every 1,000 tokens in the L2 corpus. With the relative frequencies, it may be easily concluded that the L2 speakers appear to use 就是 *jiushi* less than the L1 speakers. This aggregate data methodology, as has been discussed in Chapter 2, runs the risk of ignoring intra-group variation: this method cannot tell us whether the result of the L1 speakers using 就是 *jiushi* more frequently is caused by some L1 speakers who strongly prefer 就是 *jiushi*, or if the result represents performance of the whole group. Therefore, it is imperative to use some statistical measures to represent within group variation.

The boxplot in Figure 18 shows the distribution of the item 就是 *jiushi* in each corpus. In the graph, the inside boxes represent the interquartile ranges, which are the intervals between the lower and the upper quartiles. For instance, in terms of the L2 data, the interquartile range is the interval between the lower quartile 2.17585275 and the upper quartile 17.5338175. The interquartile range can display the spread of the values of a variable in a corpus. As Figure 18 shows visually, 就是 *jiushi* is dispersed in a similar way in the L1 and L2 data (an alternative dispersion measure will be introduced in Chapter 6). It is also observable that L2 speakers of Chinese appear to use 就是 *jiushi* less often that L1 speakers.

**Figure 18**

*Distribution of the Item 就是 Jiushi in the Two Corpora*




*Note*. This graph was created based on the relative frequencies of 就是 *jiushi* given in Tables K-2 and L-2 in Appendices K and L respectively. In this graph, the horizontal lines inside the boxes represent the medians, the means are labelled with × inside the boxes, and the 'whiskers' above and below the boxes show the minimum and maximum values. For example, the minimum and maximum value of the L1 use of 就是 *jiushi* are 3.438342 and 36.38178 respectively, representing that 就是 *jiushi* occurs at least 3.438342 times per 1,000 words and appears in 36.38178 instances per 1,000 words of speech at maximum.

In particular, there is one isolated dot which stands out in the L1 group, representing a speaker who exhibits a relative frequency for 就是 *jiushi* of 45.1148 instances per 1000 words of speech. Outliers are problematic for many statistical analyses, since they may obscure the general tendency in the data or cause the test to miss significant findings. On the other hand,

outliers can be informative about the data collection process, and it is important to understand how outliers occur and whether they are a natural part of the target language phenomenon under examination. In this study, it seems that the outlier in the L1 group indicates that this L1 speaker strongly prefers using 就是 *jiushi* in conversation. Since the L1 corpus is small in size, this feature of the speaker's preference shows up clearly. When comparing the frequencies of the use of the marker 就是 *jiushi* in Chapter 6, it should be borne in mind that the use of 就是 *jiushi* to some extent reflects speakers' personal styles and it is not necessarily the case that all L2 speakers should be recommended to use it as often as the L1 speakers in conversations.

### 5.5.2 Gender and Age

As we can see clearly in Table 20 below, neither corpus is gender-balanced, and the difference in terms of gender between the two corpora is considerable (see Figure 19): the L1 corpus contains more same-sex conversations as more female L1 participants took part in the research, while the majority of the conversations in the L2 corpus take place in a mixed-sex environment (sample sizes are given in Appendices I and J). Target proportions for the gender groups were not fixed during the process of data collection; as a result, each gender group in the two corpora consists of different number of speakers and the proportions are varied. This being the case, when comparing the frequency information of the item 就是 *jiushi*, the L1 corpus tends to represent female L1 speakers' performance, while the L2 corpus is more likely to represent the male L2 speakers' usages. The problem is that it is difficult to examine systematically whether the gender factor may impact the output of the item 就是 *jiushi* by making use of this spoken Chinese corpus, as only a small number of speakers were involved in this study. Consequently, when comparing the L2 use of 就是 *jiushi* with the L1 use in Chapter 6, gender will not be considered.

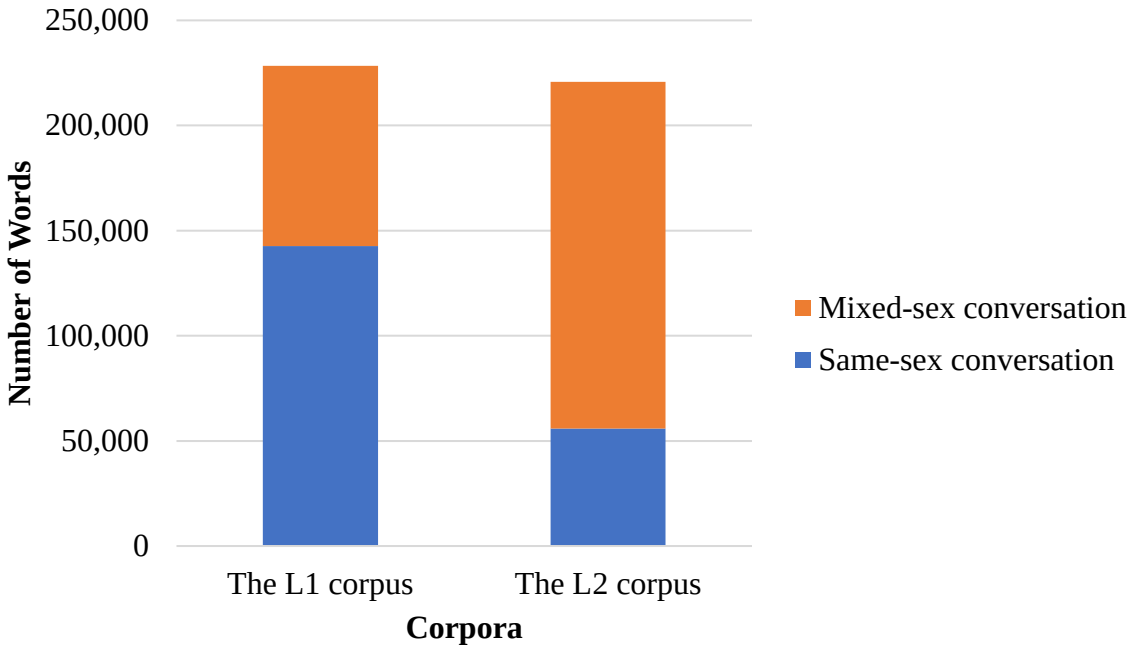
**Table 20**

*The Number of Speakers in the Spoken Chinese Corpus*

| The spoken Chinese corpus | No. of female speakers | No. of male speakers |
|---|---|---|
| The L1 corpus | 12 | 8 |
| The L2 corpus | 3 | 11 |

**Figure 19**

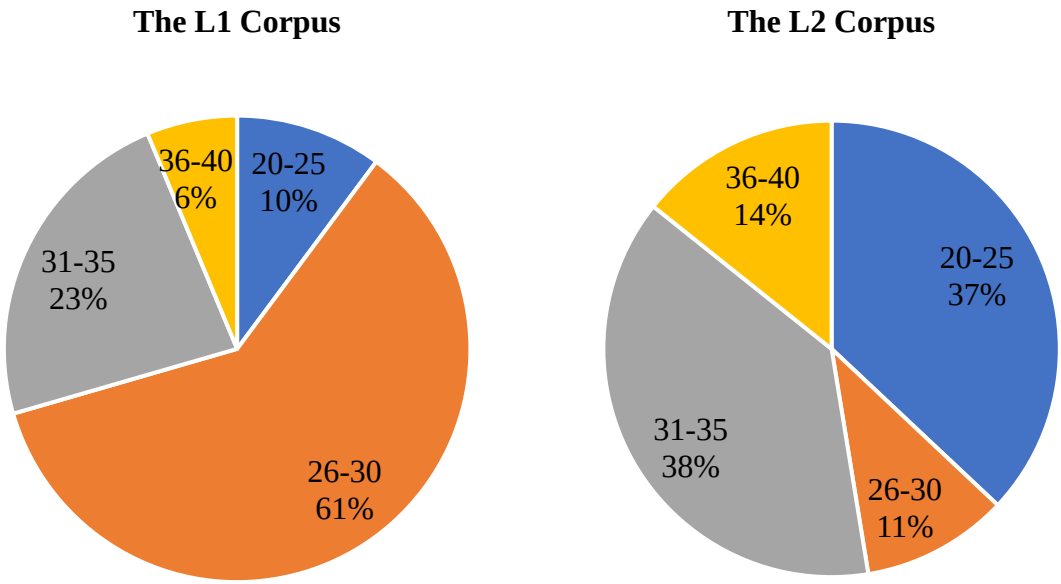
*The Difference in terms of Gender Between the L1 Corpus and the L2 Corpus*




As for the factor of age, Figure 20 below shows that the majority of conversations in the L1 corpus belong to the 26-30 age group, as the majority of L1 speakers were aged between 26-30 (see Chapter 3). In the L2 corpus, the 20-25 age group and the 31-35 age group contribute roughly equal quantities of talk. The large proportion of the 31-35 age group is caused by the production of the speaker L01 mentioned in Section 5.2, which means that the speech with regard to this group is biased towards this speaker rather than representing the performance of the whole age group. In the analysis of the use of 就是 *jiushi*, this age factor will also be ignored in this study. Nonetheless, it does not mean that the age factor has no impact on the L2 use of 就是 *jiushi*. Rather, it seems that there is no good reason to claim that the difference in gender and age between the two corpora should be automatically assumed to have contributed to the difference between L1 and L2 use. Investigating the potential influences of gender and age on the use of markers such as 就是 *jiushi* will be a goal for the future with the process of enlarging the spoken Chinese corpus.

**Figure 20**

*Proportion of Age Groups in Each Corpus*




With the above discussions, the factor of being native or non-native speakers is more likely to provide useful information in terms of L2 use, so it will be considered in the analysis of the use of the marker 就是 *jiushi* in Chapter 6.

## 5.6 Access to the Spoken Chinese Corpus

The spoken Chinese corpus has been available at https://github.com/blculyn on GitHub[42] from March 2021. GitHub is a code-hosting platform for version control and collaboration which allows users to work together on projects from anywhere in the world. Although GitHub offers a free service to archive and distribute language resources—many corpus resources can be found on GitHub, it is rarely a common approach that corpus compilers have taken to make their corpora publicly available. Therefore, in this section, I first give an account of the reasons for making the corpus data accessible via GitHub, then I introduce the first version of the spoken Chinese corpus that is released to the research community.

[42] GitHub: https://github.com/.

### 5.6.1 Using GitHub for Corpus Distribution

As McEnery and Hardie (2012) put it, "fundamentally, the corpus-based approach to language cannot do without powerful searching software" (p. 36). In practice, many corpus projects have their own web-based corpus analysis systems, such as the BNCweb interface[43] to the BNC1994 and CQPweb to the Spoken BNC2014 (Hardie, 2012). As has been mentioned in Chapter 1, the International Corpus of Learner Chinese (全球汉语学习者语料库) is also presented with a corpus analysis interface. Typically, the analysis options accessible in these corpus analysis systems include concordances, keywords, collocations, frequency lists, and so on. Such web-search interfaces enable corpus builders to make their corpora accessible via the web browser that all computer users are familiar with. By being available across the web, these corpora are instantly accessible to corpus users on any operating system (McEnery & Hardie, 2012). Also, the ubiquitous analysis functions embedded in these web-based systems enable searches to be carried out very efficiently, which reduce the work that users have to do to process the data when using the corpus approach to analyse language. However, building and maintaining such a web-based interface to a corpus is highly costly and requires a high level of expertise. Considering that this study is a small doctoral project with very limited research funds, it thus was not my first choice to build such a web-based corpus analysis system to release the spoken Chinese corpus.

There are some open language archives, such as the Oxford Text Archive (OTA)[44], the Linguistic Data Consortium (LDC)[45], and the European Language Resources Association (ELRA)[46], offering professional archive services at no cost to corpus contributors. People who attempt to contribute their corpus resources to these archives only need to complete some documentation for submissions. These archives are simply repositories, and do not offer analytical interfaces to corpus data (Thompson, 2005). In addition, corpus resources that are deposited via such archives may not be freely available to people who want to access the data. The spoken Chinese corpus was expected to be accessed by users at no extra cost other than the Internet connection. This being the case, it was decided to find another way to make the spoken Chinese corpus public. Another approach to depositing corpus resources is via the Sketch Engine which is a leading corpus tool (Kilgarriff et al., 2014). The Sketch Engine

[43] The BNCweb: http://corpora.lancs.ac.uk/BNCweb/.
[44] The OTA: https://ota.bodleian.ox.ac.uk/repository/xmlui/.
[45] The LDC: https://www.ldc.upenn.edu/.
[46] The ELRA catalogue: http://catalog.elra.info/en-us/.

website offers many ready-to-use corpora, and tools for users to build, upload, and share their own corpora. Acknowledging the multifunctionality of the Sketch Engine, it is commercial and accordingly corpus contributors need to pay for the services, so it would be costly to deposit the spoken Chinese corpus via the Sketch Engine for a long term preservation purpose. Thus, due to the limited research funds, I decided to make the spoken Chinese corpus public on GitHub.

GitHub has some advantages for the initial release of the spoken Chinese corpus. Firstly, GitHub enables me to make the spoken Chinese corpus openly available at no cost. Secondly, researchers who are interested in this corpus are free to access the data online and download the data to their computers. Additionally, GitHub allows me to upload the corpus data without any paperwork once I registered my GitHub user account. On GitHub, detailed information about the corpus can be provided by using a README file and/or create new files containing metadata of speakers and texts (see 5.6.2). Moreover, GitHub allows me to upload new files, delete, and revise uploaded files whenever possible. Acknowledging the advantages of GitHub, it should be admitted that GitHub also has some drawbacks for corpus distribution. As a code-hosting platform, GitHub may be not familiar to users who are not working with computer sciences. So, non-technical users may find GitHub is unapproachable or unfriendly. Given this, in the following, I will give an introduction of the version of the spoken Chinese corpus that is available on GitHub.

### 5.6.2 The Version of the Spoken Chinese Corpus

The first version of the spoken Chinese corpus which was made available on GitHub includes two repositories: The-spoken-L1-corpus and The-spoken-L2-corpus (see Figure 21). All the files contained in the two repositories are freely downloadable (information such as how to download the files is given in Appendix M). A brief introduction of each corpus is provided in a README file in each repository. Also, each repository has entire transcriptions of complete speech events (see Figures 22 and 23), while in an explicit attempt to create a dataset of interest to researchers working exclusively on L2 language, transcripts which excluded the researcher's turns were provided (i.e., the file of transcripts of L2 speech). Accordingly, transcribed L1 texts which excluded the researcher's turns (i.e., the file of transcripts of L1 speech) were provided as well, which enables researchers to conduct L1–L2 contrastive studies. The corpus study conducted in Chapter 6 was based on the data stored in these two repositories. This being the

case, the release of the transcribed texts containing L1 and L2 speech only makes it easier for researchers to replicate the results of the discourse marker analysis which are given in Chapter 6.

**Figure 21**

*Screenshot from GitHub: The L1 Corpus and the L2 Corpus*

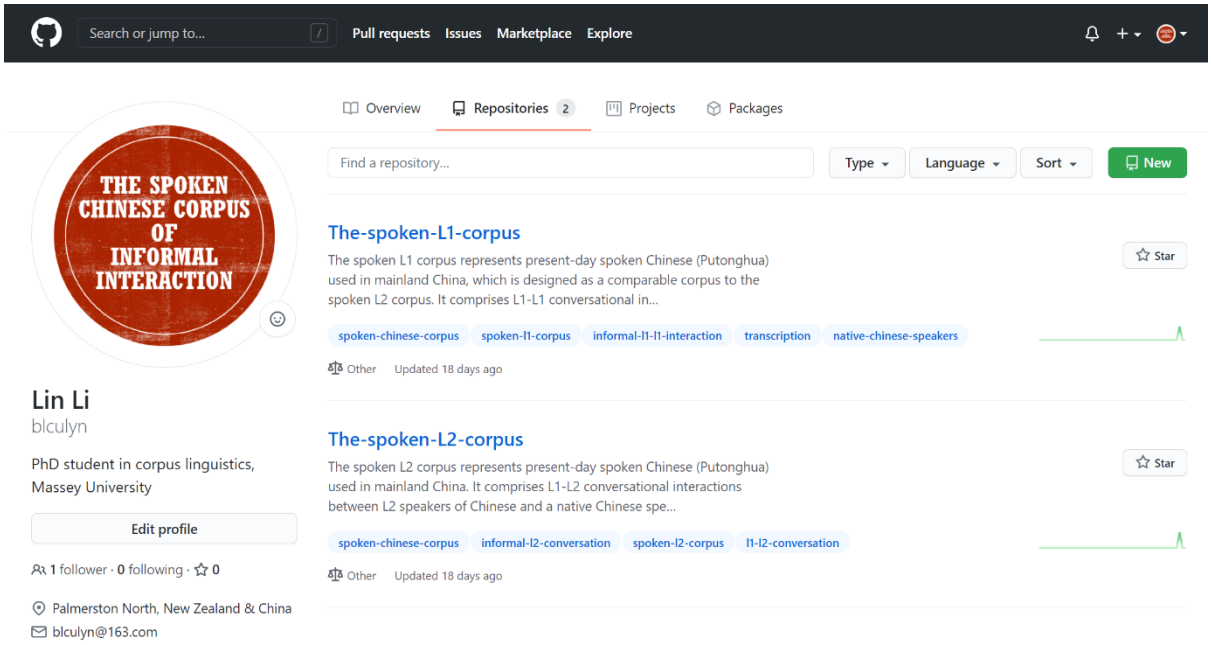



**Figure 22**

*Screenshot from GitHub: The Spoken L1 Corpus*

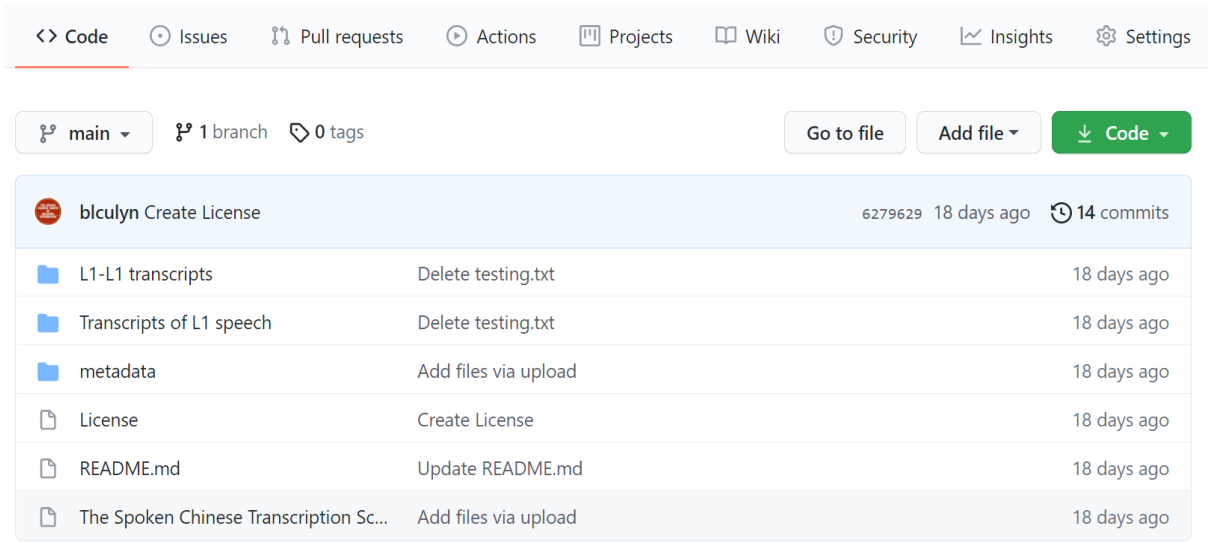


**Figure 23**

*Screenshot from GitHub: The Spoken L2 Corpus*

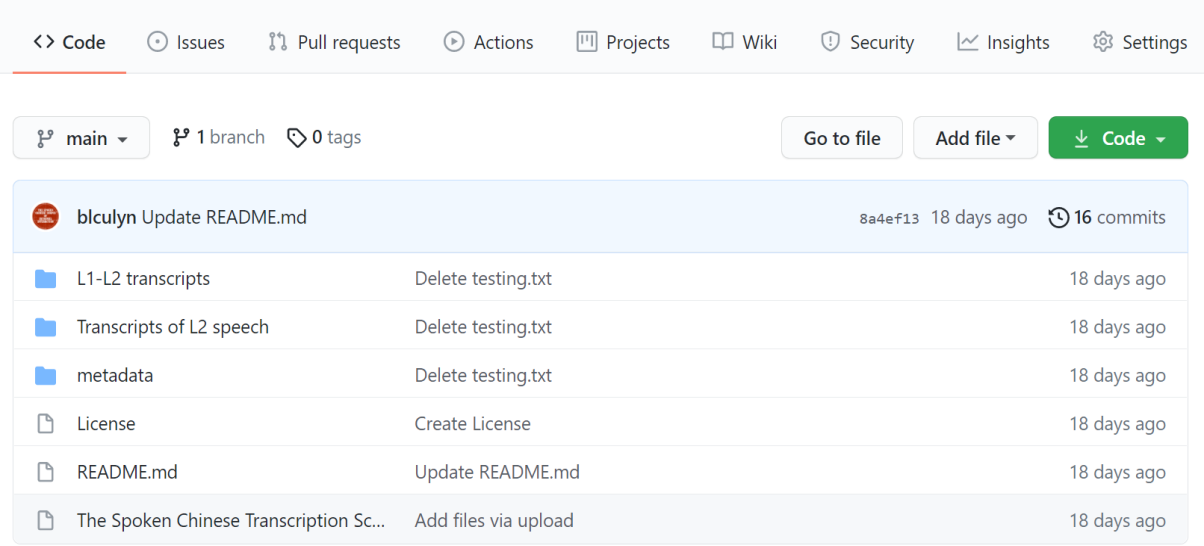



Figure 24 gives an example of the transcribed texts containing speech produced by L2 speakers of Chinese only, while Figure 25 shows an example of the L2 corpus that consists of entire transcriptions of complete speech events. It should be noted that all the transcribed texts that are distributed on GitHub contain raw data rather than annotated data. As a reminder, since there are no spaces between Chinese characters, dividing characters into meaningful units is an inevitable step taken to process the data automatically. In the absence of appropriate annotation, researchers who want to use the spoken Chinese corpus need to find annotation tools to process the data before carrying out studies. This being the case, the availability of the raw data may fail to make the analysis of spoken data easier for users who are not familiar with annotation software. More importantly, it is not surprising that researchers may use different software to annotate the raw data in their studies. Give this, it is important to provide explicit and detailed documentation about the annotations, such as which tool is used and what annotation scheme is applied (Leech, 2005), so that other users will be able to evaluate the validity of the findings of the studies using the spoken Chinese corpus. In the study, as we will see in Chapter 6, the preliminary discourse marker analysis can be undertaken by employing the original raw data; however, by offering annotated data, the research potential of the spoken Chinese corpus can be significantly expended. Bearing this in mind, a tagged version of the spoken Chinese corpus will be made available in the future.

**Figure 24**

*Screenshot from GitHub: An Example of the Texts Containing L2 Speech*

419 lines (419 sloc) | 28 KB

Raw | Blame

```
<S00> erm啊啊我有一个首先有一个问题是你的微信你为什么叫关关
<L01> 啊因为这个这个名字是你你看过啊中国的一个电视剧叫啊国门英雄吗
<S00> 什么英雄
<L01> 叫国门英雄
<S00> 过门英雄没有看过
<L01> 哦这是这个这个电视剧是一个中文电视剧是在在湖北省拍的在武汉市
<S00> 啊啊
<L01> 然后就是这个是关于啊中国的海关
<S00> eng
<L01> ran那个人里面有个关长TA姓关而且TA也是关长所以也叫TA关关我觉得这个名字就是说很快就是很好听的呀关关erm
<S00> 啊就是因为这样所以你叫关关吗
<L01> 欸还有还有因为我是在新西兰做海关的呀
<S00> 哦你是做海关
<L01> eng eng对对对然后我的岗位也可以说是关长的那种的位子所以我觉得关关也很合适所以我就是这样子
<S00> eng那我eng很好那你能给我讲一讲你你你是怎么样就是你是怎么样知道有这部电视剧的吗因为我我没有看过
<L01> 哦因为你啊你和你用YouTube多不多呀YouTube的呀
<S00> eng我是来新西兰之后会用
<L01> 啊然后啊好像我一天是在看一个就是新西兰也有中文电视台的
<S00> eng
<L01> 有CC CCTV8的呀
<S00> eng
<L01> 然后有一天我就看了一下这个有什么这个海关那个er那个电视剧的呀我说eng这个很有意思的然后就马上就看完了erm然后我就在网上查一下然后我发现有一个人把所有的那个电视剧放在YouTube就是在放在YouT
<S00> eng eng
<L01> 所以我觉得我可以从第一集就开始就开始所有的然后就是这样子发现的就是就是我就是有一天我就突然在看电视就发现这个电视剧的呀
<S00> eng
<L01> 然后觉得很挺有意思的呀
<S00> eng
<L01> erm
<S00> 那那那你的汉语是在是怎么学的
<L01> erm我二零零二零零四年就是最后就是坐高中的时候对吧
```

**Figure 25**

*Screenshot from GitHub: An Example of the L2 Transcripts Containing Complete Speech Events*

## 5.7 Summary

This chapter has introduced the spoken Chinese corpus built in this thesis. It has also addressed the fact that although the L1 corpus and the L2 corpus were designed to be comparable, there were differences between and within them. Drawing on the discussions with respect to the comparison of the two corpora, this chapter responds to the second research question proposed in Chapter 1: to what extent extent is the spoken L1 corpus comparable to the spoken L2 corpus? Given that the L1 corpus and the L2 corpus differ from each other in composition due to the exigencies of sampling and availability, it was necessary to be familiar with the corpora that were being used before carrying out the L2 study on the use of 就是 *jiushi*.

In the first section of this chapter, I examined the sample texts contained in each corpus. Here I pointed out that individual contributions were diverse, and all the sample texts differed considerably in size. In consequence, it was argued that there was a risk that the two groups tended to represent some speakers rather than the whole group when comparing the differences between the two corpora. This observation revealed the shortcoming of aggregating data in L2 corpus studies, and prompted meticulous interpretation of the differences in L2 use of 就是 *jiushi* in comparison with the L1 use in Chapter 6.

The second and third sections of this chapter discussed the researcher's contribution to and effect on participants' outputs, in particular the L2 participants' language use. Based on the spoken Chinese corpus, it was argued that there was a negative correlation between the researcher's use and the L1 participants' use of 就是 *jiushi*, while there was little or no relation between the researcher's use and the L2 participants' use of 就是 *jiushi*. The discovery of the lack of correlation in terms of the use of 就是 *jiushi* between the researcher and the L2 participants enabled the analysis in Chapter 6 to be carried out without taking into account the researcher's speech. What should be noted however was that this correlation was limited to the use of 就是 *jiushi* as it occurred in the small L2 corpus, and it was unclear to what extent this correlation would be able to be generalised into a larger population. In addition, although the researcher's use of 就是 *jiushi* has no influence on the L2 use, it was evident that the researcher took different roles in different conversations; as a result, some conversations were more dialogic, while some were highly monologic. Likewise, the interactivity of conversations varied in the L1 corpus. It is worth bearing in mind this variation when discussing the influences of the differences between the corpora on the differences between L1 and L2 use.

The fourth section of this chapter turned to an examination of the variation with respect to the speakers. It placed some emphasis on three factors related to the participants: being native or non-native, gender, and age. By discussing the use of 就是 *jiushi* in interaction, it turned out that the L1 speakers had a tendency to use it more often than the L2 speakers. As regards the gender and age factor, there were marked differences within and between the two corpora, which made it difficult to decide whether the differences between the two corpora could be attributed to the differences between L1 and L2 use. Given that, the contrastive study on the L2 use of 就是 *jiushi* in Chapter 6 will prioritise the factor of being native or non-native.

To sum up, this chapter documented the variation within and between the two corpora in a rather basic way, mainly by using the frequency information and visualising the data. It also discussed how to access the spoken Chinese corpus. Based on the discussion in this chapter, the features of the L2 use of 就是 *jiushi* can be interpreted in a more precise way.

# Chapter 6 Investigating the L2 Use of 就是 *Jiushi* in the Corpus

## 6.1 Introduction

In this chapter the focus of the thesis shifts from the spoken Chinese corpus construction to the discourse marker analysis. The linguistic patterns which are called 'discourse markers' in this thesis, are also known by a variety of other names, such as 'discourse particles' (e.g., Fischer, 2006a, 2006b), and 'pragmatic markers' (e.g., Andersen, 2001; Fraser, 1996), to name but a few (for a broader set of references, see Beeching, 2016; Fischer, 2006b; Fraser, 1999; Jucker & Ziv, 1998). In this study I use the term 'discourse markers' to refer to a group of linguistic items with the following properties: their function is to express pragmatic aspects of communication, for example by marking speakers' attitude or illocutionary force, and they do not contribute to the propositional content of the utterances in which they occur (Andersen, 2001; Schourup, 1999), such as *well* and *like* in English.

The L1 corpus abounds with Chinese discourse markers, such as 就是 *jiushi* 'be exactly' (Biq, 2001; Shi & Hu, 2013; Zhang & Gao, 2012), 然后 *ranhou* 'then' (Li, 2009; Wang, 2018; Wang & Zhou, 2005; Xu, 2009), 那个 *neige/nage* 'that' (Chen, 2017; L. Liu, 2009; Yue, 2020), 这个 *zheige/zhege* 'this' (L. Liu, 2009), and 不是 *bushi* 'to not be' (Chen, 2018, 2019). Table 21 shows the top 30 most frequent words used by the L1 speakers of Chinese in the L1 corpus[47]. As we can see clearly, 就是 *jiushi*, 然后 *ranhou*, and 那个 *neige/nage* which are given in boldface occur more than 1,000 times and are among the top 10 most frequent items in L1 speech[48]. Similarly, these three expressions are among the high frequency words used by the L2 speakers of Chinese, occurring 2,481 times, 1,454 times, and 1,272 times respectively[49].

[47] Chapter 2 has discussed that it is essential to divide Chinese characters into meaningful units prior to processing Chinese corpus data automatically. Therefore, to create the frequency list in Table 21, I segmented characters in the L1 transcripts into meaningful units by using an online software–*jieba* (https://app.gumble.pw/jiebademo/). Then I manually checked all the word-segmented texts to ensure that the occurrences of the item 就是 *jiushi* under examination were segmented accurately. No particular attention was paid to the accuracy of other words in the L1 transcripts due to time constraints.

[48] Table 21 presents the total number of the three items that occur in the L1 corpus rather than the frequencies of their occurrences as discourse markers.

[49] The frequencies refer only to the number of the occurrences of these expressions that used by the L2 participants.

**Table 21**

*Top 30 Most Frequent Words in the Spoken L1 Corpus*

| Rank | Item | Frequency | Rank | Item | Frequency |
|---|---|---|---|---|---|
| 1 | 的 de | 5,348 | 16 | 个 ge | 935 |
| **2** | **就是 jiushi** | **3,837** | 17 | 这个 zheige | 917 |
| 3 | 就 jiu | 3,601 | 18 | 说 shuo | 876 |
| 4 | 我 wo | 3,418 | 19 | 对 dui | 855 |
| **5** | **然后 ranhou** | **2,862** | 20 | 什么 shenme | 853 |
| 6 | 是 shi | 2,420 | 21 | 不 bu | 782 |
| 7 | 了 le | 2,189 | 22 | 可能 keneng | 774 |
| 8 | 你 ni | 1,760 | 23 | 吧 ba | 768 |
| **9** | **那个 neige** | **1,746** | 24 | 因为 yinwei | 754 |
| 10 | 也 ye | 1,405 | 25 | 他们 ta men | 739 |
| 11 | 在 zai | 1,173 | 26 | 觉得 juede | 695 |
| 12 | 有 you | 1,167 | 27 | 但是 danshi | 661 |
| 13 | 去 qu | 1,089 | 28 | 没有 meiyou | 644 |
| 14 | 我们 women[50] | 1,022 | 29 | 会 hui | 634 |
| 15 | 都 dou | 966 | 30 | 那种 neizhong | 634 |

*Note*. Only L1 participants' turns (a total of 174,009 words) were used to the creation of this word frequency list.

Each of these expressions as a discourse marker can serve multiple discourse functions in conversations. For example, 就是 *jiushi* as a discourse marker occurring at the beginning of a turn typically marks the onset of a new turn after successful turn-transition, and it also can serve to introduce new topics (Shi & Hu, 2013; Yao & Yao, 2012); the marker 然后 *ranhou* can mark the consequence of the prior propositions, introduce new information in speech and signal topic shifts (Ma, 2010; Wang, 2018); 那个 *neige/nage* as a discourse marker can initiate

[50] The 14th item 我们 is composed of the pronoun 我 *wo* 'I' and the suffix 们 *men* which is a plural marker for pronouns. Its English meaning is 'we' which differs from the meaning of the plural form of the English word 'woman', and the pronunciation of 我们 is distinct from the English word 'women' as well.

new topics in interaction (Yue, 2020). These functions have been identified and elaborated largely based on very limited spoken data (see Section 6.2), and corpus data and methods are still not commonly used in previous studies of Chinese discourse markers (e.g., Shi & Hu, 2013; Wang, 2018; Xu, 2008; Zhang & Gao, 2012). Against this background, this study is an empirical study which draws on data from the spoken Chinese corpus produced as part of this thesis, seeking to show how corpus data and methods can benefit the investigation of discourse markers and deepen our understanding of what they are doing in informal speech, especially in L2 interaction with L1 speakers.

It is commonplace that these Chinese discourse markers have non-discourse marker statuses. For example, 就是 *jiushi* can be a verbal phrase, a conjunction or an adverb, the canonical meaning of 然后 *ranhou* is a conjunction, and 那个 *neige/nage* can function as a demonstrative. In some cases, the lines between their discourse marker use and non-discourse marker use are not clear-cut, which indicates that demarcating their discourse marker uses is arguably labour-intensive in a study making use of corpus data. Some corpus methods can be employed to reduce the burden of identification of discourse marker uses, such as a versatile corpus tool *GraphColl*[51] (Brezina et al., 2015), which can be used to extract concordance lines that contain the target expressions. The shortcoming of concordance lines lies in the fact that it is sometimes difficult to identify discourse marker uses without taking into consideration the context where the items occur. In consequence, the task of identification requires the manual examination of each utterance that contains the target discourse markers, which makes the identification process prohibitively time-consuming even in a small corpus like the spoken Chinese corpus. This being the case, the task of providing a comprehensive description of a full set of Chinese discourse markers deserves much broader attention than the scope of the present work allows for. It then was decided to focus on one item, seeking to investigate its discourse marker use in informal speech. Given that 就是 *jiushi* is the second most frequent word used by the L1 speakers as well as being uttered highly frequently by the L2 speakers of Chinese, it was decided that this chapter would focus on the use of 就是 *jiushi* as a discourse marker in informal conversations. The discourse marker use of 就是 *jiushi* began receiving scholarly attention about two decades ago in Fang (2000) and Biq (2001), while studies on the marker 就是 *jiushi* are markedly rare and focus largely on L1 speech. Considering that

[51] *GraphColl*: https://www.clarin.eu/content/graphcoll.

discourse marker use in L2 data remains underexplored, the analysis of the L2 use of the marker 就是 *jiushi* carried out in this chapter will provide some valuable insights into the roles the marker 就是 *jiushi* has in L2 informal interaction.

This chapter is structured as follows. It begins by introducing the spoken data used in the study of 就是 *jiushi* in Section 6.2. I critically review the data employed in previous studies on the marker 就是 *jiushi*, and address the impacts the data have on the results. Then I give an account of the canonical meanings of 就是 *jiushi* in Section 6.3. Following this, I outline some pragmatic uses of 就是 *jiushi* in the corpus associated with a review of previous studies on the functions of the marker 就是 *jiushi* in Section 6.4. I also explain my decision to focus on a very restricted range of utterances that contain 就是 *jiushi*. Section 6.5 clarifies some important notions of relevance theory which are fundamental to the present study. Then I discuss the functional interpretations of the marker 就是 *jiushi* in some detail in this section. In Section 6.6 I provide a quantitative analysis of the differences between the L1 and L2 use of 就是 *jiushi*, and suggest that caution should be exercised when generalising the findings or results into a wider group of language users.

## 6.2 The Spoken Data

Although corpora are accepted as a useful source of data by the research community (McEnery et al., 2019), spoken Chinese corpora are proceeding relatively slowly in terms of their impact on both L1 and L2 pragmatic research. Table 22 lists several articles focusing mainly on the marker 就是 *jiushi*. It shows that the pragmatic versatility of the marker 就是 *jiushi* has been considered in research focused on formal or scripted speech, such as television drama productions (e.g., Shi & Hu, 2013; Yao & Yao, 2012). Undeniably, these studies provide useful findings with respect to the functions of the marker 就是 *jiushi* in specific situations, but it remains doubtful to what extent such findings can be generalised to the use of the marker 就是 *jiushi* in other contextual situations. Additionally, it is noticeable that the analysis of the marker 就是 *jiushi* in informal conversation has been carried out on the basis of very limited data. For example, Zhang and Gao (2012) discuss the use and development of the marker 就是 *jiushi* by examining conversations of approximately 55,000 words. This being the case, some actual

usages of the marker 就是 *jiushi* probably have been missed in previous studies and the findings may have failed to uncover the complexity of the roles of the marker 就是 *jiushi* in interaction.

**Table 22**

*Material Used in Previous Studies on 就是 Jiushi*

| **Reference** | **Material** |
|---|---|
| Biq (2001) | recordings of naturally occurring conversation and a written Chinese corpus in press reportage drawn from the People's Daily |
| Zhang and Gao (2012) | about 130 minutes of conversations: 55,000 words |
| Yao and Yao (2012) | a TV talk show *Dialogue*: 526,000 words<br>a TV serials *Stories in the editorial office*: 133,000 words |
| Shi and Hu (2013) | two TV talk shows: 280,000 words |
| Li (2016) | unclear |

Compared to the written to be spoken data used in previous studies on 就是 *jiushi*, the present study makes use of corpus data consisting of informal conversations. This means that the corpus data used in this study were gathered and organised in a rather systematic way, and represent an authentic manifestation of language use. As a result, this study is capable of uncovering some usages of 就是 *jiushi* in more natural settings which have not been discussed in previous studies (see 6.4.3). Moreover, as has been mentioned in Chapter 5, the corpus data that were used in this study are available on GitHub, making it possible for other researchers to replicate the findings of this discourse marker analysis of 就是 *jiushi*. Another contribution of the present study is that the spoken Chinese corpus enables this study to look at 就是 *jiushi* on a larger scale of spoken data in comparison to the conversational data on which other studies were based.

Table 23 shows the frequencies of the item 就是 *jiushi* both in the L1 corpus and the L2 corpus (for the distribution of 就是 *jiushi* in each corpus see Appendices K and L). Given that there is no observable relation between the researcher's use and the L2 speakers' use of 就是 *jiushi* (see Section 5.4 in Chapter 5), this study only takes into consideration the usages of 就是 *jiushi* by L2 participants in the L2 corpus. For the sake of comparability, occurrences of

就是 *jiushi* used by the researcher in the L1 conversations are also excluded, only L1 participants' uses are considered.

**Table 23**

*Frequencies of 就是 Jiushi in the Spoken Chinese Corpus*

| The spoken Chinese corpus | 就是 *jiushi* | | |
|---|---|---|---|
| | **Total (words)** | **Participant's use** | **Researcher's use** |
| The L1 corpus | 4,691 (100%) | 3,837 (82%) | 854 (18%) |
| The L2 corpus | 3,588 (100%) | 2,481 (69%) | 1,107 (31%) |
| Total | 8,279 (100%) | 6,318 (76%) | 1,961 (24%) |

As has been discussed in Chapter 4, all the transcripts exclude punctuations and I have taken speaker turns as the basic units. However, as we will see in the following sections, the utterance is an important concept in the theoretical analysis of 就是 *jiushi* and should be clarified. So, in this chapter the term 'utterance' refers to a sequence of words uttered communicating a complete proposition, but it does not necessarily do so in accordance with a grammatical sentence. In this study, the spoken Chinese corpus includes only raw data/unannotated data; therefore, the identification of the discourse marker use and the functional analysis of 就是 *jiushi* were carried out manually.

## 6.3 Canonical Meanings of 就是 *Jiushi*

The item 就是 *jiushi* consists of the adverb 就 *jiu* (use for emphasis) 'just' and the copula 是 *shi* 'be'. Traditionally, 就是 *jiushi* is attributed to a verbal phrase, which literally means 'A is precisely B'. For example, in (9), 'kiwifruit' and 'mi-hou-tao', linked by 就是 *jiushi*, are two different names for the same kind of fruit. This example can be called an "equational sentence" (Li & Thompson, 1977, p. 419), in which identification or a member/class relationship is expressed between two noun phrases.

(9) 奇异果**就是**咱们国内的猕猴桃呗(N18-C01) [52]

[52] Examples that are extracted from the spoken Chinese corpus have been labelled as 'Speaker ID + Text ID', e.g., N04-C01.

kiwifruit **jiushi** our domestic mi-hou-tao (Monkeypeaches)

*kiwifruit **is exactly** the fruit that we call mi-hou-tao in China*

It is widely known that 就是 *jiushi* can function as an adverb as well. For instance, in (10) 就是 *jiushi* is used as an adverb. The speaker in (10) was talking about a child in her class who was a bit egocentric and overindulged by his parents. In the conversation, the speaker said that this child refused to accept other children's different opinions and insisted he was very much right in what he did or said. On this occasion, the adverb 就是 *jiushi* is used for emphasis.

(10) 然后他就说我**就是**对的我**就是**对的 (N05-C01)

then he said I **jiushi** right I **jiushi** right

*then he said I am right I am right*

The adverb 就是 *jiushi* also has other syntactic functions in interaction. In (11), it is used as a downtoner and means 'is as little as' (Biq, 2001; Lv, 1999). Conversely, 就是 *jiushi* in example (12) is used as a uptoner emphasizing that the frequency of being drunk was quite high (Lv, 1999). In this sense, 就是 *jiushi* typically occurs with 一 *yi* 'one, once' to constitute a fixed expression 一…就是 *yi…jiushi* 'so long as…then'. Moreover, the adverb 就是 *jiushi*, as exemplified in (13), also appears at the end of the speaker's speech to make conclusions.

(11) 所以我当时就是去听 TA 的那个辅导课可能也**就是**三四个学生的样子 (N02-C02)

so I at that time jiushi go listen to his/her tutorial maybe **jiushi** three four students

*so I went to his/her tutorial (there were) probably **only** three or four students*

(12) 一回家**就是**要醉一回 (N24-C01)

once go home **jiushi** be drunk

***so long as** he gets together with his family **then** he must be drunk*

(13) 对总的来说**就是**这样啦 (N01-C01)

right overall **jiushi** like this la

*yeah overall this is what **exactly** happened (to me)*

In addition to being a verbal phrase or an adverb, 就是 *jiushi* can function as a conjunction. The conjunction 就是 *jiushi* can be associated with the negative copula 不是 *bushi* 'to not be' to constitute a rather fixed expression 不是…就是 *bushi…jiushi* 'either…or…' in

a choice construction. For example, the speaker in (14) was recalling her interview for a position as a volunteer Chinese teacher. She mentioned that the interviewer asked her a question about how to deal with unexpected events in classroom. However, the speaker could not remember more details of the situation in which this question was asked. Then the speaker used the expression 不是…就是 *bushi…jiushi* 'either…or…' to show her uncertainty about the situation of the event.

(14) 我忘了它**不是**最后一个环节**就是**倒数第二个环节(N02-C02)

I forgot it was **not** the last part **jiushi** the second to last part

*I could not remember it was **either** the last part **or** the second to last part*

In (15), 就是 *jiushi* is used as a conjunction as well. The speaker in this example was explaining her impression of New Zealand to the hearer. She claimed that she felt very safe in her community, to support this statement she then stressed that people did not need to lock their doors even if they went out. In this case, 就是 *jiushi* occurs together with the adverb 也 *ye* 'also' to constitute a fixed expression 就是…也 *jiushi…ye* 'even if' to express concession.

(15) 我觉得大家**就是**出门**也**不用锁自己的房门啊 (N02-C01)

I think people **jiushi** go out **ye** no need to lock their doors

*I think there is no need to lock your door **even if** you are not home*

Furthermore, the conjunction 就是 *jiushi* can be used to provide explanations to the preceding proposition in context, and this function is developed from the verbal phrase 就是 *jiushi* due to the process of grammaticalization (Zhang, 2002). For example, in (16) the speaker started by sharing the problem he realised after he got into a relationship and the segment following 就是 *jiushi* explained what exactly the problem was. According to Zhang (2002), 就是 *jiushi* is a straightforward conjunction in (16), because (i) the two segments connected by 就是 *jiushi* are syntactically independent, (ii) each segment can express a complete proposition, and (iii) the omission of 就是 *jiushi* would not cause ungrammaticality.

(16) 然后我自己自我感知里面会发现一个问题**就是**在你单身状态下你是因为每个人他可能自律的程度不一样 (N17-C01)

ranhou myself self-aware inside can find a problem **jiushi** you are single you are because everyone his/her self-discipline degree is different

*and myself I am self-aware to know there is a problem **that is** when you are single you are because the degree of self-discipline of everyone is different*

Whenever 就是 *jiushi* does not serve one of the above syntactic functions, it is considered a discourse marker. Fang (2000) states that the discourse marker use of Chinese conjunctions is treated as evidence that a process of semantic reduction is affecting the roles of conjunctions in speech. In other words, some Chinese conjunctions in conversation do not affect the truth conditions of utterances that contain them, and they are only taken to signal coherence relations. In the case of 就是 *jiushi*, according to Biq (2001), its semantic meaning is reduced to a connective marking textual coherence, and its connective sense becomes even more semantically empty when it serves as a "mere pause filler" (p. 61) in conversation. In what follows, I will discuss the functions of 就是 *jiushi* as a discourse marker recognised in the literature in some detail associated with my observations of the typical uses of 就是 *jiushi* in the spoken Chinese corpus.

## 6.4 Use of 就是 *Jiushi* in the Spoken Chinese Corpus

This section attempts to succinctly introduce the typical usages of the marker 就是 *jiushi* in the spoken Chinese corpus. Most of the Chinese examples in this section were taken from the spoken L1 corpus. Some usages found in the L1 corpus have received some attention in the Chinese literature. It is certain that discussions in previous studies contribute to our growing understanding of 就是 *jiushi* as a discourse marker; the disadvantages of those studies however are obvious in the sense that there is a lack of adequate theoretical basis for the analyses of the marker 就是 *jiushi*. I will return to this issue later in this section. In addition to those functions that have been identified, there are some usages which have so far rarely been discussed in the literature but were rather observable in the L1 corpus. So, I will give a brief description of these usages of 就是 *jiushi* at the end of this section. Although the focus was not on these usages, their existence shows researchers the direction for the future corpus studies on the marker 就是 *jiushi*.

Due to space constraints, this chapter is concerned with one function of the marker 就是 *jiushi* (see 6.5.3). It does not mean that the other uses of the marker 就是 *jiushi* are of little

value to the Chinese research community. Rather, emphases should be placed on all the functions of 就是 *jiushi* as a discourse marker in conversational interactions. This being the case, it is my intention to discuss these functions systematically in other writings rather than to illuminate all the usages in one single chapter. Therefore, this study is mostly concerned with one function of the marker 就是 *jiushi* that has received some attention in the literature. Other usages are then discussed briefly[53]. As we will see in Section 6.5, my analysis of the relations that 就是 *jiushi* mark are based on relevance theory (Sperber & Wilson, 1986b, 1995), and the taxonomies and classifications of relations are grounded in "a cognitive theory of the way in which linguistic meaning and context interact in discourse understanding" (Blakemore, 1997, p. 4). Clearly, the marker 就是 *jiushi* has English equivalents. To facilitate better understanding of the functions of the marker 就是 *jiushi*, I therefore provide some examples containing discourse markers *you know*, *I mean* and *like* as well.

### 6.4.1 就是 *Jiushi* in Apposition

One notable function of 就是 *jiushi* as a discourse marker that has been identified in the Chinese research literature is that its use in marking explanations to the preceding proposition in context (Biq, 2001; Li, 2016; Shi & Hu, 2013; Zhang & Gao, 2012). According to previous studies, (17) is classified as an instance of elaboration and 就是 *jiushi* is an example of an elaboration marker.

(17) 但当时还好有一个这个中文的老师在那儿所以可以帮助他稍微辅助一下**就是**如果他困难的话可以去直接用中文去求助 (N07-C01)

but at that time fortunately there was one zhege Chinese teacher over there so can help him a bit assist **jiushi** if he had problems can go directly use Chinese to ask help

*but at that time fortunately there was a Chinese teacher in the kindergarten so (the teacher) could help him **you know** if he had problems he could ask for help in Chinese*

[53] It should be noted that there are some occurrences containing 就是 *jiushi* which cannot be identified since the utterances are incomplete.

In examples like (17), Zhang and Gao (2012) argue that the hearer can still understand the speaker without the second segment that contains 就是 *jiushi*; however, the utterance that contains 就是 *jiushi* is taken to complement the original segment to make sure the hearer can interpret accurately what the speaker is communicating. It seems that in spoken English some markers such as *you know* have the similar function as 就是 *jiushi*. For instance, (18) is taken from the Spoken BNC2014, in which the segmentation following *you know* may be interpreted as an alternative means for classifying what is communicated by the preceding utterance 'when you're like left out'.

(18) (The Spoken BNC2014, S5YC)[54]

A: she was like yeah when I was at school I always used to get jealous of friends and you and <name> seem to have all these amazing times together when I'm not there I'm like what what are you on about like

B: oh dear

A: like she was like well you know like just little things that I can't be involved in I was like yeah but you're never gonna be involved in everybody's conversation there's always gonna be a time when you're like left out ***you know*** when you're not involved in something

Recall (16) which is repeated below, in which 就是 *jiushi* is taken as a conjunction to elaborate or explain the preceding utterance. This relation of elaboration seemingly differs from the one justified by Zhang and Gao (2012). However, without explicit criteria, previous studies on the elaboration function of 就是 *jiushi* fail to sufficiently distinguish the conjunction 就是 *jiushi* from the discourse marker 就是 *jiushi* on some indeterminable occasions with respect to the elaboration functions. As we will see in 6.5.3, within the relevance theoretic approach, 就是 *jiushi* in examples like (17) signals that the utterance following 就是 *jiushi* is an alternative means for communicating what is communicated by the preceding utterance. According to this view, the function of 就是 *jiushi* in (17) is distinct from that in (16) in which the utterance following 就是 *jiushi* describes the very problem the speaker found rather than rephrases the first segment.

[54] S5YC is the text ID used in the Spoken BNC2014. For the sake of readability, all examples taken from this spoken corpus were simplified by excluding speaker IDs (which were replaced with capital letters A, B, and C), punctuation markers and all tagged labels (e.g., <pause dur="short"/>).

(16) 然后我自己自我感知里面会发现一个问题**就是**在你单身状态下你是因为每个人他可能自律的程度不一样 (N17-C01)

ranhou myself self-aware inside can find a problem **jiushi** you are single you are because everyone his/her self-discipline degree is different

*and myself I am self-aware to know there is a problem **that is** when you are single you are because the degree of self-discipline of everyone is different*

In addition, Zhang and Gao (2012) hold the view that an explicature function which is claimed by Shi and Hu (2013) is in accordance with their analysis of the elaboration function of the marker 就是 *jiushi*[55]. The explicature relation signalled by the marker 就是 *jiushi* is that the speaker uses 就是 *jiushi* consciously to constrain the contextual assumptions of the hearer in order to enable the hearer to understand discourse successfully. Using this analysis, (19) is an instance of this function. The speaker in (19) explained why she chose to learn French rather than Dutch. She claimed that there were unlikely to be many chances to use Dutch in China, so it would be a waste of time learning Dutch. According to Shi and Hu (2013), in which the theory that the study is based on is unclear, the segment following 就是 *jiushi* in (19) enables the hearer to interpret the first segment in one specific contextual situation: there is no need for Dutch. Consider the two segments connected by *I mean* in (20), 'he weren't employed at all' can be taken as the very explanation or assumption in terms of 'all for nothing' the speaker would like to communicate.

(19) 回国之后我觉得荷兰语的话也不会用到**就是**需求量比较小 (N14-C01)

return to China after I think Dutch cannot be used **jiushi** the need is relatively small

*if I come back to China I do not think I will use Dutch **I mean** there is no need for Dutch*

(20) (The Spoken BNC2014, SYHW)

A: you know it was just along the pavements we used to have a bloke er called <name> who used to clear it (snow) all our pavements didn't he

B: yeah

[55] It should be noted that Shi and Hu's (2013) analysis of explicature concerns the expression 就是说 *jiushi shuo* rather than 就是 *jiushi*. 就是说 *jiushi shuo* consists of the item 就是 *jiushi* and the verb 说 *shuo* 'say', which is literally translated as 'that is to say'. In previous studies 就是 *jiushi* is often associated with the expression 就是说 *jiushi shuo*, as they to some extent are identical in meaning on some occasions. However, the present study did not take 就是说 *jiushi shuo* into consideration due to the very limited examples (only 38 occurrences) in the L1 corpus based on which reliable results were unlikely to be reached.

A: all for nothing you know ***I mean*** he he weren't employed at all

C: yeah

A: he just used to clear the pavements all the way up to the school didn't he

B: mm

From the perspective of relevance theory (see 6.5.1), the uses of 就是 *jiushi* in (17) and (19) are slightly different. Although Zhang and Gao's (2012) analysis is to some extent in line with my observation of the use of 就是 *jiushi* in utterances like (17) which is numbered as (43) in 6.5.3, the notion of elaboration in the previous studies is inaccurate because it does not provide an adequate theoretical basis for the analysis of the marker 就是 *jiushi*, and there is a lack of criteria as to what counts as an instance of this relation. Therefore, the elaboration relation ignores the differences between the functions of 就是 *jiushi* in (17) and (19). Within the relevance theoretic framework, following Blakemore (1997), I argue that the term 'elaboration' used in previous studies on the functional interpretations of 就是 *jiushi* does not define a single class of phenomena and there indeed is a subset of utterances classified in the literature in terms of elaboration that must be given a different sort of analysis.

In addition to the above uses of 就是 *jiushi*, it is clear that (21) is not accommodated in Zhang and Gao's analysis of elaboration or in any studies on the marker 就是 *jiushi* in the Chinese research literature. It is not difficult to see that 就是 *jiushi* in (17) can be said to signal sequential relationship between one segment of discourse and the preceding text. However, it is not clear whether this analysis of 就是 *jiushi* applies to examples like (21) where 就是 *jiushi* seems to mark not a sequential connection between an utterance and the preceding text, but rather a connection between a parenthetical constituent and its host utterance. Similar usages can also be found in English informal conversations. Consider (22), the segment 'no I mean you can't get a grade' is the host utterance, 'a good grade' following the marker *you know* can be seen as a parenthetical constituent.

(21) 国内的压力相对小一点**就是**舆论的压力(N11-C01)

domestic pressure relatively small a bit **jiushi** public opinion pressure

*people have relatively less pressure in China **you know** the pressure of public opinion*

(22) (The Spoken BNC2014, S2ZU)

A: are you calling <name> an idiot

B: no I mean you can't get a grade ***you know*** a good grade

The segments following 就是 *jiushi* and *you know* in (21) an (22) are parentheticals in the sense that they are discontinuous constituents licensed by grammar, and yet they bear no obvious syntactic relationships to the utterances that contain them (Blakemore, 2005; Espinal, 1991). A further explanation of this use of 就是 *jiushi* will be given in 6.5.3. Here I argue that examples like (21) have so far been ignored in studies on the functions of the marker 就是 *jiushi*; however, this type of use is common in informal conversations and it is worth an exhaustive investigation.

In conclusion, the notion of elaboration used in the studies on the marker 就是 *jiushi* is problematic and it plays no role in the explanation of the way in which these utterances that contain 就是 *jiushi* are interpreted. Since this study is a tentative analysis of the discourse marker by employing the relevance theoretic approach, in this chapter I shall be primarily concerned with the interpretive use of the discourse marker 就是 *jiushi* in utterances like (17), (19), and (21). Nevertheless, to get a good understanding of the functions the marker 就是 *jiushi* has in conversational interaction, in what follows, I will demonstrate some functions that have been recognised by researchers in 6.4.2 and some found in the spoken Chinese corpus but which have rarely been discussed in previous studies in 6.4.3.

### 6.4.2 就是 *Jiushi* as a Hesitational Device/Self-Repair

In the literature, some functions in addition to the elaborative function have been attributed to the marker 就是 *jiushi*. According to previous studies, 就是 *jiushi* can sometimes be interpreted as a hesitational device. When functioning in this way, 就是 *jiushi*, for example in (23), is used to hold the floor (Biq, 2001; Li, 2016; Shi & Hu, 2013; Yao & Yao, 2012; Zhang & Gao, 2012).

(23) 就是它那个山上**就是就是就是**很多拐 (N04-C01)
jiushi it that on the mountain **jiushi jiushi jiushi** many turns
*you know on that mountain road* ***it's like you know*** *there are many turns*

It should be noted that 就是 *jiushi* used in multiples is among the ways in which 就是 *jiushi* may be taken as a hesitational device in conversations, and yet not all the incidents of 就是

*jiushi* being used in multiples can be identified as discourse markers. For instance, it seems to me that the first 就是 *jiushi* in (24) is a verbal phrase 'be exactly' which bears a syntactic function in the utterance.

(24) 然后然后那个我第一反应**就是就是就是**正常来说吧就是<name>可能每周 er 就是周六的时候因为我不是歇班嘛是吧 (N01-C02)

ranhou ranhou neige I first reaction **jiushi jiushi jiushi** normally speaking jiushi <name> may every week er jiushi Saturday because I do not work right

*and then my first reaction **was like you know** normally you know <name> may (come) every week er you know on Saturday because I am off duty (on Saturdays) you know*

As Table 24 below shows, 5.39% of the occurrences of 就是 *jiushi* are used in multiples in L1 speech, and this type of use of 就是 *jiushi* accounts for 3.39% in the total number of the occurrences of 就是 *jiushi* used by the L2 speakers (see Appendices K and L). It may be an interesting task to investigate how 就是 *jiushi* used in multiples can contribute to utterance interpretation. However, because of space constraints, an exhaustive explanation of the hesitational device 就是 *jiushi* is beyond the scope of this chapter.

**Table 24**

*Frequencies of 就是 Jiushi Used in Multiples in the Spoken Chinese Corpus*

| **就是 *jiushi* used in multiples** | **The spoken Chinese corpus** | |
|---|---|---|
| | The L1 corpus | The L2 corpus |
| 就是就是 *jiushi jiushi* | 191 | 80 |
| 就是就是就是 *jiushi jiushi jiushi* | 16 | 4 |
| Total | 207 | 84 |

*Note*. Only participants' uses were calculated.

Apart from its association with planning difficulties, Shi and Hu (2013) notice that 就是 *jiushi* can also collocate with self-repairs in utterances like (25) below. Within the relevance theoretic framework, I question the adequacy of glossing the marker 就是 *jiushi* as a mere hesitational device or a repair signal. 就是 *jiushi* can be assigned meanings which pertain to "the relation between a speaker's thought and the external representation of this thought"

(Andersen, 2001, p. 228), and this meaning cannot associate with hesitational devices such as *erm* or *er*. This does not mean, however, that I reject the possibility that 就是 *jiushi* can be used to bridge gaps in conversation. Instead, following Andersen's comments on the hesitational function of *like* (2001), I tend to argue that 就是 *jiushi* in (25) has a similar function. That is, it has a capacity to provide a link between the propositional elements that may otherwise be syntactically or logically unrelated. Consider (26), *you know* also shows a hesitational linking function.

(25) 比方说卫生间提供的那些纸呀泡**就是**那个洗手液呀 (N02-C01)
for example bathrooms provide those toilet paper bu(bble) **jiushi** neige hand wash
*for example (in New Zealand) they provide toilet paper in bathrooms (and) bu(bble)* ***you know*** *hand wash*

(26) (the Spoken BNC2014, S2UJ)
A: so did your foot get bad when you went for those walks across the ***you know*** when you did long walk
B: yeah that was bad

So far, I have discussed briefly functions that have been qualified in the literature. In the following context, I will give a brief description of some uses of 就是 *jiushi* found in the spoken Chinese corpus but which have not received much attention in the literature.

### 6.4.3 Other Uses of 就是 *Jiushi* in the Spoken Chinese Corpus

There are some usages of 就是 *jiushi* that have rarely been discussed in the Chinese research literature. Within the relevance theoretic framework, the use of 就是 *jiushi* in (27) can be interpreted to signal a relation of exemplification, that is, it is taken as providing evidence to support a claim or proposition. It seems that *I mean* in (28) has the similar function as well. The segment following *I mean* is used to support the claim that they now had less snow than before.

(27) 他们都特别尊敬老师 eng 然后有一次**就是**学生还会就是在一个典礼上然后学生就扑通一下就跪下了 (N20-C01)

they all very respect teachers eng ranhou once **jiushi** students also can jiushi at one ceremony ranhou students suddenly kneeled

*they all respect their teachers very much and once **I mean** students also like at one ceremony and students suddenly kneeled to their teachers*

(28) (The Spoken BNC2014, SHYW)

A: we seemed to have lots of snow when we were younger than than we get now ***I mean*** we got nothing last year did we

B: no nothing last year

Another use of 就是 *jiushi* found in the spoken Chinese corpus is given in (29). At first sight, 就是 *jiushi* in (29) may be a hesitational device bridging the gap in speech. However, this statement is untenable. Based on the transcription and recordings, 就是 *jiushi* often occurs in the midst of a continuous and rapid flow of speech and is not prosodically separated from the rest of the utterance; hence the online production of the utterance does not cause the speaker much difficulty. In fact, it is highly common that 就是 *jiushi* occurs between elements that are constituents of the same clause and is pronounced with the same efficiency of deliverance as the 'real' constituents of that clause.

(29) 如果我们如果要想**就是**上山的话必须走另外一条道儿 (N11-C01)

if we if wanted to **jiushi** go up the hill must walk another path

*if we if (we) wanted to **you know** go up the hill (we) had to take another path*

This kind of use of markers has been noticed in English: for example, Andersen (2001), from the relevance theoretic viewpoint, argues that the marker *like* in (30) signals that the film can be collected the day after it has been brought in and not 'more or less' one day after, and it is applied to mark off metalinguistic use of expressions. Besides, it seems that *you know* in (31) also functions in the similar way as *like* and 就是 *jiushi*. While Andersen's work on the marker *like* can provide some useful insight into the use of 就是 *jiushi* in (29), this analysis is beyond the scope of this thesis.

(30) It's **like** one day developing, right, and she hasn't got round to collecting them yet.

(31) (the Spoken BNC2014, S2U9)

A: I think <place> could be down for months

B: it takes an awful long time but the the river <place> does carry everything way ***you know*** very quickly

Now consider the use of 就是 *jiushi* in (32). The second segment in (32) is understood as a reason for the proposition that the speaker had to leave at 7 o'clock, thus the role of 就是 *jiushi* is to instruct the hearer to interpret the second segment as a premise. On this analysis, it seems that 就是 *jiushi* has a similar function to *you know* in (33).

(32) 但是我们每天出发得七点钟就出发**就是**避避免那个八点的早高峰 (N18-C01)

but we everyday leave home must seven o'clock leave home **jiushi** avoid avoid that eight o'clock morning rush hour

*but everyday we have to leave home at seven o'clock **you know** to avoid the morning rush hour at eight o'clock*

(33) (The Spoken BNC2014, SZVB)

A: erm and <name> says it they don't you disturbing a hornet's nest ***you know*** they don't want you stirring up any problems she she said

B: yeah she started to tell me a little bit about <name>

The role of 就是 *jiushi* in (34) differs from the above uses. In (34), the second segment can be interpreted as an implication or conclusion of the propositions given in the preceding utterances. The role of 就是 *jiushi* in (34) is "to constrain the inferential computations that position enters into so that it is understood to be relevant as a contextual implication" of the preceding utterances (Blakemore, 1996, p. 333).

(34) 但是那儿真的很破旧都是流浪狗特别多然后不是真正的路嘛是土路然后也有很多小姐之类的**就是**真的一个很很不好的地方 (L08-C01)

but there really shabby all are homeless dogs very much ranhou is not a real road is dirt road ranhou have many sex workers and the like **jiushi** really a very very not good place

*but that place is really shabby there are lots of homeless dogs and it is not a real road it is a dirt road and there are many sex workers and the like **I mean** it really is not a very good place*

In conclusion, the above examples of the usages of 就是 *jiushi* can be found both in the L1 corpus and the L2 corpus; however, it is impossible to give a comprehensive description of

all these uses of 就是 *jiushi* in one chapter. Given that relevance theory has rarely been employed in Chinese discourse markers analysis, it seems a good start to focus on one main function to find out whether this theory can deepen our understanding of the contribution that 就是 *jiushi* makes in informal conversation. To this end, I will explain the interpretive use of 就是 *jiushi* in some detail within the framework of relevance theory in the next section.

## 6.5 The Interpretive Use of 就是 *Jiushi* in Informal Interaction

### 6.5.1 The Relevance Theoretic Framework

Relevance theory, outlined in Sperber and Wilson (1986b, 1995), is an approach to the study of human communication offering an explicitly cognitive account of utterance interpretation. Within the relevance theoretic framework, human cognitive processes are geared to achieving the greatest possible cognitive effect for the smallest possible processing effort (Sperber & Wilson, 1986b, 1995). This is the so-called Cognitive Principle of Relevance. The theory is founded on another principle as well, which is called the Presumption of Optimal Relevance. This second principle is specifically concerned with communication. According to Sperber and Wilson (1995), utterances are regarded as acts of ostensive communication, that is, they express not only information about something, but also express the speaker's intention to make this information manifest to the hearer. In other words, communication involves the speaker's intention to modify or affect the hearer's cognitive environment in some way (three ways are proposed by Sperber & Wilson, 1995: true contextual implications, warranted strengthening of existing assumptions, and revision of existing assumptions). As such, the Presumption of Optimal Relevance claims that the ostensive stimulus used by the speaker is worth the hearer's attention, and any utterance addressed to the hearer automatically conveys a presumption of its own relevance. Following this analysis, in verbal communication, by the very act of addressing someone, the speaker creates an expectation that the utterance will achieve enough contextual effects to be worth processing for the hearer and cause him/her no unnecessary processing effort (optimal relevance). The hearer's task is to construct a context which will make the utterance worth processing. Relevance, then, is seen as the result of a trade-off between contextual effects and processing costs, and an expectation of optimal relevance is seen as automatically created by utterances (Andersen, 2001; Schourup, 2011).

The task of accounting for the pragmatic functions of discourse markers amounts to the intention to specify the contributions they make to utterance interpretation. It has been demonstrated that relevance theory is sufficient to account for the ways in which discourse markers affect utterance interpretation in communication (e.g., Andersen, 2001; Blakemore, 2002; Jucker, 1993; Schourup, 2011). From the relevance theoretic viewpoint, discourse markers contribute to relevance by operating as signals which tell the hearer how an utterance is to be understood, thus reducing the processing effort that the hearer must employ in utterance comprehension or interpretation (Andersen, 2001). In this study my account rests on the basic assumption that the various uses of 就是 *jiushi* can justifiably be subsumed under a single and precise description of how the marker 就是 *jiushi* contributes to the relevance of utterances in interaction, while I did not attempt to use relevance theory directly to explain relations that 就是 *jiushi* marks. Rather, I found that it was feasible to apply Blakemore's (1993, 1996, 1997) work on reformulations and reformulation markers to the analysis of the discourse marker 就是 *jiushi* in the present study. This relation of reformulation is precisely the very function of 就是 *jiushi* that will be considered in this chapter for the reasons have been given previously in Section 6.4. Accordingly, in the following text I will first discuss Blakemore's discussions on reformulations and reformulation markers, and then I will offer interpretations of examples taken from the spoken Chinese corpus to explain how 就是 *jiushi* as a reformulation marker can contribute to utterance interpretation.

### 6.5.2 Reformulation and Interpretive Resemblance

The notion of *interpretive resemblance* proposed by Sperber and Wilson (1986b, 1995) plays a central role in Blakemore's analyses of reformulations and reformulation markers. Consider the following example (35)-(37) taken from Blakemore (1997):

(35) We will have to let her go.

(36) What did the director say?

(37) a. We will have to let her go.

b. They'll have to let her go.

c. She's fired.

According to Blakemore (1997), all the utterances in (37) can be taken as answers to (36) in a situation in which the director had produced the utterance in (35). It is clear that (37a) is a direct quotation and represents the director's utterance in virtue of resemblances in linguistic and semantic structure. (37b) still shares the common proposition with (35), although it uses the third person pronoun instead of the original first person pronoun. Unlike (37a) and (37b), (37c) has neither the same linguistic and semantic structure nor the same proposition of (35). Nonetheless, within relevance theory, the proposition of (37c) can still be said to resemble that of (35) in the sense that "it is not difficult to imagine a context in which it gives rise to the same contextual implications" (Blakemore, 1997, p. 7). In such cases where the resemblance involves the sharing of at least some logical and contextual implications, Sperber and Wilson (1995) claim that the utterance can be said to be relevant as an interpretation of a proposition or thought, or in other words, the utterance interpretively represents a representation which it resembles in content. This relationship, that is, resemblance in content between representations, is called *interpretive resemblance* (Sperber & Wilson, 1986a, 1986b, 1995). In Sperber and Wilson's analysis (1986b, 1995), a speaker who produces an utterance which is relevant as an interpretation of another utterance can be taken to be creating expectations of faithfulness. Faithfulness is a matter of degree, and the degree of faithfulness is determined by the extent to which the two propositions share logical and contextual implications (Blakemore, 2002). The question of whether an utterance P is intended as interpretation is a question about its explicit content or what Sperber and Wilson (1995) call its high-level explicatures. Blakemore (1996) states that this means that a speaker who produces an utterance which is relevant as an interpretation of a thought communicated by another utterance will be taken to be explicating a proposition of the form in (38):

(38) The speaker believes that P is a faithful representation of a thought Q. (p. 340)

For example, in (39) which is quoted from Blakemore (1997), the use of *that is* can be understood as evidence of the speaker's intention to produce the utterance (39b) whose main relevance lies in the fact that it is a faithful interpretation of the preceding utterance (39a). Given the above account, the use of the reformulation marker *that is*, can be analysed as marking a contribution to the explicit content of (39b). Specifically speaking, the use of a reformulation marker can be taken as "a distinct discourse unit or speech act whose relevance lies in the way it leads the hearer to recover a proposition of the form [(38)] as a higher level explicature of the host utterance" (Blakemore, 1996, p. 340).

(39) a. At the beginning of this piece there is an example of an anacrusis.
b. ***That is***, it begins with a unaccented note which is not part of the first full bar.

This then raises the question of why the speaker produces both the original and the reformulation. According to Blakemore (2002), the reformulation (39b) is produced due to a decision the speaker made in terms of the degree of faithfulness in order to obtain cognitive effects for the minimum processing costs (i.e., optimal relevance). This decision is made on the basis of the speaker's assessment of the hearer's vocabulary, his/her processing resources at the time, or contextual assumptions. At this point, however, how is the hearer to access the intended degree of faithfulness? This question has a bearing on the so-called "consistent with the principle of relevance" in Sperber and Wilson (1986a, 1986b, 1995). Whenever an utterance P is expressed, the hearer takes for granted that some subset of P's logical and contextual implications are also logical and contextual implications of the thought being communicated. The hearer assumes that this subset will have enough cognitive effects to make the utterance P worth his/her attention and at the same time it will cause him/her no unnecessary processing effort. In other word, the hearer aims for an interpretation consistent with these assumptions, i.e., consistent with the principle of relevance. When "this criterion selects a single interpretation (or closely similar interpretations with no significant differences between them), communication succeeds" (Sperber & Wilson, 1986a, p. 163).

Interpretive resemblance is a context-dependent notion: an utterance can be used to represent another utterance which it resemblances in meaning closely (e.g., a paraphrase or translation) in one context or more distantly (e.g., a summary) in another context (Sperber & Wilson, 1995). Recall (37) for example, which is repeated below. Reformulations are not always strict paraphrases; a summary is an example of an utterance which is relevant as an interpretation of another utterance. Consider *in other words* in (40) which is taken from Blakemore (1993), it introduces a summary rather than a paraphrase.

(37) a. We will have to let her go.
b. They'll have to let her go.
c. She's fired.
(40) A: I think it's time you thought about a new career.
B: ***In other words***, I'm fired.

There are phrasal appositions containing expressions such as *this is* and *in other words* which can be interpreted in exactly the same way as the sequences (41)-(42). For instance, the

second segment in (41) which is borrowed from Blakemore (1996) also achieves relevance as a reformulation of the first. Blakemore (1996) argues that this kind of use of *that is* cannot be straightforwardly accommodated in a framework which assumes that discourse markers encode sequential coherence relations, for *that is* in utterances like (42) involves a connection between a parenthetical constituent and its host utterance. The point here is that reformulations can involve parenthetical appositions as well.

(41) They completely clammed up. ***That is***, they refused to speak.

(42) They completely clammed up, ***that is***, refused to speak.

This point is made on the basis of two criteria for apposition proposed by Burton-Roberts (1999). The two central criteria are: (i) elements in apposition should converge in extralinguistic reference, and (ii) they should be capable of being understood as having the same syntactic function with respect to the same other elements in sentence structure. Following the second criterion, "any sentence containing an apposition of sentence constituents can be expanded into an apposition of full sentences without change of meaning" (Burton-Roberts, 1999, p. 26). For example, on this account, (41) is the expanded version of (42).

It seems that Blakemore's work on reformulations and reformulation markers makes it feasible to analyse 就是 *jiushi* along similar lines. In what follows, I shall discuss the reformulation relation 就是 *jiushi* marks in the spoken Chinese corpus.

### 6.5.3 Reformulation Marker 就是 *Jiushi*

While acknowledging Blakemore's analyses of reformulation markers provide useful insights into the present study of the marker 就是 *jiushi*, her arguments are to a large extent based on constructed examples. On the basis of spoken corpora, it is not surprising that the use of reformulation markers, such as *that is* and *in other words*, may be far more complex than that illuminated in Blakemore's work. In this study, within the relevance theoretic framework, I conclude that 就是 *jiushi* as a reformulation marker signals that the utterance following 就是 *jiushi* is an alternative means for communicating what is communicated by the preceding utterance.

In the spoken Chinese corpus, it is evident that 就是 *jiushi* can be used in discourse sequences. For example, in (43), the first segment and the second one containing 就是 *jiushi* share a common proposition, that is, the speaker's son can get help from a Chinese teacher. In this case, the use of 就是 *jiushi* performs a distinct speech act which communicates the proposition that what the second utterance communicates is communicated by the preceding utterance. On this account, it is reasonable to say that utterances like (43) are paraphrases as reformulations. Similarly, this relationship indicated by 就是 *jiushi* can also be recognised in (44). The first segment in (44), i.e., 'do not allow their children to leave them' and the segment containing 就是 *jiushi* 'do not want to separate from their children', share a common proposition.

(43) 但当时还好有一个这个中文的老师在那儿所以可以帮助他稍微辅助一下**就是**如果他困难的话可以去直接用中文去求助 (N07-C01)

but at that time fortunately there was one zhege Chinese teacher over there so can help him a bit assist **jiushi** if he had problems can go directly use Chinese to ask help

*but at that time fortunately there was a Chinese teacher in the kindergarten so (the teacher) could help him **you know** if he had problems he could ask for help in Chinese*

(44) 然后孩子长大了之后就是他们他们就是不让 TA 走**就是**不让 TA 和她分开对这个很不健康嘛 (L05-C02)

ranhou their children grew up after that jiushi they they jiushi do not allow TA to leave **jiushi** do not allow TA separates from her yes this is unhealthy

*then after their children grew up they they do not allow their sons/daughters to leave them **I mean** they do not want to separate from their children yes this (kind of relationship between parents and children) is unhealthy*

In the spoken Chinese corpus, there are a few occurrences of 就是 *jiushi* in translations. For example, the phrase 国际课程 *guoji kecheng* following the second 就是 *jiushi* in (45) is the Chinese translation of 'International Baccalaureate'. The first 就是 *jiushi* in (45) has a clear syntactic role and should be identified as a predicate. As regards the role of the second 就是 *jiushi* in (45), it should be labelled as a discourse marker which signals that the following utterance reformulates the utterance proceeding 就是 *jiushi*. Likewise, the boldface 就是 *jiushi*

in (46) links ‘standard spelling’ with its Chinese translation 标准的拼写 *biaozhun de pinxie*. So utterances containing 就是 *jiushi* like (45) and (46) are translations as reformulations.

(45) IB **就是** International Baccalaureate **就是**国际课程 (N10-C01)

IB **jiushi** International Baccalaureate **jiushi** guoji kecheng

*IB **is** International Baccalaureate **you know** International Baccalaureate*

(46) 啊对还有就是英文的那个 standard spelling **就是**它那个标准的拼写也是一样的功能 (L05-C02)

uh yes and jiushi English neige standard spelling **jiushi** its neige standard spelling is also the same function

*uh yes and the standard spelling in English **you know** its standard spelling has the same function as Chinese*

Furthermore, according to Blakemore (1997), the proposition of the second segment in (47) may still be said to resemble the proposition of the first segment because in this case the resemblance involves the sharing of some logical and contextual implications. The speaker in (47) was using an alternative means to restate her claim that Dutch was useless to her. This analysis of 就是 *jiushi* is similar to Blakemore’s analysis of *in other words*. Following Blakemore (1996), I argue that the use of 就是 *jiushi* communicates the proposition that the utterance following 就是 *jiushi* is an alternative means for communicating what is communicated by the preceding utterance.

(47) 回国之后我觉得荷兰语的话也不会用到**就是**需求量比较小 (N14-C01)

return to China after I think Dutch cannot be used **jiushi** the need is relatively small

*if I come back to China I do not think I will use Dutch **you know** there is no need for Dutch*

Reformulations can also be involved in parenthetical appositions. In nominal appositions, the parenthetical can be interpreted as providing an alternative means of reference. Consider (48)–(50) for example. The underlined segments 舆论的压力 ‘the public opinion pressure’ in (48), 研究中文研究中文方面的 ‘people who study Chinese’ in (49), and 综合课的老师 ‘teachers who teach the comprehensive courses’ that follow 就是 *jiushi* in (50) are parentheticals.

(48) 国内的压力相对小一点**就是**舆论的压力(N11-C01)

domestic pressure relatively small a bit **jiushi** public opinion pressure

*people have relatively less pressure in China **you know** the pressure of public opinion*

(49) 哦对就是在那个<university>就是有有多少像你这样的人**就是**研究中文研究中文方面的 (L02-C02)

Oh yes jiushi at neige <university> jiushi there are how many people like you **jiushi** study Chinese do Chinese-related work

*oh yes at <university> jiushi how many people just like you to **I mean** study Chinese to do Chinese-related work*

(50) 因为老师**就是**综合课的老师会告诉我说他们综合课已经讲了这个语法点 (N18-C01)

because teachers **jiushi** teachers of the comprehensive courses can tell me that they comprehensive courses already taught this grammar

*because teachers **you know** teachers who teach the comprehensive courses told me that they already taught this grammar in the comprehensive courses*

Following Blakemore (1996), 舆论的压力 'the public opinion pressure' in (48) is referentially equivalent to 压力 'the pressure', and this proposition that 'pressure refers to the public opinion pressure' is being communicated by using 就是 *jiushi*. Likewise, the proposition that 综合课的老师 'teachers who teach the comprehensive courses' is referentially equivalent to 老师 'teachers' in (50) is being communicated by using 就是 *jiushi*. Given that, the propositions communicated by the parentheticals in (48)–(50) are not about 舆论的压力 'the public opinion pressure', 研究中文研究中文方面的 'people who study Chinese', and 综合课的老师 'teachers who teach the comprehensive courses', but rather about the means of referring to these three segments. Therefore, the use of 就是 *jiushi* in the parentheticals in (48)–(50) performs an alternative means of reference. However, what is the role of the parenthetical in the relevance of the utterance? In Blakemore's (1996) analysis, to produce the utterance in (48), for example, would be to communicate and at the same time guarantee the relevance of the proposition that 舆论的压力 'the public opinion pressure' is a reformulation of the referential expression 压力 'the pressure'. On this account, the proposition communicated by the parenthetical constituents are not about the explicit content of the host clause as such "but

the REPRESENTATION of its explicit content" (Blakemore, 1996, p. 345; emphasis as original): the parenthetical constituents following 就是 *jiushi* aid the hearer in reference assignment. As such, the speaker in (48) would successfully lead the hearer to interpret 压力 'the pressure' as 舆论的压力 'the public opinion pressure' during the conversation.

As has been discussed in 6.5.2, the degree of faithfulness in interpretive resemblance varies from situation to situation. According to the principle of relevance, a less-than-literal interpretation of the speaker's thought may enable the speaker to achieve the same contextual effects but for less processing effect. For example, the utterance that contains 就是 *jiushi* in (51) is presented with a summary of it. The analysis of the use of 就是 *jiushi* in (51) is similar to the analysis of *in other words* in (40). To reiterate, a summary is an example of an utterance which is relevant as an interpretation of another utterance. The use of 就是 *jiushi* in (51) provides an explicit signal that the second segment should be construed as a faithful interpretation of the preceding utterance.

(51) 很多人就说诶为什么你有一个食物啊然后你有这种国旗和旗杆**就是**很为什么很不一样(L11-C01)

many people say ei why you have a food then you have this kind national flags and flagpoles **jiushi** very why different

*many people ask me ei why do you sell food and (at the same time) you also sell national flags and flagpoles **you know** why you do two different businesses*

(40) A: I think it's time you thought about a new career.

B: ***In other words***, I'm fired.

To sum up, the relevance theoretic approach explains that the marker 就是 *jiushi* signals an alternative means for communicating what is communicated by the preceding utterance in serval ways. Although the discussion is mainly based on the L1 use, it is clear that there are L2 utterances that contain 就是 *jiushi* that have the same contributions as the L1 uses to utterance interpretation. Based on the above theoretical analysis, in what follows, I will give a clear picture of the frequency and distribution of the reformulation function of 就是 *jiushi* both in the L1 corpus and the L2 corpus.

## 6.6 Comparative Analysis of the L2 Use of the Marker 就是 *Jiushi*

This section gives a quantitative analysis of the L2 use of the reformulation marker 就是 *jiushi* in comparison to the L1 use. First of all, the frequency of the interpretive use of 就是 *jiushi* was examined in each corpus. Table 25 provides information on the raw frequencies and relative frequencies (which are normalised to 1,000 words) of the marker 就是 *jiushi* both in the L1 corpus and the L2 corpus. The relative frequency represents that 就是 *jiushi* in this function occurs one time per 1,000 words of speech in the L1 corpus, while it occurs less than one time per 1,000 words in the L2 corpus. Overall, the L1 speakers tend to use 就是 *jiushi* twice as often as the L2 speakers.

**Table 25**

*Frequencies of the Marker 就是 Jiushi in the Spoken Chinese Corpus*

| Frequency | L1 use | L2 use |
|---|---|---|
| Raw frequency | 245 | 112 |
| Relative frequency | 1.073121 | 0.507265 |

*Note*. The raw frequency and relative frequency of 就是 *jiushi* in each text are given in Appendix N.

Although this difference between the L1 and L2 group in frequency of the use of 就是 *jiushi* is obvious, it should be cautioned however that it is not necessarily the case that L1 speakers of Chinese prefer using 就是 *jiushi* to signal an alternative means for communicating what is communicated by the preceding utterance, while L2 speakers of Chinese seem to fail to acquire this function of 就是 *jiushi*. As has been noted previously, interlocutors in the L1 corpus were friends and acquaintances, while interlocutors in the L2 corpus were strangers. This being the case, there is a possibility that the L2 speakers used 就是 *jiushi* less often than the L1 speakers not only because they were non-native speakers of Chinese, but also because they were engaged in conversations with a stranger. The effect of interlocutors' relationship on the use of 就是 *jiushi* is mostly a reasonable guess on the basis of the difference in corpus design. Furthermore, since a range of topics were involved in both the L1 corpus and the L2 corpus, and some of the topics were not shared by both groups, it is possible that the L2

speakers would not use 就是 *jiushi* exactly as the L1 speakers used it in terms of exactly the same topics and that the frequency difference may be due to different topic frequencies (Gries & Deshors, 2014). These aspects may be the possible reasons for the distributional differences observed in this thesis, and can be investigated further by adopting more sophisticated statistical methods. While the investigation of the use of 就是 *jiushi* in this chapter remains at a purely descriptive level, factors that may affect the L2 use will not be discussed in this thesis.

In corpus quantitative analyses, frequency data should be augmented with information on the dispersion of the items in question (Gries, 2008). Dispersion can tell researchers about the distribution of the items in question throughout the corpus (Brezina, 2018). A very basic and rather crude measure of dispersion is *range* (*r*), which shows researchers the number of corpus parts containing the items under examination. Range is formally expressed as (Brezina, 2018, p. 48):

$$range = no.of\ parts\ with\ word\ w\ (or\ phrase\ p)$$

As Table 26 shows clearly, the range for 就是 *jiushi* in this function in the L1 corpus is 21, because there are 21 texts in the L1 corpus in which 就是 *jiushi* appears. This can be expressed as: $r_1 = 21$. Accordingly, the range for 就是 *jiushi* in the L2 corpus is 28, i.e., $r_2 = 28$. We then can say that 80.8% of the L1 texts include 就是 *jiushi* when it functions as a discourse marker representing an alternative means for communicating what is communicated by the preceding utterance, while 82.4% of texts contain that use of 就是 *jiushi* in the L2 corpus. However, range is not a very good measure for quantifying the amount of dispersion throughout the corpora as it disregards the frequencies of 就是 *jiushi* in each text. This measure therefore is not efficiently sensitive to show whether 就是 *jiushi* is evenly spread across the entire corpus or distributed mainly in a few texts.

**Table 26**

*Distribution of the Marker 就是 Jiushi in the Spoken Chinese Corpus*

| Distribution | L1 text | L2 text |
|---|---|---|
| 就是 *jiushi* | 21 | 28 |
| without 就是 *jiushi* | 5 | 6 |
| Total | 26 | 34 |

In this chapter, I employed an alternative method *DP* (for deviation of proportions) to measure the dispersion of the interpretive use of 就是 *jiushi* in both the L1 corpus and the L2 corpus and then compared the differences. This measure is proposed by Gries (2008); to exemplify this measure in a simple way, I directly have taken one example used in Gries (2008, p. 416), and repeated the explanation of *DP* as follows:

> Imagine a corpus consisting of three 200-word parts, i.e. 600 words. Imagine further one is interested in a word *a* that occurs 9 times in the corpus, 3 times in each of the three corpus parts. In this case, the computation of the three steps can be summarized as in Table 2 [i.e., Table 27 in this thesis]. Step 1 results in the leftmost column: if *a* is distributed as one would expect given the sizes of the *n* corpus parts, *a*'s frequency in each file should be one third of its overall frequency in the corpus: 200/600=0.33. Step 2 results in the second column from the left: in each row, i.e. for each corpus part, 3/9=0.33. Step 3 requires to compute the *n* row-wise absolute differences (shown in the third column), sum them up (shown in the fourth column), and divide by 2; the result is DP. The result in the rightmost column shows that *a* is distributed perfectly evenly in the corpus, namely in exact accordance with how the corpus parts look like. (Gries, 2008, pp. 415-416)

**Table 27**

*Computation of DP*

| **Step 1** | **Step 2** | **Step 3** | | |
|---|---|---|---|---|
| Expected % | Observed % | Abs. difference | Sum of abs. diff. | Divide by 2 |
| 0.33 | 0.33 | 0 | | |
| 0.33 | 0.33 | 0 | 0 | 0 |
| 0.33 | 0.33 | 0 | | |

The value of *DP* is a number between 0 and 1, with 0 representing that the item under examination has perfectly even distribution in the corpus, and with 1 signifying that the item has extremely uneven distribution in the corpus. Based on this explanation, the *DPs* of 就是 *jiushi* in the two corpora are given in Table 28. The values of *DP* indicate that 就是 *jiushi* has an uneven distribution both in the L1 corpus and the L2 corpus, which, however, are closer to 0 than to 1. Since the value of *DP* for 就是 *jiushi* in the L2 corpus is only slightly higher than that in the L1 corpus, we can say that 就是 *jiushi* is dispersed in a similar way in each corpus.

**Table 28**

*DP for 就是 Jiushi in the Spoken Chinese Corpus*

| DP of 就是 *jiushi* | Sum of abs. diff. | Divide by 2 |
|---|---|---|
| L1 use | 0.607293 | 0.303647 |
| L2 use | 0.643811 | 0.321906 |

*Note*. Detailed calculation steps and quantitative information are given in Appendix N.

In conclusion, although the result shows that the L1 speakers tend to use 就是 *jiushi* more frequently than the L2 speakers in this study, 就是 *jiushi* has similar distribution in each corpus. Due to space limitations, this study did not take into consideration the possible factors that influence the use of 就是 *jiushi* in interaction. It is expected that a further scientific quantitative analysis of the use of 就是 *jiushi* can be conducted in the future.

## 6.7 Summary

This chapter has provided an initial investigation of the use of the marker 就是 *jiushi* within the framework of relevance theory. Though this chapter has reflected on a very restricted range of utterances that contain 就是 *jiushi*, it shows that relevance theory can provide a sufficient explanatory framework for the study of the marker 就是 *jiushi* by making use of corpus data. It then concluded that 就是 *jiushi* as a reformulation marker represents an explicit signal that the segment following 就是 *jiushi* can be entertained as a representation of the first segment which it resembles. Methodologically, it first investigated the frequencies of the use of 就是 *jiushi* in each group, and found that the L1 speakers tend to use 就是 *jiushi* more frequently than the L2 speakers in this study. Then, I adopted the measure *DP* to examine the dispersion of 就是 *jiushi* in each corpus. It turned out that 就是 *jiushi* has similar distribution in each corpus. Although this chapter focused exclusively on the interpretive use of 就是 *jiushi* in the corpus, and the investigation of 就是 *jiushi* remained mainly at the descriptive level, it seems to me that both the theoretical and statistical approaches employed in this study can provide

valuable insights into studies on 就是 *jiushi* and be extended to research on other Chinese discourse markers.

# Chapter 7 Discussion

## 7.1 Introduction

The preceding chapters have provided a detailed account of the design, compilation, as well as the research potential of the spoken Chinese corpus. This chapter primarily discusses in what ways the work in the preceding chapters was addressed the research questions proposed in Chapter 1. To begin with, the first research question that is given as follows focused on spoken corpus design and construction, which has been investigated in detail in Chapters 3 and 4. Although the L2 corpus was the main focus of attention, for the sake of the comparability of the L1 corpus and the L2 corpus, Chapters 3 and 4 have provided detailed descriptions of the design and compilation of both corpora. In the practice of corpus design and compilation, different tasks were assigned to the two corpora which corresponded to the sub-questions of the first research question. Comprehensive discussions with respect to the compilation of the two corpora will be given in Sections 7.2 and 7.3.

RQ1: What are the main considerations in building a spoken Chinese corpus of informal interaction?

(i) What are the design features in creating a spoken L2 Chinese corpus of L1–L2 conversational interaction?

(ii) Which strategies should be employed to compile a spoken L1 corpus to ensure that it will be comparable to the spoken L2 corpus as far as possible?

The reliability of the investigation depends crucially on the quality of the data. So, the second research question repeated below was concerned with the comparability of the L1 corpus and the L2 corpus. Chapter 5 has compared the components of the L1 corpus and the L2 corpus by examining the results of the decisions and compromises made during the creation of the two corpora. The aim of Chapter 5 was to examine the ways in which the two corpora were comparable and the possible reasons behind the differences between them. The findings of the comparison will be discussed carefully in Section 7.4.

RQ2: To what extent is the spoken L1 corpus comparable to the spoken L2 corpus?

The last research question that is given below reflected the research potential of the spoken Chinese corpus, which has been addressed in Chapter 6. By adopting the relevance theoretical approach, Chapter 6 has investigated the pragmatic role of the marker 就是 *jiushi* in interaction. In addition to the theoretical analysis of the function of 就是 *jiushi*, it has also investigated the frequency and dispersion of the occurrences of 就是 *jiushi* used with this function in both the L1 corpus and the L2 corpus. Both the qualitative and quantitative analyses and the results of the L2 use of 就是 *jiushi* will be reviewed in Section 7.5.

RQ3: What are the features which characterise the L2 use of the Chinese discourse marker 就是 *jiushi* in the L1–L2 interactions?

(i) What is the role of 就是 *jiushi* when it is used as a reformulation marker in conversational interaction?
(ii) How frequent is the marker 就是 *jiushi* among L1 and L2 production?
(iii) How does the distribution of the marker 就是 *jiushi* differ between the two subcorpora?

At the end of this chapter, I also consider the implications of the spoken Chinese corpus as well as of the findings of the corpus analysis of the L2 use of 就是 *jiushi* in Section 7.6. This chapter closes with a brief summary.

## 7.2 Key Considerations in Building the Spoken L2 Corpus of L1–L2 Interaction

The first research sub-question this thesis seeks to address is about the compilation of the spoken L2 corpus. Consistent with previous corpus practices, when building the L2 corpus consisting of L1–L2 conversational interaction, variables regarding the L2 speakers of Chinese, the L1–L2 interaction, and transcription were considered and documented in Chapters 3 and 4. Following the discussions given in the previous chapters, Table 29 summarises the features of the spoken L2 corpus. In this section, I shall give an account of how these variables in terms of features were considered in practice to address the first research sub-question.

**Table 29**

*Features of the Spoken L2 Corpus*

| Aspects | Features |
|---|---|
| L2 speakers | native English speakers<br>non-Chinese ethnicity<br>educated adults<br>at intermediate to advanced proficiency level<br>female and male speakers |
| L1–L2 interaction | the researcher as the L1 speaker<br>two speakers in each interaction<br>L2 speakers participated in two or more interviews<br>the length of each interview varies<br>all interviews were recorded in audio formats |
| Transcription | orthographical transcription<br>recordings will not be publicly available<br>a bespoke transcription scheme |

### 7.2.1 L2 Speakers of Chinese

The spoken L2 corpus comprises 220,792 words, produced by only 14 L2 speakers of Chinese across 34 transcribed recordings. Table 30 shows the number of speakers and number of words in each group of the three demographic categories in this corpus: nationality, age, and gender.

**Table 30**

*Speaker and Word Counts for the Demographic Groups in the Spoken L2 Corpus*

| Category | Group | No. of speakers | No. of words |
|---|---|---|---|
| Nationality | New Zealand | 12 | 196,667 |
| | Australia | 2 | 24,125 |
| Age | 20-25 | 6 | 78,503 |
| | 26-30 | 2 | 23,301 |
| | 31-35 | 3 | 84,307 |
| | 36-40 | 3 | 34,681 |
| Gender | female | 3 | 55,804 |
| | male | 11 | 164,988 |

Table 31 presents the criteria that have been considered in recruiting the L2 speakers of Chinese in this study. In what follows, I will discuss how these criteria were employed when building this corpus.

**Table 31**

*Considerations for Recruiting the L2 Speakers of Chinese*

| Factors | Selection criteria | Descriptive criteria |
|---|---|---|
| Speakers | native English speakers<br>non-Chinese ethnicity<br>educated adults<br>females and males<br>20-50 years old<br>speak *Putonghua*<br>at the intermediate to advanced proficiency levels | New Zealanders or Australians<br>living in New Zealand or mainland China<br>learners or speakers of Chinese<br>duration in China<br>relationship with the researcher<br>occupations |

First of all, this L2 corpus was designed purposely to cover L2 speakers of Chinese who were native English speakers of non-Chinese ethnicity. Heritage speakers of Chinese are certainly valuable contributors to spoken L2 corpora, and as such it is important to treat them as a separate group when designing a spoken L2 Chinese corpus. As such, the effects of the variable in terms of being heritage speakers on L2 production can be evaluated efficiently. Although the target L2 participants focused on New Zealanders, it was decided to include Australians who met the selection criteria as well. It then turned out that two cultural backgrounds are recognised from the point of view of the national affiliation: 12 New Zealanders and two Australians. Involving two different cultural backgrounds in this L2 corpus is a reflection of the compromise made in the corpus creation practice in order to recruit enough L2 speakers of Chinese. Due to the limited sampling size, it is less possible to provide a view on the potential influences of cultural backgrounds on L2 use of Chinese in this thesis. In addition to these criteria, the L2 corpus was expected to be a gender-balanced corpus. The gender information was gathered and recorded by me on the basis of their biological aspects of identity. Even though I still used the male/female binary tradition in this thesis, it is arguably more appropriate to allow participants to describe their gender beyond this binary fashion. Additionally, during the procedure of data collection, I made little effort to achieve the balance of the L2 corpus in gender because of the difficulties of recruiting L2 participants. It then turned out that the ratio of male participants to female participants is 4:1. The higher proportion of the male participants in the corpus is not necessarily a reflection of the higher number of males than females who are proficient speakers of Chinese in reality. Rather, the imbalance between the male and female participants should be seen as a ramification of the limited size of the L2 corpus.

The most challenging task with respect to the L2 participants in the design of the spoken L2 corpus was to determine their proficiency levels. It was intended that all the L2 participants would be at intermediate to advanced proficiency levels. In this study, methods used to assess L2 proficiency in corpus design, on Thomas's (1994) terms, primarily include impressionistic judgement, institutional status, and standardised tests. Since some L2 participants were recommended by their Chinese teachers, their proficiency levels were evaluated by the teachers on the basis of the L2 participants' institutional status, such as the number of years of exposure to Chinese at university. When I talked with certain L2 participants who were treated as intermediate L2 speakers of Chinese by their Chinese teachers, it turned out that not only did they rely heavily on English to clarify their ideas, but they also had difficulties understanding my questions. Interestingly, these L2 participants were aware of their proficiency and reminded me before the interviews that they did not qualify themselves as the target L2 speakers of Chinese that I needed. This result at least indicates that caution should be exercised when using institutional status to assess proficiency of L2 participants in corpus practice.

In addition to the measure of institutional status, certain L2 participants were assessed at advanced levels by their Chinese teachers, because they had passed the high level of the HSK test (i.e., HSK5 and HSK6). As has been discussed in Chapter 3, the scores of the HSK test are more likely to represent L2 speakers' writing and reading abilities, but are questionable if used to assess their oral skills. Chapter 2 has mentioned that the HSKK test is developed to test L2 speakers' speaking skills; however, it is not as popular as the HSK test among L2 speakers of Chinese. In practice, none of the L2 participants in this study had taken the HSKK test or any oral tests. As a result, it would be even harder to recruit enough L2 speakers of Chinese if adding the scores of the HSKK test as one criterion for the target participants. More importantly, the reason that I did not employ the measure of standardised tests to assign L2 proficiency levels was that it was difficult to decide to what extent the scores of the HSKK test or any oral tests could provide reliable assessment of L2 speakers' proficiency in real-life situations. That is, the HSSK test or any oral tests may not give fine-tuned information about whether L2 speakers have specific kinds of knowledge of spoken Chinese. Additionally, in the spoken L2 corpus, although some L2 participants had not taken any standardised oral tests, they confidently justified themselves as intermediate or advanced speakers of Chinese. It turned out that communications with these L2 participants went quite smoothly. Although the measure of impressionistic judgement tends to be seen as one's causal evaluation (Thomas, 1994), in my case, it in fact was more useful than the other measures. As a native Chinese speaker and

the only interviewer in the interviews, I talked with all the L2 participants which enabled me to compare their proficiency and to make relatively objective judgments. Undeniably, my reflection on L2 participants' proficiency to some extent was still subjective and impressionistic, but there is certainly reason to believe that it is beneficial to be reminded that, since L2 proficiency is a fuzzy variable, there is no need to undermine our ability as native speakers and researchers to assess L2 participants' proficiency when creating spoken L2 corpora.

To sum up, we do not know enough about L2 proficiency: this fact provides the impetus for much work to be done in this field. Our techniques for investigating the nature and content of proficiency are still crude, but it is only by putting such techniques into practice that it will be possible to improve upon them in the future. However, one comment that I want to share is about the necessity of using L2 proficiency to select appropriate L2 participants for any spoken L2 corpus. When recruiting L2 participants for the L2 corpus, one unexpected feature was that most of the L2 speakers in this thesis were not L2 learners but L2 speakers in the sense that they were not learning Chinese at any institution but using Chinese in daily life. Therefore, in some cases, when an L2 speaker of Chinese does not qualify him/herself as a learner of Chinese, and uses Chinese every day or quite often in daily life, do we need to ask ourselves whether we can find an alternative way to describe their capabilities in communication, rather than keep characterising them as less proficient speakers whose oral skills need to be assessed as intermediate or advanced? It is time to think thoroughly about whether we instinctively compare L2 participants to some idealised L1 speakers when assessing their proficiency and overvalue the communicative abilities of L1 speakers.

### 7.2.2 L1–L2 Interaction

By using the unstructured interviewing method, 34 transcribed recordings of L1–L2 informal conversational interactions were conducted for the spoken L2 corpus. Variables considered when gathering the L1–L2 interactions for the spoken L2 corpus are presented in Table 32. In this spoken L2 corpus, the features of the L1–L2 interactions will be discussed in order.

**Table 32**

*Variables Considered During the Collection of the L1–L2 Interactions*

| Factor | Conducted criteria | Uncontrolled variables |
|---|---|---|
| Interviews | L1–L2 interaction<br>avoid to use English as far as possible<br>two or three interviews per person<br>record all the interviews in audio formats | topics<br>the length of each interview<br>interviewing mediums<br>the location of each interview |

The L1–L2 interactions had two speakers, and I as the interviewer and the researcher conducted all the conversational interactions[56]. One advantage of using the same interviewer in each interaction is that it is feasible to observe the effects of the interviewer's personal style. Another advantage is, as has been noted in 7.2.1, that being involved in each conversation enabled me to make rather reliable judgements of the proficiency levels of the L2 participants. Although the unstructured interviewing method was adopted to gather spoken data, some interviews were conducted in a rather casual way. Some L2 participants preferred to have more interaction with me during the interviews; conversely, some were very active, and the conversation tended to be more monologic. Inevitably, my contribution varied considerably in different conversations. In the interviews, I was not concerned too much about the varied degree of interactivity between the interlocutors, so each interview was conducted as naturally as possible.

With the data collecting goal in mind, I prompted the L2 participants and made efforts to encourage them to avoid communicating in English as far as possible. It was acceptable that L2 participants used a few English words on some occasions during the conversations, such as asking me how to say certain English words in Chinese, or pet phrases (e.g., *yeah*) were used. Nevertheless, there were no stringent controls on the use of English during the interviews. In practice, I did not remind the participants explicitly or implicitly to speak Chinese when they transferred Chinese to English suddenly in order to express their ideas more clearly. As a result, a few transcripts include some English sentences which account for a very limited part of the texts. On the basis of the experience of collecting L2 speech, it seems to me that L2 speakers were able to avoid depending on English if they were proficient speakers of Chinese. Therefore,

[56] There was one interaction that had two L1 speakers of Chinese and one L2 participant.

the transcripts containing English sentences to some degree indicate the proficiency levels of those L2 speakers of Chinese.

All the L1–L2 interactions were recorded in audio formats, as the spoken L2 corpus was designed to be a monomodal corpus rather than a multimodal corpus. Both monomodal and multimodal corpora have abilities to meet specific research needs and are valuable sources to the research community. As a monomodal corpus, the quality of the recordings was a matter of concern in this thesis. To ensure the quality of the recordings were as good as possible, I made efforts to conduct and record all the interviews in quiet environments, but L2 participants were not required to choose a quiet place to attend the interviews. As a result, a few recordings had caught external noises which made the transcription a bit harder.

Another important feature of the L1–L2 interaction was that the L2 participants and I were unknown to each other until I contacted them for the interviews. It was not part of the intentional design to recruit L2 speakers of Chinese who were strangers to me. Rather, it was a direct reflection of the reality that the interlocutors in the L1–L2 interactions had never met each other before this study. The importance of emphasising this kind of relationship lies in the fact that the L1 participants and I were friends and acquaintances, so in this regard, the L1 corpus is distinct from the L2 corpus.

There are other variables which were not controlled when collecting the data. Ideally, each L2 speaker of Chinese should contribute about the same amount of speech, so no single L2 speaker's speech idiosyncrasies skew the data. However, it was realised that finding L2 speakers of Chinese was a challenging and time-consuming task. To gather as much data as possible, the length of each interview was not controlled, and the L2 participants were encouraged to participate in two or more interviews if they wished to. In consequence, the L1–L2 interactions in the L2 corpus varied in length. This result was reasonable and acceptable, as it was a reflection of the nature of real conversations. Furthermore, topics were not prepared before the interviews, though I did suggest some topics that the participants might be interested in or that related to their experience (e.g., your experience in China). Specific topics might have influences on vocabulary in L2 production. Given that it was intended to gather informal L2 speech, the wide range of topics was a positive feature of the L2 corpus.

To sum up, the L1–L2 conversational interactions are more informal in comparison with the LINDSEI which has been discussed in Chapter 2. Although the L1–L2 interactions gathered in this spoken L2 corpus differ from the spontaneous conversations in daily life, it is

my conviction that a step forward has been taken to compile a spoken L2 corpus of naturally occurring L1–L2 conversational interaction in the future.

## 7.3 The Spoken L1 Corpus: Maximising Representativeness and Comparability

The spoken L1 corpus comprises 228,306 words, produced by 22 L1 speakers of Chinese across 26 transcribed recordings. Table 33 shows the number of speakers and number of words in each group of the three demographic categories in this corpus: region, age, and gender.

**Table 33**

*The Number of Speakers and Word Counts for the Demographic Groups in the L1 Corpus*

| Category | Group | No. of speakers | No. of words |
|---|---|---|---|
| Region | northern China | 17 | 170,335 |
| | southern China | 5 | 57,971 |
| Age | 18-25 | 3 | 24,025 |
| | 26-30 | 11 | 132,567 |
| | 31-35 | 6 | 57,775 |
| | 36-40 | 2 | 13,939 |
| Gender | female | 8 | 142,596 |
| | male | 14 | 85,716 |

Although ensuring comparability with the L2 corpus was an important consideration when building the spoken L1 corpus, another crucial goal of the spoken L1 corpus was to make it possible to say something about the L1 speech of informal conversation. In this section I will clarify to what extent the decisions and compromises taken during the compilation ensured the achievement of the representativeness and comparability of the spoken L1 corpus.

### 7.3.1 L1 Speakers of Chinese

The L1 participants in the spoken L1 corpus were educated adult native Chinese speakers who came from mainland China and were either monolinguals or bilinguals at the time of interviewing. Table 34 below shows the variables considered when recruiting the target L1 participants for this corpus.

**Table 34**

*Considerations for Recruiting the L1 Speakers of Chinese*

| Factors | Selection criteria | Descriptive criteria |
|---|---|---|
| **Speakers** | native Chinese speakers<br>educated adults<br>females and males<br>20-50 years old<br>speak *Putonghua* | living in New Zealand or mainland China<br>relationship with the researcher<br>occupations |

The selection criteria for L2 participant recruitment, namely speak *Putonghua*, gender, age, and educational level, were also considered in the recruitment of L1 participants in corpus design. Consistent with L2 participant recruitment, the current residential countries of the L1 participants were not limited to China or New Zealand. It turned out that three L1 participants had been studying/living in New Zealand for over two years at the time of data collection. In terms of participant recruitment, it seems reasonable to claim that the L1 corpus is comparable to the L2 corpus. With a retrospective look at the L1 corpus design, one variable relevant to L1 participants that was disregarded in design but then exposed its effect in the subsequent analysis is the prior acquaintance with the L1 participants. The relationships between participants and me had not been taken into consideration when designing the L1 corpus. To gather spoken data as readily as possible, I invited my friends and acquaintances to participate in this study. In this regard, the L1 corpus differs from the L2 corpus in which the participants and me had never met or talked to each other before I invited them to take part in this study.

Inevitably, the goal of being comparable in L1 corpus design had to be tempered by the realities during the procedure of the L1 participant recruitment and data collection. Due to time limitations, the tasks of recruiting L1 and L2 participants were conducted simultaneously at the beginning, and gathering enough data as soon as possible was the main concern at that time. As a result, no effort was made to control the gender ratio of males and females nor to balance the number of participants in the age groups in the L1 corpus. In conclusion, the L1 participants were expected to be sampled with the similar criteria as the L2 participants in corpus design; however, compromises were taken in practice which to some extent made the L1 corpus differ from the L2 corpus in several aspects. The comparability between the L1 corpus and the L2 corpus will be discussed in Section 7.4.

### 7.3.2 L1–L1 Interaction

Table 35 shows considerations for gathering the L1–L1 informal interactions. To make the two corpora as comparable as possible, the L1–L1 interactions were carried out in a similar way to the L1–L2 interactions. In practice, however, it was difficult to find and keep a reasonable balance between corpus representativeness and comparability when gathering data for the spoken L1 corpus. In the following, I will give a brief account of the dilemmas in terms of comparability and representativeness encountered in the process of L1–L1 data collection.

**Table 35**

*Considerations for Collecting the L1–L1 Interactions*

| Factor | Criteria | Uncontrolled variables |
|---|---|---|
| **Interviews** | L1–L1 interaction | topics |
| | one interview per person | the length of each interview |
| | the unstructured interviewing method | interviewing channels |
| | record all the interviews in audio formats | the locations of interviews |

To achieve corpus comparability, ideally, the two corpora should have an equal number of participants. However, in practice, the L1 speakers of Chinese produced more words than the L2 participants in the interviews. Given that the L1 corpus was intended to be similar to the L2 corpus in size, the approach of featuring L1 and L2 speakers in equal numbers in the spoken Chinese corpus would impede the representativeness of the L1 corpus, because a very limited number of L1 participants would be needed if, as intended in the original design, each of them attended two or three interviews. This being the case, it was decided to maximise the representativeness of the spoken L1 corpus as much as possible by recruiting more L1 speakers of Chinese to attend the interviews. With reference to the total number of interactions in each corpus, it seems that the L1 corpus does not match the L2 corpus perfectly. Additionally, the L1–L1 interaction contains a range of topics which were not identical to the topics contained in the spoken L2 corpus. This decision to some extent led to the achievement of representativeness of the L1 corpus at the expense of comparability.

Without taking into consideration the relationship between the participants and me, the L1 corpus is more representative of interaction between friends or acquaintances. It should be stressed that this difference was not a ramification of prioritising the corpus representativeness over comparability. Rather, at the time of data collection, it was merely a strategy made to

gather enough L1 speech as soon as possible. It yet unexpectedly became one factor that impedes the achievement of corpus comparability.

To sum up, in the practice of building the spoken L1 corpus, many decisions were made to achieve representativeness of the spoken L1 corpus rather than achieve comparability. These decisions reflect the compromises that had been made between what is theoretically desirable and practical constraints. In the following section, I will discuss to what extent the L1 corpus is comparable to the L2 corpus.

### 7.3.3 Transcription

To make the spoken L1 corpus comparable with the spoken L2 corpus, a number of strategies were employed to represent speech in written forms in appropriate ways. Chapter 4 has provided a rather detailed description of the transcription procedure of the spoken Chinese corpus. The main work that has been done in terms of transcription was to write down what I exactly heard in the recordings by using the standardised Chinese characters and *Pinyin* wherever necessary. It was noted that L2 speech of Chinese shared a number of features with L1 speech of Chinese, Chapter 4 therefore did not discuss L2 transcription separately.

The only difference between L1 and L2 transcription in this thesis was the features of the non-standard or incorrect tones used by L2 speakers. A number of prosodic features which occurred in L2 speech were omitted in the transcripts, e.g., the non-standard or incorrect tones. I represented these non-standard tones with the correct words without paying special attention to the sounds of the non-standard or incorrect tones of L2 speakers, because in many cases, the L2 speakers knew exactly what they were talking about but only pronounced with the incorrect tones. It may be criticised that using the correct characters to represent some incorrect tones harms the validity and reliability of the transcripts. True, this decision indicated one obvious disadvantage of the orthographic transcription. But it seemed to me that it was difficult to represent the feature of the non-standard tones in a more appropriate way in orthographically transcribed texts. Since I took part in all the conversational interactions, I was familiar with the interactions. Therefore, I had the confidence to claim that I transcribed the non-standard tones properly. On the other hand, it is not uncommon that L2 speakers of any language have foreign accents. It thus would be extremely labour-extensive and time-consuming if transcribers were required to represent all non-standard tones in an appropriate way. In short, bearing my

research purposes in mind, there was no need to carry out the transcription at a higher level of granularity. In addition to the non-standard tones, I followed some transcription conventions observed in the Spoken BNC2014 and the TLC to maximise the reliability and consistency of the transcripts. It has proved that this decision enabled me to signal some minimal response tokens efficiently.

As all the audio recordings were not publicly accessible, unfortunately, the reliability and consistency of the transcripts cannot be attested. However, until reliable ways can be found to protect participants' privacy in the recordings, it seemed reasonable to keep the recordings to the researcher. I hope that my careful attention to the discussion of the detailed processes of planning, designing, and compiling this corpus will present me as worthy of trust in this regard.

In conclusion, Chapters 3 and 4 have transparently documented the compilation of the spoken L2 corpus. Many aspects involved in the design and data collection should be improved (e.g., size) and users need to be cautious when approaching the data to conduct studies on L2 speech. Acknowledging the drawbacks of this corpus, it should be pointed out that this corpus can be of use to researchers who are interested in L2 conversational interactions. Also, the variables that have been considered during the corpus design and data collection can benefit researchers who aim to build their own spoken L2 corpora.

## 7.4 Corpus Similarity Between the L1 Corpus and the L2 Corpus

Before discussing the research implications of the spoken Chinese corpus, this thesis compared the L1 corpus and the L2 corpus in some depth in Chapter 5 to address the second research question: to what extent is the spoken L1 corpus comparable to the spoken L2 corpus? It shows that while ensuring comparability with the spoken L2 corpus was a major consideration when building the spoken L1 corpus, there are noticeable differences between the two corpora other than the native/non-native distinction. Even so, the spoken L1 corpus is more comparable than any other spoken L2 corpora in existence to the spoken L2 corpus built in this study. The similarities and differences between the L1 corpus and the L2 corpus are summarised in Table 36.

**Table 36**

*Similarities and Differences between the L1 Corpus and the L2 Corpus in Composition*

| **Factor** | | **The L1 corpus** | **The L2 corpus** |
|---|---|---|---|
| **Data** | language | *Putonghua* used in mainland China | |
| | genre | informal conversations between two interlocutors | |
| | method of data collection | unstructured interviews conducted by the researcher | |
| **Corpus size** | overall size | 228,306 words | 220,792 words |
| | participants' output | 174,009 words | 146,598 words |
| | the researcher's output | 54,297 words | 74,194 words |
| **The number of participants** | | 22 | 14 |
| **The number of sample texts** | | 26 | 34 |
| **Gender** | female | 14 | 3 |
| | male | 8 | 11 |
| **Age groups** | 20-25 | 3 | 6 |
| | 26-30 | 11 | 2 |
| | 31-35 | 6 | 3 |
| | 36-40 | 2 | 3 |
| **Relationship with the researcher** | | friends and acquaintances | strangers |
| **Interview per person** | | one time | two to four times |

*Note*. Chapter 3 has discussed that two L1 participants attended two interviews per person, and one L1 participant took part three times in this study.

Table 36 shows that the two corpora differ from each other in some respects. First of all, there were more L1 participants than L2 participants who took part in this study, which directly reflected the fact that L1 speakers were much easier to recruit than L2 speakers of Chinese. This being the case, efforts were primarily made to encourage the L2 participants to conduct as many interactions as possible. As a result, the L2 participants contributed more interactions than the L1 participants. Furthermore, the ratio of males with females as 2:3 in the L1 corpus is not identical to that of L2 participants. As a result, this L2 corpus appears to represent male speech rather than female speech, while the L1 corpus is more gender-balanced. Any potential effects of gender on language use were not discussed in this thesis due to the very limited number of participants; however, it is expected that when making use of the L1 corpus and the L2 corpus to conduct contrastive studies, careful consideration can be given to this matter, bearing in mind the difference in gender between the two corpora. In addition, the two corpora shared varied topics. However, this difference is not a drawback, because the two corpora still

share the same genre—informal interaction. In this sense, it is appropriate to claim that the L1 corpus and the L2 corpus are comparable.

As has been mentioned in Chapter 1, the two comparable corpora were designed to carry out contrastive studies. For any corpus studies employing the L1–L2 comparative approach, it is essential to assess the comparability of the L1 and L2 corpora, as this comparison enables researchers to assess the variability of the results that obtain against the differences between the L1 and L2 corpora in design and compilation. It is noticeable that there are differences between the two corpora, however, there is a question of the granularity of the evaluation criteria of corpus similarity. It has been pointed out in Chapter 2 that no clear criteria can be used to evaluate the comparability between an L2 corpus and its L1 counterpart. Therefore, this study simply presented the component of each corpus and pointed out the differences. Given that the two corpora contain informal interactions and are similar in size, it seems reasonable to say that the L1 corpus is comparable to the L2 corpus, although they have some differences caused by design and data collection.

## 7.5 Characteristics of the L2 Use of 就是 *Jiushi* in L1–L2 Interaction

Chapter 6 has investigated the use of 就是 *jiushi* as a discourse marker in informal conversation on the basis of the spoken Chinese corpus. By using the L1–L2 comparative approach, Chapter 6 presented the validity of the research implication of the spoken Chinese corpus built in this thesis and addressed the third research question proposed in Chapter 1: what are the features which characterise the L2 use of the Chinese discourse marker 就是 *jiushi* in the L1–L2 interactions? The characteristics of the L2 use of the reformulation marker 就是 *jiushi* in the L1–L2 interactions are summarised as follows:

(i) the marker 就是 *jiushi* signals that the utterance following it is an alternative means for communicating what is communicated by the preceding segment;

(ii) the L2 speakers of Chinese tend to use the marker 就是 *jiushi* in the above function less often than the L1 speakers;

(iii) the marker 就是 *jiushi* has an uneven distribution both in the L1 corpus and the L2 corpus, and 就是 *jiushi* is dispersed in a similar way in each corpus.

Although the third research question focuses on the L2 use of the marker 就是 *jiushi* in the L1–L2 interactions, there is limited knowledge of the pragmatic roles of 就是 *jiushi* in the Chinese research community. As a result, the qualitative analysis in Chapter 6 was conducted largely on the basis of the L1 data. It seemed to me that this decision was necessary and did not affect the study of the L2 use in a negative way. In what follows I will discuss the above results in order.

Within the framework of relevance theory, I analysed one typical use of the marker 就是 *jiushi* by using the spoken Chinese corpus. It concluded that 就是 *jiushi* as a reformulation marker represents an explicit signal that the segment following 就是 *jiushi* can be entertained as a representation of the preceding segment which it resembles. Specifically speaking, utterances containing the marker 就是 *jiushi* in discourse sequences can be paraphrases, translations, and summaries as reformulations for the proceeding utterances. And, when the marker 就是 *jiushi* is used in parenthetical appositions, it performs an alternative means of reference. The functional analysis of 就是 *jiushi* in this study was conducted within the relevance theoretic framework which reflected a divergent research method in comparison to previous studies; therefore, the interpretation of the function of 就是 *jiushi* which has been discussed by other researchers is to some extent inconsistent with that in previous studies. However, Blakemore's analyses of reformulation markers provided useful insights into the corpus study of the marker 就是 *jiushi* in informal interaction. From the relevance theoretic viewpoint, the discussions in Blakemore (1996, 1997, 2001) raised my awareness of the use of 就是 *jiushi* in parenthetical appositions which has so far been ignored in the Chinese research literature, and more importantly, offered accounts for the use of 就是 *jiushi* on the basis of a general theory of human communication based on cognitive principles. Moreover, another advantage of the relevance theoretic approach was that it had the ability to distinguish the conjunction 就是 *jiushi* from the marker 就是 *jiushi*. As has been pointed out in 6.4.1, researchers who tend to interpret the marker 就是 *jiushi* as an elaboration/explanation marker failed to sufficiently distinguish it from the conjunction 就是 *jiushi* on some occasions. Alternatively, within the relevance theoretic framework, the line between the reformulation marker 就是 *jiushi* and the conjunction 就是 *jiushi* is rather clear. As a result, it is appropriate

to say that relevance theory was sufficient to illuminate pragmatic functions of the marker 就是 *jiushi* in informal speech.

Drawing on the qualitative analysis of 就是 *jiushi*, a quantitative L1–L2 contrastive analysis was then conducted. The results showed that (i) the L2 speakers of Chinese tended to use the reformulation marker 就是 *jiushi* less frequently than the L1 speakers in interaction, and (ii) the reformulation function of 就是 *jiushi* has similar distribution in each corpus. To date, as there is no systematic L1–L2 contrastive research on the use of 就是 *jiushi* that has been carried out in the literature, it remains unclear whether the findings of this study reflect the features of L2 use of the reformulation marker 就是 *jiushi* in speech in general, or represent the performance of the specific L2 group that this study used. This being the case, it is essential to make the statistical information and procedure available to the academic community so that other researchers can examine or replicate the quantitative results (detailed information can be found in Appendix N). Here, I will discuss the factors that may have something to do with the results.

As has been discussed in Chapter 3, the L1 corpus contains L1–L1 interaction between friends and acquaintances while the L2 corpus includes L1–L2 interaction between strangers. This being the case, it is possible that the distinct interlocutors' relationships between the two corpora cause the differences in frequency of the use of 就是 *jiushi* between the two groups. In the literature on English, some researchers claim that the relationship between the speakers in interactions has an influence on discourse marker frequency (e.g., Jucker & Smith, 1998; Redeker, 1990). These studies on English discourse markers focus mainly on L1 use, and are conducted on the basis of limited data. Accordingly, it is impossible that the statements of these studies made can be applied to explain the difference between the L1 and L2 use of 就是 *jiushi* in frequency. In addition, the difference in frequency between the two groups may be largely caused by the status of being native or non-native speakers of Chinese. However, as has been discussed in Chapter 5, due to the design of the two corpora, it is difficult to simply attribute this difference to the variable of being native or non-native speakers. Although the L1 speakers show a tendency to use the reformulation marker 就是 *jiushi* more often than the L2 speakers in speech, as has been claimed in Chapter 5, there is one L1 speaker who strongly prefers using 就是 *jiushi* in the conversations, which indicates that the higher frequency of the use of the marker 就是 *jiushi* in L1 speech to some extent can be attributed to certain L1 speakers'

personal styles. In this regard, it seems that being a native or non-native speaker may not be the key factor that leads to the difference in frequency between the two groups.

The quantitative analysis also shows that there is no difference in the distribution of the marker 就是 *jiushi* between L1 and L2 production. Due to space constraints, I only examined the dispersion of one function of the marker 就是 *jiushi* in the spoken Chinese corpus, so this result cannot be generalised to the dispersion of the other functions of 就是 *jiushi*.

In conclusion, this corpus study on 就是 *jiushi* sufficiently explains one typical usage that the marker 就是 *jiushi* displays in informal interactions both qualitatively and quantitatively. Consequently, the initial findings about the use of the reformulation marker 就是 *jiushi* in informal interaction may encourage further research in the future. On the other hand, the scope of this analysis can be criticised: the importance of individual variation among speakers has been disregarded. Additionally, the challenges that this analysis met (e.g., the effect of the relationship between the interlocutors) reveal the variables which have been ignored in spoken corpus design. For further corpus projects, these issues need to be considered carefully. Finally, the corpus study demonstrates one of the ways in which the spoken Chinese corpus can be employed. In what follows, I will discuss the implications of the corpus in some detail.

## 7.6 Implications

One of the principle contributions of this study is the spoken Chinese corpus at its centre. Given that it contains informal conversational interaction, this corpus can be used to compare with other types of spoken data to investigate language use in different genres. Another possible research implication is that it documents searchable variables which can be a reliable resource for SLA researchers. Given that L1–L2 interactions have been popular in SLA research (see Chapter 2), the L2 corpus can be used in SLA research to study L1–L2 interactions which can lead to firm inferences that can be drawn from a much wider empirical basis. Moreover, both the spoken L1 and L2 data can constitute the raw material for a wide range of Natural Language Processing based tasks, such as training of part of speech taggers and parsers. By using the spoken data, researchers may be able to develop more useful automatic annotation tools to process spoken Chinese and to improve the accuracy of the results of automated annotation.

Research into contrastive interlanguage analysis is another way in which this study can be extended. Although this study focuses primarily on the L1–L2 contrastive approach, the spoken L2 corpus can be compared with other L2 Chinese corpora (both spoken and written) to study L2 production. In addition, as the L2 corpus mainly includes speech produced by New Zealanders who are native English speakers, it can be employed to compare spoken L2 corpora featuring L2 speakers of Chinese who come from other English-speaking countries (e.g., the UK, Canada). Or it can be employed to compare L2 speakers of Chinese with different L1 backgrounds to explore the potential influences of various L1 backgrounds on L2 production.

The spoken Chinese corpus can be a rich source of data for pedagogical applications as well. Here, the spoken Chinese corpus can be seen as a corpus for "delayed pedagogical use (DPU)" (Granger, 2009, p. 20), as its L2 part is not used directly as teaching/learning material by the L2 speakers who have produced the data. Since the L2 corpus represents speech of native English speakers of Chinese recorded in non-academic settings, the data can be used to help teachers design some classroom exercises with similar topics or situations to practice L2 learners' communicative skills. Then, Chinese teachers can teach L2 learners how to converse with L1 speakers in daily life by helping them analyse examples taken from the L2 corpus. From a pedagogical point of view, the L1 corpus can show what is common in conversations between friends and acquaintances, which can be useful in providing more empirical and authentic examples to base pedagogical practices on. It should be noted that features of L1 production uncovered in the L1 corpus need not necessarily lead to targeted actions in the classroom, but can be simply presented as some useful strategies for L2 learners of Chinese who aim to improve their communicative skills.

In addition to the implications of corpus data to Chinese teaching, the L1–L2 contrastive analysis conducted in Chapter 6 can help Chinese teachers identify the discourse features that differentiate L2 production from L1 speakers. However, the fact that the L2 use of discourse markers does not approach L1 frequency should not be interpreted as indicating that L2 speakers are deficient users of Chinese. When offering pedagogical suggestions, it is essential to focus on whether L2 speakers react in an appropriate manner, rather than focusing on the frequency of discourse markers and searching for 'norms' based on L1 performance (Gablasova, Brezina, McEnery, et al., 2017). More importantly, the pragmatic function of 就是 *jiushi* identified in this study can be taught to L2 learners of Chinese in classrooms; however,

it should be borne in mind that there is no need to bring L2 learners entirely into conformity with L1 controls in discourse marker training.

## 7.7 Summary

This discussion chapter has provided answers to the research questions by drawing on the findings and discussing them with reference to the relevant literature and analytical frameworks. It began with some considerations of the L2 corpus compilation and has argued that variables in terms of L2 participant recruitment, L1–L2 interaction, and transcription have been considered in corpus design. Then, it turned to discuss decisions and measures taken with respect to the L1 corpus to ensure comparability with the L2 corpus. It was pointed out that when building the L1 corpus, I put representativeness at a more important place. Following this, I addressed the second research question. Due to practical constraints, there were differences between the two corpora in some respects, e.g., the ratio of males to females in each corpus. Acknowledging the differences between the two corpora, it was argued that the L1 corpus is essentially comparable to the L2 corpus. Drawing on the discussions on the L1–L2 corpus comparison, the third research question was addressed in some detail. Within the framework of relevance theory, I focused exclusively on the interpretive use of the marker 就是 *jiushi* and found that the L1 speakers tend to use 就是 *jiushi* more frequently than the L2 speakers, and the reformulation function of 就是 *jiushi* has similar distribution in each corpus. At the end of this chapter, I also gave a brief discussion of some implications of the findings of this thesis. It is hoped that the spoken Chinese corpus can benefit the Chinese research community and provide available corpus evidence to support linguistic research.

# Chapter 8 Conclusion

## 8.1 Overview of This Thesis

In this thesis, I have presented an account of the design, compilation, and analysis of a spoken Chinese corpus which is made up of an L1 corpus and an L2 corpus. My aim has been to make clear the major decisions taken during the procedures of participant recruitment, data collection, and transcription, as well as to demonstrate the research potential of this corpus in pragmatics. The two subcorpora of the spoken Chinese corpus can be of use to researchers who attempt to study L2 production by employing the L1–L2 comparative approach. Also, each subcorpus can be used alone to investigate L1 or L2 speech from different perspectives.

The spoken Chinese corpus contains informal conversational interactions, specifically, the spoken L1 corpus comprises L1–L1 interaction while the spoken L2 corpus includes L1–L2 interaction. Chapter 1 gave some background on this design, providing the justification for the focus of this thesis on the creation of the spoken Chinese corpus. At the time of writing, this corpus is the first spoken Chinese corpus resource that fulfils the following main strengths:

(i) it is made up of an L2 corpus and its counterpart–an L1 corpus
(ii) the L2 corpus contains L1–L2 interaction gathered in non-academic settings
(iii) the L1 corpus consists of L1–L1 informal conversational interaction
(iv) it contains orthographically transcribed data and a bespoke transcription scheme

To date, most of the spoken L2 Chinese corpora in existence, if not all, represent L2 academic data. Thus, my work on the spoken Chinese corpus was intended to achieve the following aims:

(i) to build a spoken L2 corpus of informal conversational interaction which can broaden the traditional databases for L2 studies;
(ii) to create a spoken L1 corpus which matches the spoken L2 corpus as much as possible; and, in achieving the first two aims,
(iii) to provide a new kind of data source for L1–L2 contrastive studies on a wide range of Chinese linguistic phenomena.

With these aims in mind, Chapter 2 first reviewed some key considerations in spoken L2 corpus design, including L2 participant selection, L2 data types and collection, corpus size and

representativeness, transcription, and annotation. These discussions provided both theoretical and methodological guidelines to the compilation of the spoken Chinese corpus. Subsequently, Chapters 3 and 4 presented the transparent procedures of the spoken Chinese corpus to achieve the first two objectives. In Chapter 2, I also paid attention to the L1–L2 comparative approach which has been widely employed in L2 corpus studies and discussed some methodological issues in L2 corpus studies.

The stages of the corpus creation have been illuminated in the chronological order in which they occurred: the procedures of the design and data collection of both the L1 and L2 corpora have been demonstrated in Chapter 3, and the decisions and compromises made on the transcription and the development of the bespoke transcription scheme have been presented in some detail in Chapter 4. One reason for combining the same stages of the L1 and L2 corpus building into the same chapters was to highlight the fact that the two corpora were designed with the similar criteria. The chronological order also illustrated that both chapters have had as their central themes certain issues involved in spoken corpus compilation that this thesis has attempted to address. In Chapter 3, I demonstrated prominent aspects that should be considered when designing the spoken Chinese corpus consisting of (L1–L1 and L1–L2) conversational interaction, including the target population, the target corpus size, participant recruitment, the data collection method, as well as the procedure of data gathering. The practice of this very basic description of the procedures in terms of spoken corpus design and data collection has so far been overlooked in the majority of studies on spoken Chinese corpus compilation at the time of writing. This being the case, the detailed descriptions of the spoken corpus design and data collection offered in Chapter 3 made the constructions of both the L1 and L2 corpora more tractable, and encouraged other users to access the data with sensitivity to the methodological issues identified while constructing the two corpora.

With the gathered spoken data, Chapter 4 discussed the decisions and compromises made with respect to the transcription of the audio recordings, and produced a bespoke transcription scheme on the basis of the purposes of this thesis. Since the gathered L1 and L2 speech shared a number of features, it was sufficient to create one transcription scheme covering the common features in both the L1–L1 and L1–L2 conversations. The transcription scheme also documents certain specific features belonging to L2 production. Additionally, to make the transcription scheme more reliable, I combined some conventions in the transcription systems of the Spoken BNC2014 and the TLC with the specific features of spoken Chinese. Even though a number of features were omitted (e.g., non-verbal sounds) in this thesis due to practical reasons, Chapter

4 explored in some detail the rationale behind the selection of some important features for this study, which to some degree deepened the understanding of the characteristics of informal spoken Chinese. Given that transcription has long been disregarded in the Chinese research literature, the decision to set transcription as the focus of Chapter 4 expressly placed emphasis on the significant role that transcription has in spoken corpus building; accordingly, the theoretical and methodological discussions noticeably filled the gap in transcription in the Chinese research literature.

Since the L1 corpus and the L2 corpus were designed to be comparable to conduct L1–L2 contrastive studies, I evaluated the comparability of the two corpora in Chapter 5. The imperative of corpus comparability between L1 and L2 corpora in contrastive research has received some attention; however, practical actions are rarely taken to compare and evaluate L1 and L2 corpora which were designed with the same principle in contrastive studies. To obtain a rather comprehensive overview of the similarities and differences between the two corpora, and more importantly, to inform whether the observed differences which resulted from some compromises made during the data collection may affect the results of the L1–L2 contrastive analysis conducted in Chapter 6, Chapter 5 compared the L1 corpus and the L2 corpus from the point of view of corpus components. It first demonstrated that the two corpora were similar in size. On the other hand, Chapter 5 discussed that the two corpora differed from each other in some respects, although they were designed with the similar criteria. For instance, in contrast to the L1 corpus which contains more female speakers of Chinese, the L2 corpus unexpectedly underrepresented female speakers due to the relatively small number of L2 participants who participated in this study. In this regard, discrepancies in terms of gender between the L1 corpus and the L2 corpus made the two corpora less comparable. Additionally, the sample texts differed considerably in size in each corpus and it was realised that the varied proportions of the sample texts might have some influences on the representativeness and balance of the two corpora. The differences between and within the L1 corpus and the L2 corpus were the ramifications of decisions and compromises made in order to recruit enough participants and to collect sufficient spoken data. In doing this, it attempted to argue that (i) the differences between the L1 and L2 use caused by corpus design might have influences on the differences observed between L1 and L2 use in studies, and that (ii) it was of paramount importance to take into consideration the variation within each corpus when interpreting L2 production in L1–L2 contrastive studies.

Although there are some issues involved in the design and compilation of the spoken Chinese corpus, Chapter 6 showed that it can be used to analyse discourse markers. Drawing on the data gathered for the spoken Chinese corpus, Chapter 6 analysed the L2 use of a Chinese discourse marker 就是 *jiushi* in comparison with the L1 use within the framework of relevance theory. Due to space constraints, Chapter 6 does not attempt to produce a full analysis of the pragmatic functions of 就是 *jiushi* in informal conversations. Rather, it focused exclusively on the reformulation function of 就是 *jiushi* in speech. The result shows that the reformulation marker 就是 *jiushi* has the capacity to represent that the utterance following it is an alternative means for communicating what is communicated by the preceding utterance. On the basis of the functional analysis, in Chapter 6, I also conducted a quantitative analysis of the reformulation marker 就是 *jiushi* to examine the differences in frequency and dispersion between the L1 and L2 use. The findings include that (i) the L1 speakers tend to use 就是 *jiushi* more frequently than the L2 speakers in informal interaction, and (ii) 就是 *jiushi* has similar distribution in each corpus.

## 8.2 Contributions and Limitations of This Thesis

The main contribution of the thesis to the research community is the spoken Chinese corpus. For researchers who are interested in conversational interactions of Chinese and who have attempted to use existing spoken Chinese corpora to carry out their studies, the spoken Chinese corpus built in this thesis may meet their research needs once it is released. Given that the spoken Chinese corpus contains an L1 part and an L2 part, it enables L2 researchers to investigate L2 language making use of two comparable spoken corpora, which may lead to more valid and reliable findings in comparison to using two corpora designed with different criteria. Also, the spoken L2 corpus as a new addition to the research community can complement evidence from existing spoken L2 Chinese corpora that represent academic data.

In addition to the obtainable data, this thesis gives transparent and documented accounts of each step of the construction of both the spoken L1 and L2 corpora. Therefore, the important decisions and compromises made when constructing the spoken Chinese corpus can provide some valuable suggestions to researchers who want to build their own spoken Chinese corpora. The detailed discussions of transcription presented in this thesis in particular can, alongside the

creation of the bespoke transcription scheme for the corpus, not only inform users and future corpus compilers of spoken Chinese corpora of the importance of well-organised transcription for corpus analyses, but also provide some useful guidelines underlying transcription to the Chinese research community. Moreover, the previous chapters have shown that this spoken Chinese corpus was built by following state-of-the-art methodological approaches as well as corpus practices in English corpus linguistics. This thesis therefore represents an example of good practice by employing theoretical and methodological approaches that have been used in the compilation of English corpora to inform spoken Chinese corpus building. In this respect, this thesis to some extent can encourage Chinese researchers and future corpus compilers of Chinese corpora (both written and spoken) to pay more attention to and learn more from studies carried out in the field of English corpus linguistics to facilitate the development of Chinese corpus linguistics and L2 corpus research on Chinese.

The third major contribution relates to the investigation of the use of the marker 就是 *jiushi*. It is the first corpus-based contrastive analysis of the L2 use of 就是 *jiushi* in informal conversation at the time of writing, which can contribute noticeably to the research community in some respects. First of all, this corpus analysis has shown that relevance theory has the ability to provide more sufficient insights into the ways in which the marker 就是 *jiushi* affects utterance interpretation in communication than other theoretical frameworks used in previous studies. Accordingly, this thesis offers an alternative approach to study the pragmatic functions of Chinese discourse markers, and deepens the understanding of the role that the marker 就是 *jiushi* has in informal interactions. Secondly, the corpus analysis applied relative frequency and the dispersion measure *DP* to investigate the features of the L2 use of 就是 *jiushi* in comparison with L1 use, which contributes to the understanding of the discrepancies between L1 and L2 use.

Moreover, this thesis places some emphasis on the potential influences of the differences between the L1 corpus and the L2 corpus in design and creation on the subsequent corpus analysis by examining the discrepancies of the components of the two corpora. In doing this, it encourages users to access the data with sensitivity to the methodological issues identified while constructing the two corpora. All of these contributions in turn prove that small corpora are valuable resources and can contribute to linguistic research.

This thesis also has some limitations. From the point of view of corpus design, the L1 corpus does not match the L2 corpus perfectly; as a result, the differences in corpus design made it impossible to evaluate to what extent the observed differences in the use of 就是 *jiushi* between the L1 and L2 participants were caused by the variable of being native or non-native speakers of Chinese or relevant to the relationship between the interlocutors. Another drawback in terms of corpus design is about the gathering of metadata. When gathering speaker metadata, rather than asking participants to provide their own information prior to the conversations, I mainly collected speaker metadata information during the conversations. Moreover, gender information was gathered based on my personal judgement. Collecting metadata in this way is rather subjective and may affect the subsequent corpus-based analyses.

One limitation in terms of the spoken Chinese corpus is that, compared to some existing spoken corpora, by today's standard, it is small in size and consists of a limited number of (L1 and L2) participants. Consequently, there is a fairly sizable gender imbalance favouring male speakers of Chinese in the L2 corpus which hinders the achievement of corpus balance. In this regard, the L2 corpus tends to represent mixed-sex informal interactions while the L1 corpus is more gender-balanced. Furthermore, the spoken Chinese corpus overrepresents certain participants' speech, which has the risk to skew the data when conducting corpus studies. Therefore, due to the limited size of the spoken Chinese corpus, caution should be exercised when generalising the findings of the marker 就是 *jiushi* to a wider population group.

In terms of transcription, this thesis has not prepared the audio recordings for public release, so corpus users have to depend fully on the transcripts given by the researcher. This decision makes it impossible for other users to evaluate the validity and reliability of the transcripts. Additionally, since the spoken Chinese corpus represents my domain of research interest, many features (e.g., non-verbal sounds, overlaps) were omitted in the transcripts according to the research questions that this thesis attempted to address. In this regard, the spoken Chinese corpus may be criticised because the transcripts may not meet some researchers' needs. Moreover, the orthographical transcripts made limited contributions to users who are interested in prosodic features in spoken Chinese. However, no single corpus can answer all the research questions that researchers attempt to address (Egbert et al., 2020; Granger, 2021). This thesis documents each step of corpus design and compilation and identifies where mismatches may arise, which allows researchers to make their judgments about whether the spoken Chinese corpus is an appropriate source for their research.

For researchers and users who support corpus annotation, another limitation of the spoken Chinese corpus is that it only contains raw data. It should be noted that this study does not object to corpus annotation, but it seems that there are good reasons to provide raw data rather than annotated data to the research community. Aside from the reason that has been discussed in Chapter 2, that is, the concern of the accuracy of the results of annotation (see 2.2.4), corpus compilers cannot annotate texts in a way that would be exactly suitable for all research purposes. Consequently, "it is really valuable to pass on or give access to plain text" (Sinclair, 2014, p. 102). There are many available tools that can be employed to annotate corpus data on the basis of specific research purposes. By using these publicly available free tools, researchers are able to process the raw data of the spoken Chinese corpus in ways that meet their research needs.

## 8.3 Looking Ahead

This thesis is a pilot study for the compilation of a spoken Chinese corpus containing informal interactions. Since there are several L1 and L2 recordings of informal conversations that have not been transcribed due to practical reasons, in the future I will transcribe these recordings to extend the size of this corpus. More importantly, since the L2 corpus is imbalanced in favour of male speakers of Chinese, efforts will be made to gather more speech produced by female speakers of Chinese. Furthermore, I aim to convert all the text files into XML files, and then release the XML files to the public in the future. Admittedly, there are many things I did not do which would add more value to the spoken corpus data. So, I welcome and encourage other researchers to refine the data (e.g., by adding layers of annotation) and supplement both the L1 and L2 subcorpora so that the spoken Chinese corpus may become more useful in the future.

Also, this thesis has proved that the spoken Chinese corpus has the ability to conduct quantitative and qualitative analyses of Chinese discourse markers. It is expected that the study of the L2 use of 就是 *jiushi* from the perspective of relevance theory can be a good starting point for Chinese discourse marker studies on L2 production. I hope this study can promote an interest in the study of Chinese discourse markers in L2 informal interaction. Once the corpus is released, I also hope the findings of the corpus analysis of the L2 use of 就是 *jiushi* in this thesis can be replicated by researchers with an interest in Chinese discourse markers.

## References


Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research* (pp. 35-53). Rodopi.

Ädel, A. (2015). Variability in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 379-400). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.018

Adolphs, S., & Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. Routledge.

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38-52). Routledge.

Adolphs, S., Knight, D., & Carter, R. (2015). Beyond modal spoken corpora: A dynamic approach to tracking language in context. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies: Intergrating discourse and corpora* (pp. 41-62). Palgrave Macmillan.

Aijmer, K. (2002). *English discourse particles: Evidence from a corpus*. John Benjamins Publishing. https://doi.org/10.1075/scl.10

Aijmer, K. (2011). Well I'm not sure I think… The use of well by non-native speakers. *International Journal of Corpus Linguistics, 16*(2), 231-254. https://doi.org/10.1075/ijcl.16.2.04aij

Aijmer, K. (2013). *Understanding pragmatic markers: A variational pragmatic approach*. Edinburgh University Press.

Allwood, J. (2010). Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 207-225). Walter de Gruyter.

Andersen, G. (2001). *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents*. John Benjamins Publishing. https://doi.org/10.1075/pbns.84

Andersen, G. (2016). Semi-lexical features in corpus transcription: Consitency, comparability, stadardisation. *International Journal of Corpus Linguistics, 21*(3), 323-347. https://doi.org/10.1075/ijcl.21.3.02and

Aston, G. (2008). "It's only Human...". In A. Martelli & V. Pulcini (Eds.), *Investigating English with corpora: Studies in honour of Maria Teresa Prat* (pp. 343-354). Polimetrica International Scientific Publisher.

Atkins, S., Clear, J., & Ostler, N. (1991). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1-16. https://doi.org/10.1093/llc/7.1.1

Atkinson, J. M., & Heritage, J. (1999). Transcript notation-structures of social action: Studies in conversation analysis. *Aphasiology, 13*(4-5), 243-249. https://doi.org/10.1080/026870399402073

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum.

Baker, P. (2014). *Using Corpora to Analyze Gender*. Bloomsbury.

Bardovi - Harlig, K. (2013). Developing L2 pragmatics. *Language Learning, 63*, 68-86. https://doi.org/10.1111/j.1467-9922.2012.00738.x

Beeching, K. (2016). *Pragmatic markers in British English: Meaning in social interaction*. Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press.

Biber, D., & Jones, J. K. (2008). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (Vol. 1, pp. 1287-1304). De Gruyter.

Biq, Y.-O. (1990). Conversation, continuation, and connectives. *Text-Interdisciplinary Journal for the Study of Discourse, 10*(3), 187-208. https://doi.org/10.1515/text.1.1990.10.3.187

Biq, Y.-O. (2001). The grammaticalization of "jiushi" and "jiushishuo" in Mandarin Chinese. *Concentric: Studies in Linguistics, 27*(2), 53-74.

Birdsong, D. (2005). Nativelikeness and non-nativelikeness in L2A research. *IRAL-International Review of Applied Linguistics in Language Teaching, 43*, 319-328. https://doi.org/10.1515/iral.2005.43.4.319

Birdsong, D. (2006). Why not fossilization. In Z. Han & T. Odlin (Eds.), *Studies of fossilization in second language acquisition* (pp. 173-188). Multilingual Matters. https://doi.org/10.21832/9781853598371-011

Birdsong, D., & Gertken, L. M. (2013). In faint praise of folly: A critical review of native/non-native speaker comparisons, with examples from native and bilingual processing of French complex syntax. *Language, Interaction and Acquisition, 4*(2), 107-133. https://doi.org/https://doi.org/10.1075/lia.4.2.01bir

Blakemore, D. (1993). The relevance of reformulations. *Language and Literature, 2*(2), 101-120. https://doi.org/10.1177/096394709300200202

Blakemore, D. (1996). Are apposition markers discourse markers? *Linguistics, 32*, 325-347. https://doi.org/10.1017/S0022226700015917

Blakemore, D. (1997). Restatement and exemplification: A relevance theoretic re-assessment of elaboration. *Pragmatics and Cognition, 5*(1), 1-19. https://doi.org/10.1075/pc.5.1.04bla

Blakemore, D. (2001). Discourse and relevance theory. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 100-118). Blackwell.

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. Cambridge University Press.

Blakemore, D. (2005). And-parentheticals. *Journal of Pragmatics, 37*(8), 1165-1181. https://doi.org/10.1016/j.pragma.2005.04.003

Bley - Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning, 33*(1), 1-17. https://doi.org/10.1111/j.1467-1770.1983.tb00983.x

Boersma, P. (2014). The use of Praat in corpus research. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 342-360). Oxford University Press.

Bolton, K., & Graddol, D. (2012). English in China today. *English Today, 28*(3), 3-9. https://doi.org/10.1017/S0266078410000118

Breiteneder, A., Pitzl, M.-L., Majewski, S., & Klimpfinger, T. (2006). VOICE recording—Methodological challenges in the compilation of a corpus of spoken ELF. *Nordic Journal of English Studies, 5*(2), 161-187. https://doi.org/10.35360/njes.16

Brezina, V. (2018). *Statistics in corpus linguistics*. Cambridge University Press. https://doi.org/10.1017/9781316410899

Brezina, V., Gablasova, D., & McEnery, T. (2019). Corpus-based approaches to spoken L2 production: Evidence from the Trinity Lancaster Corpus. *International Journal of Learner Corpus Research, 5*(2), 119-125. https://doi.org/10.1075/ijlcr.00008.int

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139-173. https://doi.org/10.1075/ijcl.20.2.01bre

Brezina, V., & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics, 19*(1), 1-28. https://doi.org/10.1075/ijcl.19.1.01bre

Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). *#LancsBox v. 5.x. [Computer software]*. Lancaster University. http://corpora.lancs.ac.uk/lancsbox

Burnard, L. (2002). Where did we go wrong? A retrospective look at the British National Corpus. In K. Bernard & M. Georg (Eds.), *Teaching and learning by doing corpus analysis* (pp. 51-70). Rodopi.

Burnard, L. (2005). Metadata for corpus work. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 30-46). Oxbow Books.

Burton-Roberts, N. (1999). Apposition. In K. Brown & J. Miller (Eds.), *Concise encyclopedia of grammatical categories* (pp. 25-29). Elsevier.

Busse, B., & Kleiber, I. (2020). Realizing an online conference: Organization, mangagement, tools, communication, and co-creation. *International Journal of Corpus Linguistics, 25*(3), 322-346. https://doi.org/10.1075/ijcl.00028.bus

Buysse, L. (2012). So as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics, 44*(13), 1764-1782. https://doi.org/10.1016/j.pragma.2012.08.012

Buysse, L. (2015). 'Well it's not very ideal…' The pragmatic marker well in learner English. *Intercultural Pragmatics, 12*(1), 59-89. https://doi.org/10.1515/ip-2015-0003

Buysse, L. (2017). The pragmatic marker you know in learner Englishes. *Journal of Pragmatics, 121*, 40-57. https://doi.org/10.1016/j.pragma.2017.09.010

Callies, M. (2013). Advancing the research agenda of interlanguage pragmatics: The role of learner corpora. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013* (pp. 9-36). Springer. https://doi.org/10.1007/978-94-007-6250-3_2

Callies, M. (2015). Using learner corpora in language testing and assessment: Current practice and future challenges. In E. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment* (pp. 21-35). Peter Lang.

Cameron, D. (2001). *Working with Spoken Discourse*. SAGE Publications.

Cao, X. (2013). On the construction of longitudinal Chinese inter-language corpus (留学生汉语中介语纵向语料库建设的若干问题). *Applied Linguistics*(2), 127-134.

Cao, X., & Wu, C. (2013). Measuring L2 Chinese and building a longitudinal Chinese interlanguage corpus (汉语中介语综合测度指标及中介语发展语料库的建设). In X. Cui & B. Zhang (Eds.), *Selected papers from the 2th international symposium on the construction and application of Chinese interlanguage corpus (第二届汉语中介语语料库建设与应用国际学术讨论会论文选集)* (pp. 88-98). Beijing Language and Culture University Press.

Carlsen, C. (2012). Proficiency level—A fuzzy variable in computer learner corpora. *Applied Linguistics, 33*, 161-183. https://doi.org/10.1093/applin/amr047

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics, 16*(2), 141-158. https://doi.org/10.1093/applin/16.2.141

Chafe, W. (1988). Linking intonation units in spoken English. In J. Haiman & S. A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 1-27). John Benjamins Publishing.

Chafe, W. L. (2014). Adequacy, user-friendliness, and practicality in transcribing. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application* (pp. 54-61). Routledge.

Chafe, W. L., Du Bois, J. W., & Thompson, S. A. (2013). Towards a new corpus of spoken American English. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik* (pp. 64-82). Routledge.

Chen, H., & Xu, H. (2019). Quantitative linguistics approach to interlanguage development: A study based on the Guangwai-Lancaster Chinese Learner Corpus. *Lingua*, 1-15. https://doi.org/10.1016/j.lingua.2019.102736

Chen, J. (2018). (Inter)subjectification at the left and right periphery: Deriving Chinese pragmatic marker bushi from the negative copula. *Language Sciences, 66*, 83-102. https://doi.org/10.1016/j.langsci.2018.01.002

Chen, J. (2019). *Pragmatic function and diachronic evolution of Chinese discourse markers (汉语话语标记的语用功能与历时演变)*. Fudan University Press.

Chen, L. (2017). Pragmatic and discursive function of "na""nage""name" (话语标记"那""那个""那么"语用语篇功能辨析). *Journal of Shenyang University (Social Science), 19*(6), 729-738.

Chen, X. (1996). An introduction to the Chinese Interlanguage Corpus ("汉语中介语语料库系统"介绍). Proceedings of the 5th International Conference on Chinese Language Teaching, Beijing, China.

Chu, C., & Chen, X. (1993). Basic assumption for establishing the Chinese Interlanguage Corpus system (建立"汉语中介语语料库系统"的基本设想). *Chinese Teaching in the World*(3), 199-205.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics, 22*, 75-95. https://doi.org/10.1017/S0267190502000041

Cook, G. (2014). Theoretical issues: Transcribing the untranscribable. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer: Transcription, markup and applications* (pp. 35-53). Routledge.

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching, 5*(4), 161-170.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing, 8*(4), 259-265. https://doi.org/https://doi.org/10.1093/llc/8.4.259

Crowdy, S. (1994). Spoken corpus transcription. *Literary and Linguistic Computing, 9*(1), 25-28. https://doi.org/10.1093/llc/9.1.25

Crowdy, S. (2014). The BNC spoken corpus. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application* (pp. 224-234). Routledge.

Crystal, D. (1988). Another look at, well, you know…. *English Today, 4*(1), 47-49. https://doi.org/10.1017/S0266078400003321

Crystal, D., & Davy, D. (1975). *Advanced Conversational English*. Longman. https://www.davidcrystal.com/Files/BooksAndArticles/-5312.pdf

Crystal, D., & Quirk, R. (1964). *Systems of prosodic and paralinguistic features in English*. Mouton.

Cucchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. *Clinical Linguistics & Phonetics, 10*(2), 131-155. https://doi.org/10.3109/02699209608985167

Cui, X., & Zhang, B. (2011a). Building the International Corpus of Learner Chinese ("全球汉语学习者语料库"建设方案). In X. Xiao & W. Zhang (Eds.), *Selected papers from the 1st international symposium on the construction and application of Chinese*

*interlanguage corpora (首届汉语中介语语料库建设与应用国际学术讨论会论文选集)* (pp. 23-32). World Publishing Corporation.

Cui, X., & Zhang, B. (2011b). Principles for building the International Corpus of Learner Chinese (全球汉语学习者语料库建设方案). *Applied Linguistics, 2*, 100-108. https://doi.org/10.16499/j.cnki.1003-5397.2011.02.017

Dang, J., Wang, L., & Su, J. (2017). Vocabulary borrowing from Putonghua in Taiwan Mandarin (台湾国语吸收大陆普通话词语趋势研究). *Applied Linguistics, 4*, 113-121.

Davies, A. (2011). Does language testing need the native speakers? *Language Assessment Quarterly, 8*(3), 291-308. https://doi.org/10.1080/15434303.2011.570827

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL), 2*(1), 225-246.

Deng, D., Shi, F., & Lv, S. (2006). A contrastive study on tones of Putonghua and Taiwan Mandarin (普通话与台湾国语声调的对比分析). *Acta Acustica, 31*(6), 536-541.

Diao, Y. (2015). An investigation of the vocabulary convergence of Taiwan Mandarin and Putonghua (台湾"国语"词汇与大陆普通话趋同现象调查). *Studies of the Chinese Language, 3*, 278-288.

Ding, A., Guo, Y., & Zhao, Y. (2000). Chinese, Putonghua, Guoyu and Huayu—Which term should be used to represent the offical language in China? (应该怎么样称呼现代中国的官方语言？—从英汉对比看"汉语"，"普通话"，"国语"与"华语"等概念的使用). *Journal of Henan Normal University (Philosophy and Social Sciences), 27*(3), 96-102.

Dong, S., & Huang, C.-R. (2020). A comparative study on haplology of Putonghua and Taiwan Mandarin (两岸同音删略现象对比研究). *Chinese Linguistics, 1*, 14-24.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. *International Pragmatics Association, 1*(1), 71-106. https://doi.org/10.1075/prag.1.1.04boi

Du Bois, J. W., Cumming, S., Schuetze-Coburn, S., & Paolino, D. (1992). *Discourse transcription: Santa Barbara papers in linguistics* (Vol. 4). Department of Linguistics, University of California.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & Paolino, D. (1993). Outline of discourse transcription. In J. A. Edwards & M. D. Lampert (Eds.), *Talking data: Transcription and coding in discourse research* (pp. 45-89). Lawrence Erlbaum Associates.

Duan, L. (2010). Transcription as a approach to represent verbal communication (口语交际书面化的转写桥梁). *Social Scientist, 10*, 159-161.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching, 47*(2), 157-177. https://doi.org/10.1515/iral.2009.007

Edwards, J. A. (2014). Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application* (pp. 19-34). Routledge.

Egbert, J., Larsson, T., & Biber, D. (2020). *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge University Press.

Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.

Espinal, M. T. (1991). The representation of disjunct constituents. *Language, 67*(4), 726-762. https://doi.org/10.2307/415075

Fang, M. (2000). Reduced conjunctions as discourse markers (自然口语中弱化连词的话语标记功能). *Studies of the Chinese Language*(5), 459-480.

Fang, Q. (2013). A corpus-based contrastrative study on mood markers of Taiwan Mandarin and Putonghua (基于口语库统计的两岸话语语气标记比较研究). *TCSOL Studies*(3), 58-65.

Farr, F., Murphy, B., & O'Keeffe, A. (2004). The Limerick Corpus of Irish English: Design, description and application. *Teanga: The Irish Yearbook of Applied Linguistics*, 5-29. http://hdl.handle.net/10344/4712

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal*, *81*(3), 285-300. https://doi.org/10.2307/329302

Fischer, K. (Ed.). (2006a). *Approaches to discourse particles*. Elsevier.

Fischer, K. (2006b). Towards an understanding of the spectrum of approaches to discourse particles: Introduction to the volume. In K. Fischer (Ed.), *Approaches to discourse particles* (pp. 1-20). Elsevier.

Francis, W. N. (1980). A tagged corpus–problems and prospects. In S. Greenbaum, G. Leech, & J. Svartvik (Eds.), *Studies in English Linguistics for Randolph Quirk* (pp. 192-209). Longman.

Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics*, *14*(3), 383-398. https://doi.org/10.1016/0378-2166(90)90096-V

Fraser, B. (1996). Pragmatic markers. *Pragmatics*, *6*(2), 167-190. https://doi.org/10.1075/prag.6.2.03fra

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, *31*(7), 931-952. https://doi.org/10.1016/S0378-2166(98)00101-5

Fuller, J. M. (2003). The influence of speaker roles on discourse marker use. *Journal of Pragmatics*, *35*(1), 23-45. https://doi.org/10.1016/S0378-2166(02)00065-6

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, *28*(3), 410-439. https://doi.org/10.1093/applin/amm030

Gablasova, D., & Brezina, V. (2015). Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics* (pp. 117-136). Springer.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 130-154. https://doi.org/10.1111/lang.12226

Gablasova, D., Brezina, V., & McEnery, T. (2019a). The Trinity Lancaster Corpus: Applications in language teaching and materials development. In S. Götz & J. Mukherjee (Eds.), *Learner corpora and language teaching* (pp. 7-28). John Benjamins Publishing.

Gablasova, D., Brezina, V., & McEnery, T. (2019b). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126-160. https://doi.org/10.1075/ijlcr.19001.gab

Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2017). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, *38*(5), 613-637. https://doi.org/10.1093/applin/amv055

Gao, H., & Tao, H. (2021). Fanzheng 'anyway' as a discourse pragmatic particle in Mandarin conversation: Prosody, locus, and interactional function. *Journal of Pragmatics*, *173*, 148-166. https://doi.org/10.1016/j.pragma.2020.12.003

Garrard, P., Haigh, A.-M., & de Jager, C. (2011). Techniques for transcribers: Assessing and improving consistency in transcripts of spoken language. *Literary and Linguistic Computing*, *26*(4), 389-405. https://doi.org/10.1093/llc/fqr018

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3d ed.). Routledge.
Gass, S. M., & Varonis, E. M. (1985). Variation in native speaker speech modification to non-native speakers. *Studies in Second Language Acquisition*, *7*(1), 37-58. https://doi.org/10.1017/S0272263100005143
Gass, S. M., & Varonis, E. M. (1986). Sex differences in NNS/NNS interaction. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 327-351). Newbury House Publishers.
Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, *16*(3), 283-302. https://doi.org/10.1017/S0272263100013097
Ghadessy, M., Henry, A., & L. Roseberry, R. (2001). *Small corpus studies and ELT: Theory and practice*. John Benjamins Publishing.
Giles, H., & Smith, P. (1979). Accommodation theory: Optimal levels of convergence. In H. Giles & R. N. St. Clair (Eds.), *Language and social psychology* (pp. 45-65). Blackwell.
Gilquin, G., & De Cock, S. (2013). Errors and disfluencies in spoken corpora: Setting the scene. In G. Gilquin & S. De Cock (Eds.), *Errors and disfluencies in spoken corpora* (pp. 1-32). John Benjamins Publishing.
Gilquin, G., De Cock, S., & Granger, S. (Eds.). (2010). *The Louvain International Database of Spoken English Interlanguage*. Presses Universitaires de Louvai.
Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, *1*(1), 41-61. https://doi.org/10.1075/etc.1.1.05gil
Gilquin, G. t. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: Detection, explanation, evaluation. In G. Gilquin, S. Papp, & M. B. Diez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research* (pp. 3-33). Rodopi. https://doi.org/10.1163/9789401206204_002
Glenn, P. (2010). Interviewer laughs: Shared laughter and asymmetries in employment interviews. *Journal of Pragmatics*, *42*(6), 1485-1498. https://doi.org/10.1016/j.pragma.2010.01.009
Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological inquiry*, *50*(3 - 4), 272-302. https://doi.org/10.1111/j.1475-682X.1980.tb00023.x
Gráf, T. (2017). The story of the learner corpus LINDSEI_CZ. *Studie Z Aplikované Lingvistiky*, *2*, 22-35.
Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Language in contrast: Papers from a symposium on text-based cross-linguistic studies, Lund, March 1994* (pp. 37-51). Lund University Press.
Granger, S. (1998a). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). Longman.
Granger, S. (Ed.). (1998b). *Learner English on computer*. Longman.
Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora: Second language acquisition and foreign language teaching* (pp. 3-33). John Benjamins Publishing. https://doi.org/10.1075/lllt.6.04gra
Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, *37*(3), 538-546. https://doi.org/10.2307/3588404
Granger, S. (2004). Computer learner corpus research: Current status and future prospects. *Language and Computers*, *52*, 123-146. https://doi.org/10.1163/9789004333772_008

Granger, S. (2008). Learner corpora. In L. Anke & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 259-275). Walter de Gruyter.
Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). John Benjamins Publishing.
Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 7-29). Blackwell. https://doi.org/10.1002/9781444347340.ch2
Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7-24. https://doi.org/10.1075/ijlcr.1.1.01gra
Granger, S. (2021). Have learner corpus research and SLA finally met? In B. L. Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 243-257). Cambridge University Press.
Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437. https://doi.org/10.1075/ijcl.13.4.02gri
Gries, S. T., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/taget: Two suggestions. *Corpora, 9*(1), 109-136. https://doi.org/10.3366/cor.2014.0053
Gries, S. T., & Wulff, S. (2013). The genitive alternation in Chinese and German ESL learners. *International Journal of Corpus Linguistics, 18*(3), 327-356. https://doi.org/10.1075/ijcl.18.3.04gri
Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain And Language, 36*, 3-15. https://doi.org/10.1016/0093-934x(89)90048-5
Hardie, A. (2012). CQPweb - combing power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380-409. https://doi.org/https://doi.org/10.1075/ijcl.17.3.04har
Haslerud, V., & Stenström, A.-B. (2014). The Bergen Corpus of London Teenager Language (COLT). In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application* (pp. 235-242). Routledge.
Hasselgren, A. (2002). Learner corpora and language testing: Smallwords as markers of learner fluency. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 143-174). John Benjamins Publishing.
Hasund, K. (1998). Protecting the innocent: The issue of informants' anonymity in the COLT corpus. In A. Renouf (Ed.), *Explorations in corpus linguistics* (pp. 13-28). Rodopi.
He, F. (2020). A discussion of the applicability of the English CA transcription notation to spoken Chinese (英语会话转写体系对汉语的适用性研究). *Journal of Henan College of Finance & Taxation, 34*(5), 90-96.
Holt, E. (2010). The last laugh: Shared laughter and topic termination. *Journal of Pragmatics, 42*(6), 1513-1525. https://doi.org/10.1016/j.pragma.2010.01.011
Hu, F., & Wang, X. (2011). *Transcription of the HSK Dynamic Spoken Corpus (HSK 动态口语语料库的语料转写研究)*. The 7th National Applied Linguistics Symposium, Hunan, China.
Hu, X., & Xu, X. (2020). The construction and significance of the Spoken Interlanguage Corpus of Korean Chinese Learners (韩国汉语学习者中介语口语语料库的建设及意义). *TCSOL Studies, 1*, 53-59.
Huang, B., & Liao, X. (2017). *Contemporary Chinese (现代汉语)* (6d ed., Vol. 1). Higher Education Press.

Huang, L.-f. (2014). Constructing the Taiwanese component of the Louvain International Database of Spoken English Interlanguage (LINDSEI). *Taiwan Journal of TESOL, 11*(1), 31-74.

Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.). (2007). *Corpus linguistics and the web*. Rodopi.

Hunston, S. (2008). Collection strategies and design decisions. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 154-168). Walter de Gruyter.

Hyltenstama, K., & Abrahamsson, N. (2012). Introduction: High-level L2 acquisition, learning, and use. *Studies in Second Language Acquisition, 34*, 177-186. https://doi.org/10.1017/S0272263112000010

Iacono, V. L., Symonds, P., & Brown, D. H. (2016). Skype as a tool for qualitative research interviews. *Sociological Research Online, 21*(2), 1-15. https://doi.org/10.5153/sro.3952

Irvine, A. (2011). Duration, dominance and depth in telephone and face-to-face interviews: A comparative exploration. *International Journal of Qualitative Methods, 10*(3), 202-220. https://doi.org/10.1177/160940691101000302

Irvine, A., Drew, P., & Sainsbury, R. (2013). 'Am I not answering your questions properly?' Clarification, adequacy and responsiveness in semi-structured telephone and face-to-face interviews. *Qualitative Research, 13*(1), 87-106. https://doi.org/10.1177/1468794112439086

Janghorban, R., Roudsari, R. L., & Taghipour, A. (2014). Skype interviewing: The new generation of online synchronous interview in qualitative research. *International Journal of Qualitative Studies on Health and Well-being, 9*(1), 24152. https://doi.org/10.3402/qhw.v9.24152

Jefferson, G. (1979). A technique for inviting laughter and its subsequent acceptance eclination. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 79-96). Irvington.

Jefferson, G. (1983). Issues in the transcription of naturally-occurring talk: Caricature versus capturing pronunciational particulars. *Tilburg Papers in Languge and Literature, 34*, 1-12. http://liso-archives.liso.ucsb.edu/Jefferson/Caricature.pdf

Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens "yeah" and "mm hm". *Taylor & Francis Group, 17*(2), 197-216.

Jenks, C. J. (2009). When is it appropriate to talk? Managing overlapping talk in multi-participant voice-based chat rooms. *Computer Assisted Language Learning, 22*(1), 19-30. https://doi.org/10.1080/09588220802613781

Jenks, C. J. (2011). *Transcribing talk and interaction: Issues in the representation of communication data*. John Benjamins Publishing. https://doi.org/10.1075/z.165

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 33-44). John Benjamins Publishing.

Jones, C., Byrne, S., & Halenko, N. (2018). *Successful spoken English: Findings from learner corpora*. Routledge.

Jucker, A. H. (1993). The discourse marker well: A relevance-theoretical account. *Journal of Pragmatics, 19*(5), 435-452. https://doi.org/10.1016/0378-2166(93)90004-9

Jucker, A. H., & Smith, S. W. (1998). And people just you know like 'wow': Discourse markers as negotiating strategies. In A. H. Jucker & Y. Ziv (Eds.), *Discourse markers: Descriptions and theory* (pp. 171-201). John Benjamins Publishing.

Jucker, A. H., & Ziv, Y. (1998). Discourse markers: Introduction. In A. H. Jucker & Y. Ziv (Eds.), *Discourse markers: Descriptions and theory* (pp. 1-12). John Benjamins Publishing.

Kasper, G. (Ed.). (1995). *Pragmatics of Chinese as native and target langauge*. University of Hawai'i Press.
Kasper, G., & Rose, K. R. (1999). Pragmatics and SLA. *Annual Review of Applied Linguistics, 19*, 81-104. https://doi.org/10.1017/S0267190599190056
Kennedy, G. (2007). An under-exploited resource: Using the BNC for exploring the nature of language learning. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 151-166). Rodopi.
Kennedy, G. D. (1998). *An introduction to corpus linguistics*. Longman. https://doi.org/10.4324/9781315843674
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7-36. https://doi.org/https://doi.org/10.1007/s40607-014-0009-9
Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics, 29*(3), 333-347. https://doi.org/10.1162/089120103322711569
Kirk, J. M. (1992). The Northern Ireland transcribed corpus of speech. In G. Leitner (Ed.), *New directions in English language corpora: Methodology, results, software developments* (pp. 65-73). Mouton de Gruyter.
Kirk, J. M., & Andersen, G. (2016). Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics, 21*(3), 291-298. https://doi.org/10.1075/ijcl.21.3.001int
Kjellmer, G. (2009). Where do we backchannel?: On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics, 14*(1), 81-112. https://doi.org/10.1075/ijcl.14.1.05kje
Knight, D. (2011). The future of mulitmodal corpora. *RBLA, Belo Horizonte, 11*(2), 391-415. https://doi.org/10.1590/S1984-63982011000200006
Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). Routledge.
Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of phonetics, 35*(2), 162-179. https://doi.org/10.1016/j.wocn.2006.04.001
Lakshmanan, U., & Selinker, L. (2001). Analysing interlanguage: How do we know what learners know? *Second Language Research, 17*(4), 393-420. https://doi.org/10.1177/026765830101700406
Lapadat, J. C. (2000). Problematizing transcription: Purpose, paradigm and quality. *International Journal of Social Research Methodology, 3*(3), 203-219. https://doi.org/10.1080/13645570050083698
Lardiere, D. (2013). Nativelike and non-nativelike attainment. In J. Herschensohn & M. Young-Scholten (Eds.), *The Cambridge hanbook of second language acquisition* (pp. 670-691). Cambridge University Press.
Larsen-Freeman, D. (2014). Another step to be taken: Rethinking the endpoint of the interlanguage continuum. In Z. Han & E. Tarone (Eds.), *Interlanguage: Forty years later* (pp. 203-220). John Benjamins Publishing. https://doi.org/10.1075/lllt.39.11ch9
Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991* (pp. 105-122). Mouton de Gruyter.
Leech, G. (1993). 100 million words of English. *English Today, 9*(1), 9-15. https://doi.org/10.1017/S0266078400006854
Leech, G. (1998). Learner corpora: What they are and what can be done with them. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). Longman.

Leech, G. (2005). Adding linguistic annotation. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 21-36). Oxbow Books. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf
Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 133-149). Rodopi.
Leech, G. (2011). Frequency, corpora and language learning. In F. Meunier, S. De Cock, G. t. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 7-32). John Benjamins Publishing.
Leech, G., Myers, G., & Thomas, J. (Eds.). (2014). *Spoken English on computer: Transcription, mark-up and application*. Routledge.
Lenk, U. (1998). Discourse markers and global coherence in conversation. *Journal of Pragmatics, 30*(2), 245-257. https://doi.org/10.1016/S0378-2166(98)00027-7
Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x
Li, C. N., & Thompson, S. A. (1977). A mechanism for the development of copula morphemes. In C. N. Li (Ed.), *Mechanisms of syntactic change* (pp. 419-444). University of Texas Press.
Li, D. C. S. (2006). Chinese as a lingua franca in greater China. *Annual Review of Applied Linguistics, 26*, 149-176. https://doi.org/10.1017/S0267190506000080
Li, S. (2016). The diversity of the original modal of discourse markers (话语标记来源模式的多样性). *Chinese Language Learning, 2*, 18-29.
Li, X. (2009). Do they tell stories differently?: Discourse marker use by Chinese native speakers and nonnative speakers. *Intercultural Communication Studies, 18*(2), 150-170.
Lin, T. (1998). From Guanhua, Guoyu to Putonghua (从官话、国语到普通话). *Language Planning*(10), 6-10.
Lindsay, J., & O'Connell, D. C. (1995). How do transcribers deal with audio recordings of spoken discourse? *Journal of Psycholinguistic Research, 24*(2), 101-115. https://doi.org/10.1007/BF02143958
Liu, B. (2009). Chinese discourse markers in oral speech of mainland Mandarin speakers. Proceedings of the 21st North American conference on Chinese linguistics (NACCL-21), Smithfield, USA.
Liu, L. (2009). "Zhege" and "nage" as discourse markers (作为话语标记的"这个"和"那个"). *Language Teaching and Linguistic Studies, 1*, 89-96.
Liu, Y. (2016). A comparison of several common transcription systems (常用口语转写系统的比较). In X. Lin, X. Xiao, & B. Zhang (Eds.), *Selected papers from the 3rd international symposium on the construction and application of Chinese interlanguage corpora* (pp. 199-204). World Publishing Corporation.
Liu, Y. (2020). Two key issues in the compilation of spoken Chinese interlanguage corpora (汉语口语中介语语料库建设中的两个关键问题). *TCSOL Studies, 1*, 47-52. https://doi.org/10.1613/j.cnki.cn44-1669/g4.2020.01.008
Lobe, B., Morgan, D., & Hoffman, K. A. (2020). Qualitative data collection in an era of social distancing. *International Journal of Qualitative Methods, 19*, 1-8. https://doi.org/10.1177/1609406920937875
Long, M. H. (1983a). Linguistic and conversational adjustments to non-native speakers. *Studies in Second Language Acquisition, 5*(2), 177-193. https://doi.org/10.1017/S0272263100004848

Long, M. H. (1983b). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics, 4*(2), 126-141. https://doi.org/10.1093/applin/4.2.126

Love, R. (2017). *The Spoken British National Corpus 2014: Design, compilation and analysis* [Unpublished Doctoral dissertation]. Lancaster University.

Love, R. (2020). *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. Routledge.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344. https://doi.org/10.1075/ijcl.22.3.02lov

Lu, J. (2009). My version of translating "Putonghua" as "Mandarin" (“普通话”译为“Mandarin”之我见). *Chinese Teaching in the World, 23*(1), 138-140.

Lu, P., Lee, J. W., & Tao, H. (2014). Discourse properties of some special sound elements and their transcription treatment (现代汉语口语中特殊话语语音成分的转写研究). *Language Sciences, 13*(2), 113-130.

Lu, Q., & Tao, J. (2011). Building a small-scale learner corpus of spoken Chinese (小型外国语学生口语中介语语料库的建立与价值). In X. Xiao & W. Zhang (Eds.), *Selected papers from the 1st international symposium on the construction and application of Chinese interlanguage corpus (首届汉语中介语语料库建设与应用国际学术讨论会论文选集)* (pp. 46-55). World Publishing Corporation.

Lv, S. (Ed.). (1999). *Eight hundred words in contemporary Chinese (现代汉语八百词)* (Revised ed.). The Commercial Press.

Ma, G. (2010). Discourse markers and pet phrases—Take "ranhou" and "danshi" as examples (话语标记与口头禅—以“然后”和“但是”为例). *Language Teaching and Linguistic Studies*(4), 69-76.

Mack, M. (1997). The monolingual native speaker: Not a norm, but still a necessity. *Studies in the Linguistic Sciences, 27*(2), 113-146. http://hdl.handle.net/2142/11591

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (Vol. 1). Lawrence Erlbaum Association.

Magliacane, A., & Howard, M. (2019). The role of learner status in the acquisition of pragmatic markers during study abroad: The use of 'like' in L2 English. *Journal of Pragmatics, 146*, 72-86. https://doi.org/10.1016/j.pragma.2019.01.026

Martinez-Garcia, M. T., & Wulff, S. (2012). Not wrong, yet not quite right: Spanish ESL students' use of gerundial and infinitival complementation. *International Journal of Applied Linguistics, 22*(2), 225-244. https://doi.org/10.1111/j.1473-4192.2012.00310.x

McCarthy, M. (1998). *Spoken Language and Applied Linguistics*. Cambridge University Press.

McCarthy, M. (2002). Good listenership made plain: British and American non-minimal response tokesn in everyday conversation. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 49-71). John Benjamins Publishing.

McCarthy, M., & Carter, R. (2001). Size isn't everything: Spoken English, corpus, and the classroom. *TESOL Quarterly, 35*(2), 337-340. https://doi.org/10.2307/3587654

McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 3-13). Routledge.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics, 39*, 74-92. https://doi.org/10.1017/S0267190519000096

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
McEnery, T., & Xiao, R. (2005). Character encoding in corpus construction. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 47-58). Oxbow Books.
McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora: What is happening? In G. Anderman & M. Rogers (Eds.), *Incorporating Corpora: The linguist and the translator* (pp. 18-31). Multilingual Matters Ltd.
McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: An advanced resource book*. Routledge. https://eprints.lancs.ac.uk/id/eprint/1836
Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge University Press.
Milroy, L., & Gordon, M. (2003). *Sociolinguistics: Method and interpretation*. Blackwell.
Müller, S. (2005). *Discourse Markers in Native and Non-native English Discourse*. John Benjamins. https://doi.org/10.1075/pbns.138
Myers, F. C. (2008). Pre-electronic corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 1-13). Walter de Gruyter.
Nelson, G. (2014). The International Corpus of English: Mark-up for spoken language. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up and applications* (pp. 220-223). Routledge.
O'Connell, D. C., & Kowal, S. (1999). Transcription and the issue of standardization. *Journal of Psycholinguistic Research, 28*(2), 103-120. https://doi.org/10.1023/A:1023265024072
Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh University Press.
Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43-72). Academic Press.
Ortega, L. (2012). Ways forward for a bi/multilingual turn in SLA. In S. May (Ed.), *The multilingual turn: Implications for SLA, TESOL and bilingual education* (pp. 32-53). Routledge.
Payne, J. (2014). The COBUILD spoken corpus: Transcription conventions. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up and applications* (pp. 203-207). Routledge.
Peppé, S. (2014). The Survey of English Usage and the London-Lund Corpus: Computerizing manual prosodic transcription. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application* (pp. 187-202). Routledge.
Peters, P., & Wong, D. (2015). Turn management and backchannels. In K. Aijmer & C. Rühlemann (Eds.), *Corpus pragmatics: A handbook* (pp. 408-428). Cambridge University Press.
Pöldvere, N. (2019). *What's in a Dialogue? On the Dynamics of Meaning-making in English Conversation* [Doctoral dissertation, Lund University]. Media-Tryck. https://portal.research.lu.se/portal/en/publications/whats-in-a-dialogue(3004710c-5d08-4069-9dc9-80544f9c3b49).html
Pöldvere, N., Paradis, C., & Johansson, V. (2019). The London-Lund Corpus 2 of spoken British English. *Lund University*. https://doi.org/https://corpora.humlab.lu.se
Psathas, G., & Anderson, T. (1990). The 'practices' of transcription in conversation analysis. *Semiotica, 78*(1-2), 75-100. https://doi.org/10.1515/semi.1990.78.1-2.75
Qi, G. Y. (2016). The importance of English in primary school education in China: Perceptions of students. *Multilingual Education, 6*(1), 1-18. https://doi.org/https://doi.org/10.1186/s13616-016-0026-0

Quan, L. (2017). A study of construction of small-sized Chinese spoken corpora (小型汉语口语语料库建设探讨). *Journal of Guandong University of Foreign Studies, 28*(4), 69-74.
Rayson, P. (2015). Computational tools and methods for corpus compilation and analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 32-49). Cambridge University Press.
Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics, 14*(3), 367-381. https://doi.org/10.1016/0378-2166(90)90095-U
Redeker, G. (1991). Linguistic markers of discourse structure. *Linguistics, 29*(6), 1139-1172. https://doi.org/10.1515/ling.1991.29.6.1139
Ritchie, H., & Roser, M. (2019). *Gender ratio*. Our World In Data. https://ourworldindata.org/gender-ratio
Rühlemann, C. (2019). How long does it take to say 'well'? Evidence from the audio BNC. *Corpus Pragmatics, 3*(1), 49-66. https://doi.org/10.1007/s41701-018-0046-y
Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Language, 50*(4), 696-735. https://doi.org/10.2307/412243
Schegloff, E. A. (1981). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing discourse: Text and talk, 71*, 93.
Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press.
Schmidt, T. (2016). Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. *International Journal of Corpus Linguistics, 21*(3), 396-418. https://doi.org/10.1075/ijcl.21.3.05sch
Schourup, L. (1999). Turorial overview: Discourse markers. *Lingua, 107*, 227-265.
Schourup, L. (2011). The discourse marker now: A relevance-theoretic approach. *Journal of Pragmatics, 43*, 2110-2129. https://doi.org/10.1016/j.pragma.2011.01.005
Schweinberger, M. (2018). The discourse particle eh in New Zealand English. *Australian Journal of Linguistics, 38*(3), 395-420. https://doi.org/10.1080/07268602.2018.1470458
Selinker, L. (2014). Interlanguage 40 years on: Three themes from here. In Z. Han & E. Tarone (Eds.), *Interlanguage: Forty years later* (pp. 229-263). John Benjamins Publishing.
Shi, J., & Hu, X. (2013). Functions of "jiushi" as a discourse marker and its grammaticalization ("就是"的话语标记功能及其语法化). *Chinese Language Learning*(4), 13-20.
Shi, R. (2020). The use of "ranhou" by intermeidate and advanced learners of Chinese (中高级水平留学生口语中"然后"的使用情况研究). *Journal of Zhejiang Sci-Tech University, 44*(2), 144-149.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
Sinclair, J. (2005). Corpus and text—Basic principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-20). Oxbow Books. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf
Sinclair, J. (2014). From theory to practice. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application* (pp. 99-109). Routledge.
Sperber, D., & Wilson, D. (1986a). Loose talk. *Proceedings of the Aristotelian Society, 86*, 153-171.
Sperber, D., & Wilson, D. (1986b). *Relevance: Communication and cognition*. Blackwell.
Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell.

Stenström, A.-B., & Svartvik, J. (1994). Imparsable speech: Repeats and other nonfluencies in spoken English. In N. Oostdijk & P. de Haan (Eds.), *Corpus-based research into language: In honour of Jan Aarts* (pp. 241-254). Rodopi.
Svartvik, J. (1980). Well in conversation. In S. Greenbaum, G. Leech, & J. Svartvik (Eds.), *Studies in English Linguistics for Randolph Quirk* (pp. 167-177). Longman.
Svartvik, J. (1982). *Survey of Spoken English: Report on research 1975-81*. Gleerup.
Svartvik, J. (Ed.). (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund University Press.
Svartvik, J., & Quirk, R. (Eds.). (1980). *A corpus of English conversation*. C.W.K. Gleerup.
Taguchi, N. (2015). Pragmatics in Chinese as a second/foreign language. *Studies in Chinese Learning and Teaching, 1*(1), 3-17.
Taguchi, N., & Roever, C. (2017). *Second language pragmatics*. Oxford University Press.
Tao, H. (2004). Fundamentals in spoken language research (口语研究的若干理论与实践问题). *Language Sciences, 3*(1), 50-67.
Temple, L. (2000). Second language learner speech production. *Studia linguistica, 54*(2), 288-297. https://doi.org/10.1111/1467-9582.00068
Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44*(2), 307-336. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x
Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279-298). John Benjamins Publishing. https://doi.org/10.1075/lllt.13.13tho
Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 59-70). Oxbow Books. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf
Thompson, P. (2010). Building a specialised audio-visual corpus. In S. Adolphs & D. Knight (Eds.), *The Routledge handbook of corpus linguistics* (pp. 93-103). Routledge.
Tian, Q. (2005). Design of the spoken corpus of learner Chinese (汉语学习者口语语料库计算机系统设计). In M. Sun & Q. Chen (Eds.), *Selected papers from the 8th Joint Symposium on Computational Linguistics (JSCL-2005)* (pp. 579-581). Tsinghua University Press.
Tono, Y. (2003). Learner corpora: Design, development and applications. Proceedings of the Corpus Linguistics 2003 Conference, Lancaster, UK.
Tottie, G. (2013). Uh and um as sociolinguistic markers in British English. In G. t. Gilquin & S. De Cock (Eds.), *Errors and disfluencies in spoken corpora* (pp. 33-57). John Benjamins Publishing.
Tseng, S.-C. (2004). Processing Mandarin spoken corpora. *Traitement Automatique des Langes*, 89-108.
Tseng, S.-C., & Gibbon, D. (2006). Discourse functions of duration in Mandarin: Resource design and implementation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy.
Varonis, E. M., & Gass, S. M. (1985). Miscommunication in native/nonnative conversation. *Language in Society, 14*(3), 327-343. https://doi.org/10.1017/S0047404500011295
Wang, D. (2016). Guanhua, Guoyu, and Putonghua: Politics and "proper names" for standard language in modern China. *Chinese Studies in History, 49*(3), 152-174. https://doi.org/10.1080/00094633.2015.1174972
Wang, J., & Halenko, N. (2019). Second language pragmatics. *East Asian Pragmatics, 4*(1), 1-9. https://doi.org/10.1558/eap.38206

Wang, L. (1999). From Guanhua to Guoyu and Putonghua—The development of modern national language (从官话到国语和普通话—现代汉语民族共同语的形成及发展). *Language Planning*(6), 22-25.
Wang, W. (2018). Discourse uses and prosodic properties of ranhou in spontaneous Mandarin conversation. *Chinese Language and Discourse, 9*(1), 1-25. https://doi.org/10.1075/cld.00006.wan
Wang, W., & Zhou, W. (2005). On extension of the word "ranhou" in modern spoken Chinese and its mechanism (“然后”一词在现代汉语口语中使用范围的扩大及其机制). *Chinese Language Learning, 8*(4), 31-39.
Wang, Y. (2016). The construction of spoken interaction corpus of Mandarin Chinese (汉语口语互动分级语料库的构建). *Computer Engineering & Science, 38*(2), 395-400.
Wang, Z., & Yang, W. (2011). *Annotation of the HSK Dynamic Spoken Corpus (HSK 动态口语语料库的标注研究)* The 7th National Applied Linguistics Symposium, Hunan, China.
Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis*. Wiley Blackwell. https://doi.org/10.1002/9781119180180
White, S. (1989). Backchannels across cultures: A study of Americans and Japanese. *Language in Society, 18*(1), 59-76. https://doi.org/10.1017/S0047404500013270
Wichmann, A. (2008). Speech corpora and spoken corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 187-207). Walter de Gruyter.
Xiao, X., & Zhou, W. (2014). Comprehensiveness and categorization issues of Chinese interlanguage corpus annotation (汉语中介语语料库的全面性及类别问题). *Chinese Teaching in the World, 28*(3), 368-377.
Xu, H., Lu, X., & Brezina, V. (2019). Acquistion of the Chinese particle le by L2 learners: A corpus-based approach. In X. Lu & B. Chen (Eds.), *Computational and corpus approaches to Chinese language learning* (pp. 197-216). Springer.
Xu, J. (2008). Discourse marker "na(ge)" and its functions in spoken Chinese (汉语自然会话中话语标记“那（个）”的功能分析). *Linguistic Sciences, 7*(1), 49-57.
Xu, J. (2009). The discourse marker "ranhou" and its functions in spoken Chinese (汉语自然会话中“然后”的话语功能分析). *Foreign Language Research*(2), 9-15.
Xu, J., Wang, C., Song, J., & Guo, D. (2021). Data collection and corpus construction (英语母语者汉语口语语料的采集分析与语料库构建). *Journal of Yunnan Normal University (Teaching and Research of Chinese as A Foreign Language), 19*(1), 13-25.
Xu, W., & Fang, S. (2019). The translations of terms and China's image in the international context: A corpus-based approach to the use of "Putonghua" (国际语境下的术语翻译与中国形象构建—基于语料库的“普通话”概念词使用实证研究). *Foreign Languages in China, 16*(2), 91-103.
Yang, Y., Li, S., Guo, Y., & Tian, Q. (2006). Constructing the spoken corpus of learner Chinese (建立汉语学习者口语语料库的基本设想). *Linguistic Research, 3*, 58-64.
Yao, S., & Yao, X. (2012). On the emerging of a new function of "jiushi" as discourse markers in authentic speech (自然口语中“就是”话语标记功能的浮现). *Chinese Teaching in the World, 26*(1), 77-84.
Yue, Y. (2020). A Chinese conversational discourse approach to the extended uses of nage (指示与非指示：汉语言谈交际中“那个”的用法). *Language Teaching and Linguistic Studies, 1*, 48-61.
Zeng, S., & Liu, Y. (2002). *The Mandarin Conversational Dialogue Corpus: Phonetic transcription conventions*.

Zhang, B. (2012). Design of the HSK Dynamic Spoken Corpus ("HSK 动态口语语料库"总体设计). In B. Zhang & J. Wang (Eds.), *Language testing: From an interdisciplinary perspective* (pp. 239-258). Sinolingua.

Zhang, B. (2013). Reconsidering the modes of annotation of general Chinese interlanguage corpora (关于通用型汉语中介语语料库标注模式的再认识). *Chinese Teaching in the World, 27*(1), 128-140.

Zhang, B. (2019). *Research on tagging of Chinese interlanguage corpus (汉语中介语语料库标注规范研究)*. Peking University Press.

Zhang, B., & Cui, X. (2013a). Challenges and solutions in the compilation of Chinese interlanguage corpus (汉语中介语语料库建设面临的任务与对策). In X. Cui & B. Zhang (Eds.), *Selected papers from the 2nd international symposium on the construction and application of Chinese interlanguage corpus (第二届汉语中介语语料库建设与应用国际学术讨论会论文选集)* (pp. 31-43). Beijing Language and Culture University Press.

Zhang, B., & Cui, X. (2013b). Design of the International Corpus of Learner Chinese ("全球汉语中介语料库建设和研究"的设计理念). *Language Teaching and Linguistic Studies*(5), 27-34.

Zhang, B., & Cui, X. (2015). On the standards of building a Chinese interlanguage corpus (谈汉语中介语语料库的建设标准). *Applied Linguistics, 5*(2), 125-134.

Zhang, L. (2016). Transcription principles for spoken Chinese interlanguge corpora (汉语中介语口语语料库转写规则初探). In X. Lin, X. Xiao, & B. Zhang (Eds.), *Selected papers from the 3rd international symposium on the construction and application of Chinese interlanguage corpus (第三届汉语中介语语料库建设与应用国际学术讨论会论文选集)* (pp. 205-215). World Publishing Corporation.

Zhang, W., & Gao, H. (2012). Functions of "jiushi(就是)" in everyday conversation and its grammaticalization (自然会话中"就是"的话语功能与语法化研究). *Language Teaching and Linguistic Studies, 1*, 91-98.

Zhang, Y. (2002). The coherent function of "jiushi" in text and the evolution of its grammaticalization ("就是"的篇章衔接功能及其语法化历程). *Chinese Teaching in the World*(3), 80-91.

Zhao, H., & Lin, J. (2019). Some thoughts on the annotation codes of Chinese interlanguage corpora (关于汉语中介语语料库标注代码的思考). *Overseas Chinese Education, 1*, 103-114.

Zheng, G. (2020). Three form factors affecting the function of Chinese discourse markers (影响汉语话语标记功能表达的三个形式因素). *Chinese Language Learning, 2*, 17-27.

Zhou, B. (2011). An overview of studies on L2 Chinese phonetics—Building a spoken L2 Chinese corpus (汉语中介音研究综述—兼谈汉语自然口语语料库的建立). In X. Xiao & W. Zhang (Eds.), *Selected papers from the 1st international symposium on the construction and application of Chinese interlanguage corpus (首届汉语中介语语料库建设与应用国际学术讨论会论文选集)* (pp. 319-329). World Publishing Corproration.

Zhu, D. (2017). Default semantics construction of the conversational narrative marker "ranhou" (自然会话叙事标记语"然后"构建的语义缺省). *Foreign Language Research*(5), 50-57.

# Appendices

## Appendix A Existing Spoken L2 Chinese Corpora

This appendix provides a survey of the existing spoken L2 Chinese corpora which covers all publicly attenable information about these corpora at the time of writing. However, this survey is of necessity far from complete, since it is unlikely to reach all spoken L2 corpus in existence. In the Chinese research literature, it is quite often the case that researchers tend to build and use their own corpora in research without giving sufficient descriptions of those corpora to the academic community. As a result, this survey only includes nine corpora that have been received much attention in the academic community. In what follows, I will introduce them briefly in order.

**The Learner Corpus of Spoken Chinese (汉语学习者口语语料库).** This corpus was compiled at Beijing Language and Culture University, which is the first corpus of spoken L2 Chinese to date (Tian, 2005; Yang et al., 2006). Although this L2 corpus has been mentioned frequently in studies on Chinese L2 corpora, Zhang (2012) rightly points out that this L2 corpus in fact has never been used to carry out any further research except the publication of Tian (2005) and Yang et al. (2006).

**Table A-1**

*The Learner Corpus of Spoken Chinese (汉语学习者口语语料库)*

| | |
|---|---|
| **Compiled at:** | Beijing Language and Culture University |
| **Details of material:** | unknown |
| **Participants:** | unknown |
| **Metadata:** | unknown |
| **No. of recordings:** | unknown |
| **Size of transcribed texts:** | unknown |
| **How transcribed:** | manually transcribed into written texts |
| **Annotation:** | unknown |
| **Organisation:** | unknown |
| **Availability:** | not available |
| **Use of corpus:** | second language acquisition |
| **How analysed:** | unknown |
| **References:** | Tian (2005); Yang et al. (2006) |

**The HSK Dynamic Spoken Corpus (HSK 动态口语语料库).** This corpus is an on-going project, which is compiled at Beijing Language and Culture University (Hu & Wang, 2011; Wang & Yang, 2011; Zhang, 2012). HSK is short for *Hanyu Shuiping Kaoshi* (Chinese proficiency test)[57], which is an international standardized test of Chinese language proficiency, aiming to assess non-native Chinese speakers' listening, reading and writing skills in using Chinese. The speaking test of HSK is called HSKK (*Hanyu Shuiping Kouyu Kaoshi*–Chinese proficiency speaking test). This corpus includes approximately 5.5 million words of HSKK tests gathered between 1993 and 2011. The main task of this corpus is to process the large proportion of data gathered from 2001 to 2010 (Zhang, 2012, p. 251). Table A-2 shows the main component of this corpus. This corpus is not accessible at the time of writing.

**Table A-2**

*Component of the HSK Dynamic Spoken Corpus (2001–2010)*

| Countries | No. of the tests | No. of the words |
|---|---|---|
| Japan and Korea | 2,000 | 1,000,000 |
| South-east Asia | 1,000 | 500,000 |
| Other Asian countries | 500 | 250,000 |
| English-speaking countries | 1,000 | 500,000 |
| Slavic groups | 500 | 250,000 |
| Other European countries and other countries in the Americas | 500 | 250,000 |
| **Total** | **5,500** | **2.75 million** |

**The International Corpus of Learner Chinese (全球汉语学习者语料库).** This L2 corpus is an on-going project which is led by Beijing Language and Culture University, including L2 Chinese production provided by many universities around the world. As Table A-3 shows, this L2 corpus contains both a written part and a spoken part which is asserted to be the largest L2 Chinese corpus in the world. This corpus is publicly accessible to the academic community via its official website (which is given in Table A-3).

[57] *Hanyu Shuiping Kaoshi* (Chinese proficiency test): www.chinesetest.cn/index.do.

**Table A-3**

*The International Corpus of Learner Chinese (全球汉语学习者语料库)*

| | |
|---|---|
| **Compiled at:** | Beijing Language and Culture University |
| **Details of material:** | written language: 45 million words; Spoken language: 5 million words |
| **Participants:** | unknown |
| **Metadata:** | name (ID), gender, age, nationality, Chinese ethnicity, first language, language proficiency, grade, learning history etc. |
| **No. of recordings:** | unknown |
| **Size of transcribed texts:** | unknown |
| **How transcribed:** | unknown |
| **Annotation:** | error tagged, and annotated correct usage |
| **Availability:** | partly available |
| **Use of corpus:** | various applications |
| **How analysed:** | unknown |
| **Website:** | http://qqk.blcu.edu.cn/#/login |
| **Reference:** | Cui and Zhang (2011a); Zhang and Cui (2013a, 2013b) |

**The Jinan Learner Corpus of Spoken Chinese (暨南大学华文学院口语语料库).** This spoken L2 corpus can be accessed at https://huayu.jnu.edu.cn/corpus5/Default.aspx. It contains recordings of learners of Chinese who come from 22 countries all around the world. These learners of Chinese are at different Chinese proficiency levels. No further information can be found.

**The Longitudinal Chinese Interlanguage Corpus (外国留学生汉语口语纵向语料库).** This corpus is an on-going project, which compiled at Nanjing Normal University (Cao & Wu, 2013). It is designed to contain one million written data and one million spoken data (see Table A-4). The spoken component of this corpus has gathered over 400 hours of recordings provided by more than 100 learners, including (i) about 200 hours of audio recordings of dialogues provided by about 10 L2 Chinese learners , and (ii) approximately 200 hours of audio and video recordings of OPI (Oral Proficiency Interview) test, oral reports and class discussions produced by more than 90 L2 learners of Chinese (Cao, 2013). Corpus compilers follow the OPI modal to conduct the conversations. Each conversation is designed as a 60-mintue interview between one researcher and one L2 learner of Chinese covering various topics (e.g., daily life, personal experience). The data collection lasts for six months to one year.

**Table A-4**

*The Longitudinal Chinese Interlanguage Corpus (外国留学生汉语口语纵向语料库)*

| | |
|---|---|
| **Compiled at:** | Nanjing Normal University |
| **Details of material:** | audio recordings of dialogues<br>audio and video recordings of oral reports, OPI tests, class discussions |
| **Participants:** | over 100 students |
| **Metadata:** | unknown |
| **No. of recordings:** | unknown |
| **Size of transcribed texts:** | transcribed 1,500,000 words of 140 hours of data produced by six students |
| **How transcribed:** | manually transcribed into written texts |
| **Annotation:** | unknown |
| **Organisation:** | dialogues, oral reports and tests (200 hours)<br>class discussions (200 hours) |
| **Availability:** | not available |
| **Use of corpus:** | second language acquisition |
| **How analysed:** | unknown |
| **Reference:** | Cao (2013); Cao and Wu (2013) |

**The Guangwai–Lancaster Chinese Learner Corpus (GLCLC).** This corpus is a 1.2-million-word corpus of learner Mandarin Chinese, which is a result of the collaboration between Guangdong University of Foreign Studies (GDUFS) and Lancaster University, represents a new addition to corpora of L2 Chinese (see Table A-5). The corpus has both a spoken part (642,385 tokens, 49.62%) and a written part (652,329 tokens, 52.38%) and covers a variety of task types and topics. The spoken texts comprise interactions typically between L1 and L2 speakers of Chinese and involve one, two or multiple speakers. This corpus is fully error tagged. It can be used to explore various theoretical and practical issues pertaining to the acquisition of Chinese as a foreign language.

**Table A-5**

*The Guangwai–Lancaster Chinese Learner Corpus (GLCLC)*

| | |
|---|---|
| **Compiled at:** | Guangdong University of Foreign Studies and Lancaster University |
| **Details of material:** | The spoken component draws spoken data from L2 learners in four task types: (1) oral tests administered by a native Chinese instructor to 1–3 test-takers (379,839 tokens), (2) interviews conducted by a native Chinese speaker with individual advanced learners (119,982 tokens), (3) free talks (monologues) on the topic "My Hometown" or "A Memorable Trip" (81,630 tokens), and (4) tutorials given by a native speaker to individual learners (60,934 tokens). |
| **L2 participants:** | 1,492 L2 Chinese learners come from 107 countries, and are students of Chinese as a second/foreign language at GDUFS |
| **Proficiency level:** | beginner, intermediate, and advanced |
| **L1 participants:** | L1 Chinese speakers are staff and students at GDUFS |
| **Metadata:** | XML format with rich metadata |
| **No. of recordings:** | unknown |
| **Size of transcribed texts:** | 621,990 words |
| **How transcribed:** | unknown |
| **Annotation:** | error tagged, and annotated correct usage |
| **Availability:** | available on Sketch Engine |
| **Use of corpus:** | various applications |
| **How analysed:** | unknown |
| **Website:** | https://www.sketchengine.eu/guangwai-lancaster-chinese-learner-corpus/ |
| **References:** | Chen and Xu (2019), Xu et al. (2019); http://cass.lancs.ac.uk/tag/guangwai-lancaster-chinese-learner-corpus/ |

**The Soochow L2 Corpus**. Lu and Tao (2011) introduce a small-scale L2 corpus of spoken Chinese compiled at Soochow University. This L2 corpus consists of recordings of oral tests produced by 39 students at the Chinese University of Hong Kong, and recordings of oral tests and impromptu speech of 74 students with various proficiency levels in School of Oversees Education at Soochow University in 2009. This corpus is not available to the public at the time of writing. More information is provided in Table A-6.

**Table A-6**

*The Soochow L2 Corpus*

| | |
|---|---|
| **Compiled at:** | Soochow University |
| **Details of material:** | oral tests and impromptu speech |
| **Participants:** | 113 students |
| **Metadata:** | unknown |
| **No. of recordings:** | 689 |
| **Size of transcribed texts:** | approximate 300,000 words |
| **How transcribed:** | written texts |
| **Annotation:** | manually annotated errors, including grammar and phonetic errors |
| **Availability:** | not available |
| **Use of corpus:** | error analysis, contrastive analysis of various learner groups |
| **How analysed:** | raw data and annotated data produced |
| **Reference:** | Lu and Tao (2011) |

**The Multimodal Corpus of L2 Chinese (外国学生多模态口语语料库)**. This is an L2 multimodal corpus. Table A-7 gives some basic information of this corpus. Unfortunately, there is no further information in terms of the creation of this multimodal corpus can be found.

**Table A-7**

*The Multimodal Corpus of L2 Chinese (外国学生多模态口语语料库)*

| | |
|---|---|
| **Compiled by:** | a team led by Xiao Xiqiang |
| **Compiled at:** | Nanjing Normal University |
| **Sampling period:** | 2010-2011 (two semesters) |
| **Details of material:** | oral tests and speaking tasks/activities in classroom |
| **Participants:** | students (learners of Chinese) in the International College of Chinese Studies at Nanjing Normal University |
| **Proficiency level:** | beginners, intermediate and advanced learners of Chinese |
| **Size:** | unknown |
| **Data collection:** | Audio and video recordings |
| **Place of data collection:** | classrooms |
| **Metadata collection:** | unknown |
| **How transcribed:** | written texts; paralinguistic features are omitted |
| **Use of corpus:** | the high-quality recordings can be used to build a phonological corpus |
| **Availability:** | not available |
| **References:** | Zhou (2011) |

**The Country–Specific Corpus of Spoken Chinese Interlanguage (Korean Learners) (韩国学习者汉语中介语口语语料库)**. This spoken corpus is compiled at Ludong University. Table A-8 shows some information about this spoken corpus. Unlike other L2 Chinese corpora which contain multiple L1 backgrounds for L2 learners of Chinese, this corpus covers L2 learners of Chinese with one single L1 background: all L2 participants are Korean learners of Chinese.

**Table A-8**

*The Country–Specific Corpus of Spoken Chinese Interlanguage (韩国学习者汉语中介语口语语料库)*

| | |
|---|---|
| **Compiled at:** | Ludong University |
| **Details of material:** | Korean HSKK (*Hanyu Shuiping Kaoshi*–Chinese Proficiency Speaking Test) |
| **Participants：** | Korean learners of Chinese at three proficiency levels (beginner, intermediate, and advanced) |
| **Metadata:** | unknown |
| **No. of recordings:** | about 1,5000 recordings |
| **Size of transcribed texts:** | unknown |
| **How transcribed:** | manually transcribed into written texts |
| **Annotation:** | annotated errors, including grammar and phonetic errors; annotated correct usage as well |
| **Availability:** | not available |
| **Storage details:** | unknown |
| **Use of corpus:** | error analysis |
| **How analysed:** | annotated data |
| **References:** | Hu and Xu (2020) |

# Appendix B Massey University Human Ethics Approval

Date: 05 January 2018

Dear Lin Li

Re: Ethics Notification - **4000018814** - **Chinese collocations used by learners in spoken texts**

Thank you for your notification which you have assessed as Low Risk.

Your project has been recorded in our system which is reported in the Annual Report of the Massey University Human Ethics Committee.

The low risk notification for this project is valid for a maximum of three years.

If situations subsequently occur which cause you to reconsider your ethical analysis, please contact a Research Ethics Administrator.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

**A reminder to include the following statement on all public documents:**
*"This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.*

*If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director - Ethics, telephone 06 3569099 ext 86015, email humanethics@massey.ac.nz. "*

Please note, if a sponsoring organisation, funding authority or a journal in which you wish to publish requires evidence of committee approval (with an approval number), you will have to complete the application form again, answering "yes" to the publication question to provide more information for one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

Yours sincerely

Dr Brian Finch
Chair, Human Ethics Chairs' Committee and Director (Research Ethics)

**Research Ethics Office, Research and Enterprise**
Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand **T** 06 951 6841; 06 95106840
**E** humanethics@massey.ac.nz; animalethics@massey.ac.nz; gtc@massey.ac.nz

## Appendix C Information Sheet for L1 Speakers of Chinese

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

**School of Humanities**
**Massey University**
**Private Bag 11222**
**Palmerston North 4442**
**New Zealand**

# *Chinese collocations used by native speakers in spoken texts*

### INFORMATION SHEET

**Researcher Introduction**

My name is Li Lin and I am a Chinese PhD candidate in Applied Linguistics in the School of Humanities at Massey University.

**Project Description and Invitation**

In every language, certain words often go together to express specific meanings. For example, we say *black tea* rather than *red tea* in English, while *红茶 hong cha* [red tea] rather than *黑茶 hei cha* [black tea] is more acceptable in Chinese. *Black* and *tea*, *红* and *茶 are* often used together, so we can call *black tea* and *红茶* collocations in linguistics. Using collocations enables learners to express ideas fluently and fulfil the communicative needs. Therefore, I am trying to find out collocations used by non-native speakers of Chinese when they communicate with Chinese native speakers.

To do this, I would like to record conversations with approximately 10 or more Chinese native speakers, and I am contacting you to join. I hope that this will be mutually interesting. We can begin with three conversations over two weeks, but if you find the process interesting we can continue after that. For each conversation, we can choose together topics that interest you and each will be 30 to 60 minutes long. If you would be willing to talk to me, we can arrange a time that is convenient to you, and we can talk online or in person, whichever you prefer.

**The Recordings**

I will record all conversations and transcribe them, but you can turn off the recorder any time you want to. All the recorded conversations will be collected and contribute to the construction of a Chinese spoken corpus (a collection of spoken language stored as written transcriptions in an electronic form) which will be publicly available. All the recordings will be stored on a password protected computer and will be listened to only by myself and my supervisors. I will use the information that I gain from the corpus in my doctoral thesis and in other articles and presentations. However, I will not use your name and I will make sure that nobody can identify the participants involved. If you want, I can give you the recordings and transcriptions.

**Participant's Rights**

You are under no obligation to accept this invitation. If you decide to participate, you have the right to:

- decline to answer any particular questions;
- withdraw from the study at any time before the recording of the first conversation;
- ask any questions about the study at any time during participation;
- provide information on the understanding that your name will not be used unless you give permission to the researcher;
- be given access to a summary of the project findings when it is concluded.

**Project Contacts**

You can ask me questions about the research before you agree to take part. You can contact me by e-mail (L.Lin4@massey.ac.nz) or telephone (+64 021 031 9987).

Or you can ask one of my supervisors:

- Dr. Gillian Skyrme
  E-mail:G.R.Skyrme@massey.ac.nz
  Telephone: +64 (06) 356 9099 ext 83572
  Campus: Manawatu
- Dr. Michael Li
  E-mail:S.Li.1@massey.ac.nz
  Telephone: +64 (09) 414 0800 ext 43368
  Campus: Albany
- Dr. Tony Fisher
  E-mail: A.Fisher@massey.ac.nz
  Telephone: +64 (04) 801 5799 ext 63572
  Campus: Wellington
- Prof. Cynthia White
  E-mail: C.J.White@massey.ac.nz
  Telephone: +64 (06) 356 9099 ext 83565
  Campus: Manawatu

This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr. Brian Finch, Director (Research Ethics), email humanethics@massey.ac.nz.

**Appendix D Information Sheet for L2 Speakers of Chinese**

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

**School of Humanities**
**Massey University**
**Private Bag 11222**
**Palmerston North 4442**
**New Zealand**

# *Chinese collocations used by non-native speakers in spoken texts*

## INFORMATION SHEET

### Researcher Introduction

My name is Li Lin and I am a Chinese PhD candidate in Applied Linguistics in the School of Humanities at Massey University.

### Project Description and Invitation

In every language, certain words often go together to express specific meanings. For example, we say *black tea* rather than *red tea* in English, while *红茶 hong cha* [red tea] rather than *黑茶 hei cha* [black tea] is more acceptable in Chinese. *Black* and *tea*, *红* and *茶 are* often used together, so we can call *black tea* and *红茶* collocations in linguistics. Using collocations is important for learners, as it enables them to express ideas fluently and fulfil the communicative needs. Therefore, I am trying to find out collocations used by learners of Chinese when they communicate with Chinese native speakers.

To do this, I would like to record conversations with approximately 20 or more intermediate and advanced learners of Chinese, and I am contacting you to join. I hope that this will be mutually interesting and beneficial. I can collect some examples of your Chinese, and you can practice Chinese talking to a native speaker. I will follow up the conversation with some written feedback to you to help you improve your speaking skills if you would like me to. We can begin with three conversations over two weeks, but if you find the process useful we can continue after that. For each conversation, we can choose together topics that interest you and each will be 10 to 15 minutes long. If you would be willing to talk to me, we can arrange a time that is convenient to you, and we can talk online or in person, whichever you prefer.

### The Recordings

I will record all conversations and transcribe them, but you can turn off the recorder any time you want to. All the recorded conversations will be collected and contribute to the construction of a Chinese spoken corpus (a collection of spoken language stored as written transcriptions in an electronic form) which will be publicly available. All the recordings will be stored on a password protected computer and will be listened to only by myself and my supervisors. I will use the information that I gain from the corpus in my doctoral thesis and in other articles and presentations.

However, I will not use your name and I will make sure that nobody can identify the participants involved. If you want, I can give you the recordings and transcriptions.

**Participant's Rights**

You are under no obligation to accept this invitation. If you decide to participate, you have the right to:

- decline to answer any particular questions;
- withdraw from the study at any time before the recording of the first conversation;
- ask any questions about the study at any time during participation;
- provide information on the understanding that your name will not be used unless you give permission to the researcher;
- be given access to a summary of the project findings when it is concluded;
- ask for feedback from the researcher to help you improve your oral skills.

**Project Contacts**

You can ask me questions about the research before you agree to take part. You can contact me by e-mail (L.Lin4@massey.ac.nz) or telephone (+64 021 031 9987).

Or you can ask one of my supervisors:

- Dr. Gillian Skyrme
  E-mail:G.R.Skyrme@massey.ac.nz
  Telephone: +64 (06) 356 9099 ext 83572
  Campus: Manawatu
- Dr. Michael Li
  E-mail:S.Li.1@massey.ac.nz
  Telephone: +64 (09) 414 0800 ext 43368
  Campus: Albany
- Dr. Tony Fisher
  E-mail: A.Fisher@massey.ac.nz
  Telephone: +64 (04) 801 5799 ext 63572
  Campus: Wellington
- Prof. Cynthia White
  E-mail: C.J.White@massey.ac.nz
  Telephone: +64 (06) 356 9099 ext 83565
  Campus: Manawatu

This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr. Brian Finch, Director (Research Ethics), email humanethics@massey.ac.nz.

## Appendix E Consent Form

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

**School of Humanities**
**Massey University**
**Private Bag 11222**
**Palmerston North 4442**
**New Zealand**

# *Chinese collocations used by native/non-native speakers in spoken texts*

### PARTICIPANT CONSENT FORM – INDIVIDUAL

I have read the Information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I agree/do not agree to the interview being sound recorded.

I wish/do not wish to have my recordings returned to me.

I wish/do not wish to have data placed in an official archive.

I agree to participate in this study under the conditions set out in the Information Sheet.

**Signature:** ..... **Date:** .....

**Full Name - printed** .....

## Appendix F Metadata Information on L1 Participants and Conversations

This appendix shows the metadata categories gathered in terms of the L1 participants and informal conversations conducted for the creation of the spoken L1 corpus. It should be noted that not all the audio recordings gathered were included in the L1 corpus. Metadata categories of L1 speakers listed sequentially are speaker ID, gender, age, birthplace (province), region, and current country of residence; metadata categories of recordings listed in turn are recording ID, date, method, location (participant-researcher), and length.

**S00 female, age c.28, Shandong, northern China, New Zealand**

**N01 female, age c.28, Hebei, northern China, China**

N01-C01 **(**02/04/2018) WeChat audio, home-home, 1:05:02

N01-C02 (09/04/2018) WeChat audio, home-home, 31:07

**N02 female, age c.28, Hunan, southern China, New Zealand**

N02-C01 (02/04/2018) WeChat audio, home-home, 37:39

N02-C02 (09/04/2018) face-to-face, campus, 27:54

N02-C03 (09/04/2018) face-to-face, office, 34:20

**N03 male, age c.28, Hebei, northern China, China**

N03-C01 (05/04/2018) WeChat audio, home-office, 21:17

**N04 male, age c.28, Shandong, northern China, Sweden**

N04-C01 (19/04/2018) WeChat audio, home-office, 42:33

N04-C02 (07/05/2018) WeChat audio, home-office, 44:33

**N05 female, age c.32, Hebei, northern China, China**

N05-C01 (24/04/2018) WeChat audio, home-office, 20:48

**N06 female, age c.36, Beijing, northern China, New Zealand**

N06-C01 (29/04/2018) face-to-face, library, 25:34

**N07 female, age c.35, Beijing, northern China, New Zealand**

N07-C01 (03/05/2018) face-to-face, library, 36:28

**N08 female, age c.36, Shaanxi, northern China, China**

N08-C01 (01/07/2018) WeChat audio, home-office, 30:53

**N09 female, age c.28, Shandong, northern China, China**

N09-C01 (08/07/2018) WeChat audio, home-office, 29:48

**N10 female, age c.31, Shandong, northern China, China**

N10-C01 (08/07/2018) WeChat audio, home-office, 36:44

**N11 female, age c.25, Beijing, northern China, New Zealand**

N11-C01 (09/07/2018) WeChat audio, home-office, 32:57

**N12 female, age c.25, Shandong, northern China, China**

N12-C01 (09/07/2018) WeChat audio, home-office, 34:04

**N13 male, age c.27, Hebei, northern China, China**

N13-C01 (09/07/2018) WeChat audio, home-home, 37:13

**N14 female, age c.28, Henan, northern China, China**

N14-C01 (13/07/2018) WeChat audio, home-office, 37:17

**N15 male, age c.35, Beijing, northern China, China**

N15-C01 (14/07/2018) WeChat audio, home-home, 39:18

**N16 female, age c.28, Jiangxi, southern China, China**

N16-C01 (14/07/2018) WeChat audio, office-home, 28:17

**N17 male, age c.27, Jiangxi, southern China, China**

N17-C01 (15/07/2018) WeChat audio, home-office, 27:49

**N18 male, age c.27, Henan, northern China, China**

N18-C01 (15/07/2018) WeChat audio, home-office, 41:28

**N19 female, age c.27, Sichuan, southern China, China**

N19-C01 (15/07/2018) WeChat audio, home-home, 32:09

**N20 male, age c.26, Liaoning, northern China, China**

N20-C01 (16/07/2018) WeChat audio, office-home, 46:33

**N21 female, age c.26, Hunan, southern China, China**

N21-C01 (17/07/2018) WeChat audio, home-home, 27:29

**N22 male, age c.35, Sichuan, southern China, China**

N22-C01 (20/07/2018) WeChat audio, office-home, 35:51

**N23 male, age c.27, Shandong, northern China, China**

N23-C01 (21/07/2018) WeChat audio, office-home, 30:05

**N24 male, age c.35, Shandong, northern China, China**

N24-C01 (21/07/2018) WeChat audio, home-home, 49:18

**N25 female, age c.24, Hunan, southern China, China**

N25-C01 (21/07/2018) WeChat audio, home-home, 27:52

## Appendix G Metadata Information on L2 Participants and Conversations

This appendix shows the metadata categories gathered in terms of the L2 participants and informal conversations conducted for the creation of the spoken L2 corpus. It should be noted that not all the audio recordings gathered were included in the L2 corpus. Metadata categories of L2 speakers listed sequentially are speaker ID, gender, age, and nationality; metadata categories of recordings listed in turn are recording ID, date, method, location (participant-researcher), and length.

**L01 male, age c.32, New Zealand**

L01-C01 **(**12/04/2018) Skype video, home-home, 30:55

L01-C02 (19/04/2018) Skype video, home-home, 48:23

L01-C03 **(**03/05/2018) Skype video, home-home, 1:01:25

L01-C04 (10/05/2018) Skype video, home-home, 42:51

**L02 male, age c.24, New Zealand**

L02-C01 **(**12/05/2018) WeChat audio, home-home, 18:26

L02-C02 (20/05/2018) WeChat audio, home-home, 26:27

**L03 female, age c.24, New Zealand**

L03-C01 **(**21/05/2018) Skype video, home-office, 18:58

L03-C02 (07/06/2018) Skype video, home-office, 30:51

**L04 female, age c.22, New Zealand**

L04-C01 **(**25/05/2018) WeChat video, office-office, 25:32

L04-C02 (07/06/2018) WeChat video, office-office, 44:43

**L05 Female, age c.24, New Zealand**

L05-C01 **(**22/05/2018) WeChat video, dorm-office, 33:06

L05-C02 (25/05/2018) WeChat video, dorm-office, 29:20

L05-C03 (28/05/2018) WeChat video, dorm-office, 37:36

**L06 female, age c.35, New Zealand**

L06-C01 **(**29/05/2018) face-to-face, library, 40:52

L06-C02 (30/05/2018) face-to-face, library, 24:51

L06-C03 (12/06/2018) WeChat video, home-office, 1:19:58

**L07 male, age c.32, New Zealand**

L07-C01 **(**29/05/2018) WeChat audio, street-home, 30:08

L07-C02 (14/06/2018) WeChat audio, street-home, 23:01

**L08 male, age c.40, New Zealand**

L08-C01 **(**01/06/2018) WeChat audio, office-home, 31:37

**L09 male, age c.24, Australia**

L09-C01 **(**02/06/2018) WeChat audio, café-office, 31:47

L09-C02 (07/06/2018) Skype video, home-home, 13:06

L09-C03 (09/06/2018) Skype video, home-office, 40:40

**L10 male, age c.32, New Zealand**

L10-C01 **(**04/06/2018) Skype video, home-home, 14:31

**L11 male, age c.40, New Zealand**

L11-C01 **(**04/06/2018) face-to-face, library, 31:24

L11-C02 (19/06/2018) WeChat audio, home-home, 36:02

L11-C03 (25/06/2018) WeChat audio, home-home, 40:37

**L12 male, age c.30, New Zealand**

L12-C01 **(**14/06/2018) WeChat audio, home-home, 44:36

L12-C02 (02/07/2018) WeChat audio, office-home, 40:20

**L13 male, age c.24, New Zealand**

L13-C01 **(**15/06/2018) WeChat audio, home-office, 30:40

L13-C02 (19/06/2018) WeChat audio, home-office, 43:07

L13-C03 (26/06/2018) WeChat audio, home-office, 47:40

**L14 male, age c.40, New Zealand**

L14-C01 **(**27/06/2018) WeChat audio, home-office, 19:48

L14-C02 (19/06/2018) WeChat audio, home-office, 23:19

L14-C03 (29/06/2018) WeChat audio, home-office, 12:59

**L15 male, age c.30, New Zealand**

L15-C01 **(**29/06/2018) WeChat audio, office-home, 12:20

L15-C02 (03/07/2018) WeChat audio, office-home, 19:15

**L16 male, age c.24, Australia**

L16-C01 **(**03/07/2018) WeChat audio, home-office, 30:38

**L17 female, age c.18, New Zealand**

L17-C01 **(**16/08/2018) WeChat audio, home-office, 27:51

**L18 male, age c.24, New Zealand**

L18-C01 **(**30/07/2018) WeChat audio, café -office, 39:38

**L19 male, age c.30, New Zealand**

L19-C01 **(**14/08/2018) WeChat audio, home-home, 42:52

## Appendix H The Spoken Chinese Transcription Scheme

This appendix provides the transcription scheme created for the spoken Chinese corpus.

**Table H-1**

*The Spoken Chinese Transcription Scheme*

| Feature | Transcription guideline | Example |
|---|---|---|
| **Speaker identification** | L2 participants are identified as <L01>, <L02> and so on; L1 participants are labelled as <N01>, <N02> and so on; use <S00> for the researcher in all the conversations.<br>Speaker labels are followed by one space. | (1) <S00> 去学习嘛<br><N01> 去了江西<br>(2) <L09> 挺难的<br><S00> 他们都说汉字很难 |
| **Pinyin** | Use Pinyin to represent backchannels, such as *eng* in (1), and non-standard pronunciation, for example *guan* in (1); to mark truncated words, e.g., *xi* in (2); and to represent tongue slips, e.g., *liu* in (3). | (1) <L01> ***eng eng*** 对对对然后我的岗位也可以说是 ***guan*** 长的那种的位子<br>(2) <N11> 我大概是 er 一六年九啊八月份的时候到 ***xi*** 第一次到的新西兰<br>(3) <N11>国内其实我也 ***liu*** 有一些 er 我同班的呀 |
| **Capitalisation** | The third singular pronouns are marked with capital letters TA in situations where the gender of the person mentioned by the participant are not clear. | <N01> 算是第一顿饭都是 TA 请我们就是这样的 |
| | No capitalisation is used to mark backchannels. | <S00> eng eng eng |
| **Punctuation** | Do not use punctuation markers. | |
| **Pauses** | All pauses are measured and marked with the length inside angle brackets. | <pause=0.7> |
| **Overlapping speech** | Do not mark overlaps. | |

| Feature | Transcription guideline | Example |
|---|---|---|
| **Backchannels** | Backchannels *eng* are marked with *Pinyin*. | <S00> eng eng eng |
| **Minimal response tokens** | Use standard Chinese characters. | (1) <S00> 对啊<br>(2) <S00> erm 对对是<br>(3) <N21> 哦哦 |
| **Uncertain words** | Mark as uncertain with a guess if possible. | <uncertain=战狼> (Wolf Warrior) |
| **Unclear speech** | Mark as unclear with a guess if possible. | <unclear=fund> |
| **Acronyms and abbreviations** | Use capital letters without spaces when spelling out a word letter by letter, e.g., (1); where acronyms and abbreviations are pronounced as words only the first letters of them are capitalised and all letters are not separated by spaces, e.g., (2) App is pronounced as 'æp'. | (1) HSK<br>(2) App |
| **Repetition** | Use standard Chinese characters. | <L11> 啊太热太热了对太热了 |
| **False starts and repairs** | Use standard Chinese characters. | <L06>那边的人也非常地 er 他们的生活节奏也非常慢 |
| **Anonymisation** | Anonymise name of person and any reference that would allow an individual to be identified from the transcription. | <name>, <university>, <city> |
| **Numbers and dates** | All numbers and dates should be spelt out. | 二零一五 *er ling yi wu* (2015) |
| **L2 language features** | Do not attempt to transcribe different accents or non-standard pronunciation. Use standard forms of words. | |
| | If an incorrect pronunciation is produced, transcribe with its correct corresponding written form. | <L11>我们看什么兰**战狼二** (Wolf Warrior II) |
| | Do not correct L2 errors. | |
| | Use standard English to record English words. | <L03>**two thousand sixteen two thousand sixteen** 我参加这个汉语桥比赛 |

| Feature | Transcription guideline | Example |
|---|---|---|
| **Pronunciation** | The word 这个 can be pronounced as '*zhege'* or '*zheige'* in spoken Chinese, and both 'neige' and 'nage' are referred to 那个 with no difference in meanings. In the transcripts, 这个 is used to represent '*zhege'* and '*zheige'*; either '*neige'* or '*nage'* is transcribed as 那个. | <N15> 然后**那个**就包括北京北京也是一样<br><L13> 因为我觉得我觉得**这个这个**学学会一个外语并不是一个一朝一夕的事情对吧 |
| | All the uses of 儿 er are kept in the transcripts. It is a non-syllabic diminutive suffix in spoken Chinese which is widely used in the northern dialects and *Putonghua*. | <N01> 三月份**儿**那会**儿**可能自己就自己那段**儿**时间也懒嘛 |

## Appendix I  The Spoken L1 Corpus

**Table I-1**

*Metadata Information of the L1 Participants*

| Speaker ID | Gender | Age | Birthplace | Region | Highest qualification | Occupation | relationship with the researcher |
|---|---|---|---|---|---|---|---|
| N01 | female | 28 | Hebei | northern China | Master's | professional | friend |
| N02 | female | 28 | Hunan | southern China | Master's | student | acquaintance |
| N03 | male | 28 | Hebei | northern China | Bachelor's | professional | stranger |
| N04 | male | 28 | Shandong | northern China | Master's | student | friend |
| N05 | female | 32 | Hebei | northern China | Master's | professional | friend |
| N06 | female | 36 | Beijing | northern China | Doctorate | student | acquaintance |
| N07 | female | 35 | Beijing | northern China | Master's | professional | acquaintance |
| N08 | female | 36 | Shaanxi | northern China | Master's | professional | stranger |
| N09 | female | 28 | Shandong | northern China | Master's | professional | friend |
| N10 | female | 31 | Shandong | northern China | Master's | professional | friend |
| N11 | female | 25 | Beijing | northern China | Master's | student | acquaintance |
| N12 | female | 25 | Shandong | northern China | Master's | professional | acquaintance |
| N13 | male | 27 | Hebei | northern China | Bachelor's | professional | acquaintance |
| N14 | female | 28 | Henan | northern China | Master's | student | acquaintance |
| N15 | male | 35 | Beijing | northern China | Master's | professional | acquaintance |
| N16 | female | 28 | Jiangsu | southern China | Master's | professional | friend |
| N17 | male | 27 | Henan | northern China | Master's | professional | acquaintance |
| N18 | female | 27 | Sichuan | southern China | Master's | student | stranger |
| N19 | male | 35 | Sichuan | southern China | Master's | professional | acquaintance |
| N20 | male | 27 | Shandong | northern China | Master's | professional | acquaintance |

| Speaker ID | Gender | Age | Birthplace | Region | Highest qualification | Occupation | Inter-speaker relationship |
|---|---|---|---|---|---|---|---|
| N21 | male | 35 | Shandong | northern China | Master’s | professional | acquaintance |
| N22 | female | 24 | Hunan | southern China | Master’s | student | acquaintance |

**Table I-2**

*The Spoken L1 Corpus*

| Text | No. of words (participants' speech only) | No. of words (researcher's speech only) | Sample size (100%) |
|---|---|---|---|
| N01-C01 | 12,958 (91.56%) | 1,194 (8.43%) | 14,152 |
| N01-C02 | 7,273 (90.95%) | 724 (9.05%) | 7,997 |
| N02-C01 | 6,775 (66.77%) | 3,372 (33.23%) | 10,147 |
| N02-C02 | 7,280 (89.97%) | 812 (10.03%) | 8,092 |
| N02-C03 | 9,427 (97.58%) | 234 (2.42%) | 9,661 |
| N03-C01 | 4,572 (76.83%) | 1,379 (23.17%) | 5,951 |
| N04-C01 | 9,681 (91.62%) | 885 (8.38%) | 10,566 |
| N04-C02 | 10,508 (95.34%) | 514 (4.66%) | 11,022 |
| N05-C01 | 6,475 (95.98%) | 271 (4.02%) | 6,746 |
| N06-C01 | 5,896 (93.01%) | 443 (6.99%) | 6,339 |
| N07-C01 | 7,056 (76.21%) | 2,203 (23.79%) | 9,259 |
| N08-C01 | 5,035 (66.25%) | 2,565 (33.75%) | 7,600 |
| N09-C01 | 2,881 (65.02%) | 1,550 (34.98%) | 4,431 |
| N10-C01 | 5,215 (57.11%) | 3,916 (42.89%) | 9,131 |
| N11-C01 | 8,428 (94.89%) | 454 (5.11%) | 8,882 |
| N12-C01 | 4,572 (56.59%) | 3,507 (43.41%) | 8,079 |
| N13-C01 | 3,880 (59.10%) | 2,685 (40.90%) | 6,565 |
| N14-C01 | 7,706 (72.48%) | 2,926 (27.52%) | 10,632 |
| N15-C01 | 8,142 (74.39%) | 2,803 (25.61%) | 10,945 |
| N16-C01 | 4,599 (75.83%) | 1,466 (24.17%) | 6,065 |
| N17-C01 | 5,374 (49.94%) | 5,387 (50.06%) | 10,761 |
| N18-C01 | 5,851 (70.38%) | 2,462 (29.62%) | 8,313 |
| N19-C01 | 6,713 (77.80%) | 1,916 (22.20%) | 8,629 |
| N20-C01 | 6,226 (75.82%) | 1,986 (24.18%) | 8,212 |
| N21-C01 | 6,771 (51.83%) | 6,294 (48.17%) | 13,065 |
| N22-C01 | 4,715 (66.75%) | 2,349 (33.25%) | 7,064 |
| **Total (words)** | **174,009 (76.22%)** | **54,297 (23.78%)** | **228,306** |

**Table I-3**

*Sample Texts Produced by the Female L1 Participants*

| Text | No. of words (participants' speech only) | No. of words (researcher's speech only) | Sample size (100%) |
|---|---|---|---|
| N01-C01 | 12,958 (91.56%) | 1,194 (8.43%) | 14,152 |
| N01-C02 | 7,273 (90.95%) | 724 (9.05%) | 7,997 |
| N02-C01 | 6,775 (66.77%) | 3,372 (33.23%) | 10,147 |
| N02-C02 | 7,280 (89.97%) | 812 (10.03%) | 8,092 |
| N02-C03 | 9,427 (97.58%) | 234 (2.42%) | 9,661 |
| N05-C01 | 6,475 (95.98%) | 271 (4.02%) | 6,746 |
| N06-C01 | 5,896 (93.01%) | 443 (6.99%) | 6,339 |
| N07-C01 | 7,056 (76.21%) | 2,203 (23.79%) | 9,259 |
| N08-C01 | 5,035 (66.25%) | 2,565 (33.75%) | 7,600 |
| N09-C01 | 2,881 (65.02%) | 1,550 (34.98%) | 4,431 |
| N10-C01 | 5,215 (57.11%) | 3,916 (42.89%) | 9,131 |
| N11-C01 | 8,428 (94.89%) | 454 (5.11%) | 8,882 |
| N12-C01 | 4,572 (56.59%) | 3,507 (43.41%) | 8,079 |
| N14-C01 | 7,706 (72.48%) | 2,926 (27.52%) | 10,632 |
| N16-C01 | 4,599 (75.83%) | 1,466 (24.17%) | 6,065 |
| N18-C01 | 5,851 (70.38%) | 2,462 (29.62%) | 8,313 |
| N22-C01 | 4,715 (66.75%) | 2,349 (33.25%) | 7,064 |
| **Total (words)** | **112,142 (78.65%)** | **30,448 (21.35%)** | **142,590** |

**Table I-4**

*Sample Texts Produced by the Male L1 Participants*

| Text | No. of words (participants' speech only) | No. of words (researcher's speech only) | Sample size (100%) |
|---|---|---|---|
| N03-C01 | 4,572 (76.83%) | 1,379 (23.17%) | 5,951 |
| N04-C01 | 9,681 (91.62%) | 885 (8.38%) | 10,566 |
| N04-C02 | 10,508 (95.34%) | 514 (4.66%) | 11,022 |
| N13-C01 | 3,880 (59.10%) | 2,685 (40.90%) | 6,565 |
| N15-C01 | 8,142 (74.39%) | 2,803 (25.61%) | 10,945 |
| N17-C01 | 5,374 (49.94%) | 5,387 (50.06%) | 10,761 |
| N19-C01 | 6,713 (77.80%) | 1,916 (22.20%) | 8,629 |
| N20-C01 | 6,226 (75.82%) | 1,986 (24.18%) | 8,212 |
| N21-C01 | 6,771 (51.83%) | 6,294 (48.17%) | 13,065 |
| **Total (words)** | **61,867 (72.18%)** | **23,849 (27.82%)** | **85,716** |

## Appendix J The Spoken L2 Corpus

This appendix provides the detailed information on the L2 participants and the sampling texts in the L2 corpus. To protect the participants' privacy, specific cities of residence are not given in Table J-1. Table J-2 lists the number of words contained in each conversation, including the proportions of the researcher's contribution in this corpus. The distribution of the same-sex conversations in the L2 corpus is given in Table J-3, while the distribution of the mixed-sex conversations is shown in Table J-4.

**Table J-1**

*Metadata Information of the L2 Participants*

| Speaker ID | Gender | Age | Nationality | Place lived in China | Current country of residence | Occupation |
|---|---|---|---|---|---|---|
| L01 | male | 32 | New Zealand | southern city | New Zealand | professional |
| L02 | male | 24 | New Zealand | northern city | China | professional |
| L03 | female | 22 | New Zealand | northern city | New Zealand | professional |
| L04 | female | 24 | New Zealand | southern city | China | student |
| L05 | female | 35 | New Zealand | northern city | New Zealand | professional |
| L06 | male | 32 | New Zealand | northern city | China | professional |
| L07 | male | 40 | New Zealand | northern city | New Zealand | professional |
| L08 | male | 24 | Australia | northern city | China | student |
| L09 | male | 40 | New Zealand | northern city | New Zealand | professional |
| L10 | male | 30 | New Zealand | northern city | China | professional |
| L11 | male | 24 | New Zealand | northern city | New Zealand | professional |
| L12 | male | 40 | New Zealand | southern city | China | professional |
| L13 | male | 30 | New Zealand | northern city | China | student |
| L14 | male | 24 | Australia | northern city | China | student |

**Table J-2**

*The Spoken L2 Corpus*

| Text | No. of words (participants' speech only) | No. of words (researcher's speech only) | Sample size (100%) |
|---|---|---|---|
| L01-C01 | 7,117 (84.47%) | 1,308 (15.53%) | 8,425 |
| L01-C02 | 9,779 (75.94%) | 3,099 (24.06%) | 12,878 |
| L01-C03 | 8,867 (55.23%) | 7,188 (44.77%) | 16,055 |
| L01-C04 | 7,048 (76.24%) | 2,197 (23.76%) | 9,245 |
| L02-C01 | 2,718 (68.65%) | 1,241 (31.35%) | 3,959 |
| L02-C02 | 3,198 (63.82%) | 1,813 (36.18%) | 5,011 |
| L03-C01 | 2,055 (77.40%) | 600 (22.60%) | 2,655 |
| L03-C02 | 4,013 (77.40%) | 1,172 (22.60%) | 5,185 |
| L04-C01 | 2,776 (44.79%) | 3,422 (55.21%) | 6,198 |
| L04-C02 | 3,392 (59.32%) | 2,326 (40.68%) | 5,718 |
| L04-C03 | 4,439 (59.20%) | 3,059 (40.80%) | 7,498 |
| L05-C01 | 5,313 (59.34%) | 3,641 (40.66%) | 8,954 |
| L05-C02 | 6,728 (48.32%) | 7,197 (51.68%) | 13,925 |
| L05-C03 | 3,420 (60.31%) | 2,251 (39.69%) | 5,671 |
| L06-C01 | 4,958 (84.09%) | 938 (15.91%) | 5,896 |
| L06-C02 | 2,927 (89.84%) | 331 (10.16%) | 3,258 |
| L07-C01 | 3,431 (76.58%) | 1,049 (23.42%) | 4,480 |
| L08-C01 | 4,176 (65.78%) | 2,172 (34.22%) | 6,348 |
| L08-C02 | 2,036 (74.09%) | 712 (25.91%) | 2,748 |
| L08-C03 | 5,023 (59.80%) | 3,376 (40.20%) | 8,399 |
| L09-C01 | 2,553 (47.44%) | 2,829 (52.56%) | 5,382 |
| L09-C02 | 3,644 (55.02%) | 2,979 (44.98%) | 6,623 |
| L09-C03 | 3,571 (46.89%) | 4,044 (53.11%) | 7,615 |
| L10-C01 | 5,386 (55.56%) | 4,308 (44.44%) | 9,694 |
| L10-C02 | 6,019 (75.63%) | 1,940 (24.37%) | 7,959 |
| L11-C01 | 3,928 (82.47%) | 835 (17.53%) | 4,763 |
| L11-C02 | 5,553 (78.74%) | 1,499 (21.26%) | 7,052 |
| L11-C03 | 5,034 (79.41%) | 1,305 (20.59%) | 6,339 |
| L12-C01 | 2,435 (65.79%) | 1,266 (34.21%) | 3,701 |
| L12-C02 | 3,558 (79.56%) | 914 (20.44%) | 4,472 |
| L12-C03 | 1,727 (71.72%) | 681 (28.28%) | 2,408 |
| L13-C01 | 1,757 (71.89%) | 687 (28.11%) | 2,444 |
| L13-C02 | 2,104 (65.67%) | 1,100 (34.33%) | 3,204 |
| L14-C01 | 5,915 (89.22%) | 715 (10.78%) | 6,630 |
| **Total (words)** | **146,598 (66.40%)** | **74,194 (33.60%)** | **220,792** |

**Table J-3**

*Sample Texts Produced by the Male L2 Participants*

| Text | No. of words (participants' speech only) | No. of words (researcher's speech only) | Sample size (100%) |
|---|---|---|---|
| L01-C01 | 7,117 (84.47%) | 1,308 (15.53%) | 8,425 |
| L01-C02 | 9,779 (75.94%) | 3,099 (24.06%) | 12,878 |
| L01-C03 | 8,867 (55.23%) | 7,188 (44.77%) | 16,055 |
| L01-C04 | 7,048 (76.24%) | 2,197 (23.76%) | 9,245 |
| L02-C01 | 2,718 (68.65%) | 1,241 (31.35%) | 3,959 |
| L02-C02 | 3,198 (63.82%) | 1,813 (36.18%) | 5,011 |
| L06-C01 | 4,958 (84.09%) | 938 (15.91%) | 5,896 |
| L06-C02 | 2,927 (89.84%) | 331 (10.16%) | 3,258 |
| L07-C01 | 3,431 (76.58%) | 1,049 (23.42%) | 4,480 |
| L08-C01 | 4,176 (65.78%) | 2,172 (34.22%) | 6,348 |
| L08-C02 | 2,036 (74.09%) | 712 (25.91%) | 2,748 |
| L08-C03 | 5,023 (59.80%) | 3,376 (40.20%) | 8,399 |
| L09-C01 | 2,553 (47.44%) | 2,829 (52.56%) | 5,382 |
| L09-C02 | 3,644 (55.02%) | 2,979 (44.98%) | 6,623 |
| L09-C03 | 3,571 (46.89%) | 4,044 (53.11%) | 7,615 |
| L10-C01 | 5,386 (55.56%) | 4,308 (44.44%) | 9,694 |
| L10-C02 | 6,019 (75.63%) | 1,940 (24.37%) | 7,959 |
| L11-C01 | 3,928 (82.47%) | 835 (17.53%) | 4,763 |
| L11-C02 | 5,553 (78.74%) | 1,499 (21.26%) | 7,052 |
| L11-C03 | 5,034 (79.41%) | 1,305 (20.59%) | 6,339 |
| L12-C01 | 2,435 (65.79%) | 1,266 (34.21%) | 3,701 |
| L12-C02 | 3,558 (79.56%) | 914 (20.44%) | 4,472 |
| L12-C03 | 1,727 (71.72%) | 681 (28.28%) | 2,408 |
| L13-C01 | 1,757 (71.89%) | 687 (28.11%) | 2,444 |
| L13-C02 | 2,104 (65.67%) | 1,100 (34.33%) | 3,204 |
| L14-C01 | 5,915 (89.22%) | 715 (10.78%) | 6,630 |
| **Total (words)** | **114,462 (63.38%)** | **50,526 (30.62%)** | **164,988** |

**Table J-4**

*Sample Texts Produced by the Female L2 Participants*

| Text | No. of words (participants' speech only) | No. of words (researcher's speech only) | Sample size (100%) |
|---|---|---|---|
| L03-C01 | 2,055 (77.40%) | 600 (22.60%) | 2,655 |
| L03-C02 | 4,013 (77.40%) | 1,172 (22.60%) | 5,185 |
| L04-C01 | 2,776 (44.79%) | 3,422 (55.21%) | 6,198 |
| L04-C02 | 3,392 (59.32%) | 2,326 (40.68%) | 5,718 |
| L04-C03 | 4,439 (59.20%) | 3,059 (40.80%) | 7,498 |
| L05-C01 | 5,313 (59.34%) | 3,641 (40.66%) | 8,954 |
| L05-C02 | 6,728 (48.32%) | 7,197 (51.68%) | 13,925 |
| L05-C03 | 3,420 (60.31%) | 2,251 (39.69%) | 5,671 |
| **Total (words)** | **32,136 (57.59%)** | **23,668 (42.41%)** | **55,804** |

## Appendix K The Lexical Item 就是 *Jiushi* in the L1 Corpus

This appendix shows frequencies of the item 就是 *jiushi* in the L1 corpus.

**Table K-1**

*Distribution of the Item 就是 Jiushi in Each L1 Conversation*

| Text | Occurrences in each text | L1 participant's use | Researcher's use |
|---|---|---|---|
| N01-C01 | 366 | 356 | 10 |
| N01-C02 | 157 | 145 | 12 |
| N02-C01 | 250 | 188 | 62 |
| N02-C02 | 178 | 171 | 7 |
| N02-C03 | 204 | 203 | 1 |
| N03-C01 | 76 | 38 | 38 |
| N04-C01 | 481 | 477 | 4 |
| N04-C02 | 405 | 401 | 4 |
| N05-C01 | 65 | 63 | 2 |
| N06-C01 | 194 | 187 | 7 |
| N07-C01 | 184 | 149 | 35 |
| N08-C01 | 91 | 39 | 52 |
| N09-C01 | 52 | 26 | 26 |
| N10-C01 | 103 | 52 | 51 |
| N11-C01 | 163 | 163 | 0 |
| N12-C01 | 142 | 98 | 44 |
| N13-C01 | 121 | 66 | 55 |
| N14-C01 | 372 | 312 | 60 |
| N15-C01 | 190 | 146 | 44 |
| N16-C01 | 84 | 69 | 15 |
| N17-C01 | 127 | 37 | 90 |
| N18-C01 | 124 | 87 | 37 |
| N19-C01 | 81 | 55 | 26 |
| N20-C01 | 166 | 138 | 28 |
| N21-C01 | 158 | 63 | 95 |
| N22-C01 | 155 | 106 | 49 |
| **Total** | **4,691** | **3,837** | **854** |

*Note*. All the occurrences of 就是说 *jiushi shuo* in the L1 corpus were excluded from this frequency list.

**Table K-2**

*Relative Frequencies of the Item 就是 Jiushi Used in Each L1 Conversation*

| Text | L1 participant's use | Researcher's use |
|---|---|---|
| N01-C01 | 25.15546 | 0.706614 |
| N01-C02 | 18.38189 | 1.500563 |
| N02-C01 | 18.52764 | 6.11018 |
| N02-C02 | 21.13198 | 0.865052 |
| N02-C03 | 21.01232 | 0.103509 |
| N03-C01 | 6.385481 | 6.385481 |
| N04-C01 | 45.1448 | 0.378573 |
| N04-C02 | 36.38178 | 0.362911 |
| N05-C01 | 9.338867 | 0.296472 |
| N06-C01 | 29.49992 | 1.104275 |
| N07-C01 | 16.09245 | 3.780106 |
| N08-C01 | 5.131579 | 6.842105 |
| N09-C01 | 5.86775 | 5.86775 |
| N10-C01 | 5.694886 | 5.585369 |
| N11-C01 | 18.35172 | 0 |
| N12-C01 | 12.13021 | 5.446219 |
| N13-C01 | 10.05331 | 8.377761 |
| N14-C01 | 29.34537 | 5.643341 |
| N15-C01 | 13.33942 | 4.020101 |
| N16-C01 | 11.37675 | 2.473207 |
| N17-C01 | 3.438342 | 8.363535 |
| N18-C01 | 10.46554 | 4.45086 |
| N19-C01 | 6.373856 | 3.013095 |
| N20-C01 | 16.80468 | 3.409644 |
| N21-C01 | 4.822044 | 7.271336 |
| N22-C01 | 15.00566 | 6.93658 |

*Note*. Relative frequencies of 就是 *jiushi* used by L1 participants (RF) were calculated as follows:

$$RF = \frac{\text{absolute frequency of each L1 participant's use}}{\text{number of tokens in conversation}} \times basis\ for\ normalisation$$

Relative frequencies of 就是 *jiushi* used by the researcher (rf) in each text were calculated as follows:

$$\text{rf} = \frac{\text{absolute frequency of the researcher's use in each text}}{\text{number of tokens in conversation}} \times basis\ for\ normalisation$$

For example, the relative frequencies of the L1 use and the researcher's use of 就是 *jiushi* in the text N01-C01 were calculated relatively as follows:

$$RF = \frac{356}{14{,}152} \times 1{,}000 \;=\; 25.15546$$

$$\mathrm{rf} = \frac{10}{14{,}152} \times 1{,}000 \;=\; 0.706614$$

## Appendix L The Lexical Item 就是 *Jiushi* in the L2 Corpus

This appendix shows frequencies of the item 就是 *jiushi* in the L2 corpus.

**Table L-1**

*Distribution of the Item 就是 Jiushi Used in the L2 Corpus*

| Text | Occurrences in each text | L2 participant's use | Researcher's use |
|---|---|---|---|
| L01-C01 | 209 | 203 | 6 |
| L01-C02 | 345 | 306 | 39 |
| L01-C03 | 369 | 242 | 127 |
| L01-C04 | 210 | 181 | 29 |
| L02-C01 | 50 | 27 | 23 |
| L02-C02 | 84 | 55 | 29 |
| L03-C01 | 9 | 6 | 3 |
| L03-C02 | 11 | 8 | 3 |
| L04-C01 | 23 | 1 | 22 |
| L04-C02 | 38 | 11 | 27 |
| L04-C03 | 40 | 8 | 32 |
| L05-C01 | 231 | 193 | 38 |
| L05-C02 | 381 | 248 | 133 |
| L05-C03 | 160 | 125 | 35 |
| L06-C01 | 63 | 50 | 13 |
| L06-C02 | 67 | 59 | 8 |
| L07-C01 | 14 | 5 | 9 |
| L08-C01 | 82 | 44 | 38 |
| L08-C02 | 31 | 23 | 8 |
| L08-C03 | 126 | 62 | 64 |
| L09-C01 | 60 | 19 | 41 |
| L09-C02 | 71 | 12 | 59 |
| L09-C03 | 99 | 12 | 87 |
| L10-C01 | 102 | 29 | 73 |
| L10-C02 | 50 | 25 | 25 |
| L11-C01 | 89 | 77 | 12 |
| L11-C02 | 98 | 84 | 14 |
| L11-C03 | 93 | 74 | 19 |
| L12-C01 | 96 | 60 | 36 |
| L12-C02 | 157 | 142 | 15 |
| L12-C03 | 52 | 42 | 10 |
| L13-C01 | 14 | 12 | 2 |
| L13-C02 | 44 | 26 | 18 |
| L14-C01 | 20 | 10 | 10 |
| **Total** | **3,588** | **2,481** | **1,107** |

*Note.* All the occurrences of 就是说 *jiushi shuo* in the L2 corpus were excluded from this list.

**Table L-2**

*Relative Frequencies of the Item 就是 Jiushi Used in Each L2 Conversation*

| Text | L2 participant's use | Researcher's use |
|---|---|---|
| L01-C01 | 24.09496 | 0.712166 |
| L01-C02 | 23.76145 | 3.028421 |
| L01-C03 | 15.07319 | 7.910308 |
| L01-C04 | 19.57815 | 3.136831 |
| L02-C01 | 6.819904 | 5.809548 |
| L02-C02 | 10.97585 | 5.787268 |
| L03-C01 | 2.259887 | 1.129944 |
| L03-C02 | 1.542912 | 0.578592 |
| L04-C01 | 0.161342 | 3.549532 |
| L04-C02 | 1.92375 | 4.721931 |
| L04-C03 | 1.066951 | 4.267805 |
| L05-C01 | 21.55461 | 4.243913 |
| L05-C02 | 17.80969 | 9.551167 |
| L05-C03 | 22.04197 | 6.171751 |
| L06-C01 | 8.480326 | 2.204885 |
| L06-C02 | 18.10927 | 2.455494 |
| L07-C01 | 1.116071 | 2.008929 |
| L08-C01 | 6.931317 | 5.986137 |
| L08-C02 | 8.369723 | 2.911208 |
| L08-C03 | 7.381831 | 7.619955 |
| L09-C01 | 3.530286 | 7.617986 |
| L09-C02 | 1.811868 | 8.90835 |
| L09-C03 | 1.575837 | 11.42482 |
| L10-C01 | 2.991541 | 7.530431 |
| L10-C02 | 3.141098 | 3.141098 |
| L11-C01 | 16.16628 | 2.519421 |
| L11-C02 | 11.91151 | 1.985252 |
| L11-C03 | 11.67377 | 2.997318 |
| L12-C01 | 16.21183 | 9.727101 |
| L12-C02 | 31.75313 | 3.354204 |
| L12-C03 | 17.44186 | 4.152824 |
| L13-C01 | 4.909984 | 0.818331 |
| L13-C02 | 8.114856 | 5.617978 |
| L14-C01 | 1.508296 | 1.508296 |

*Note*. Relative frequencies of 就是 *jiushi* used by the L2 participants (RF) were calculated as follows:

$$RF = \frac{\text{absolute frequency of each L2 participant's use}}{\text{number of tokens in conversation}} \times basis\ for\ normalisation$$

Relative frequencies of 就是 *jiushi* used by the researcher (rf) in each text were calculated as follows:

$$rf = \frac{\text{absolute frequency of the researcher's use in each text}}{\text{number of tokens in conversation}} \times basis\ for\ normalisation$$

For example, the relative frequencies of the L2 use and the researcher's use of 就是 *jiushi* in the text L01-C01 were calculated relatively as follows:

$$RF = \frac{203}{8{,}425} \times 1{,}000 = 24.09496$$

$$\text{rf} = \frac{6}{8{,}425} \times 1{,}000 = 0.712166$$

## Appendix M User Manual of the Spoken Chinese Corpus

It should be noted that all the text files can be read online, but pdf and word documents cannot be opened on GitHub. Therefore, users must download the pdf or word documents if they want to know the contents that the documents contain.

1. **Download the Corpus Data**

   If you want to download the text files, please follow the steps given below:

   **Step 1:** Open the link https://github.com/blculyn;

   **Step 2:** Click on 'The-Spoken-Chinese-Corpus-of-L1-L1-Informal-Conversation', or 'The-Spoken-Chinese-Corpus-of-L1-L2-Informal-Conversation';

   **Step 3:** Find the green button 

in the top-right corner;

   **Step 4:** Click on 

and then go to 

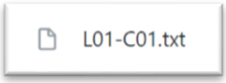

   **Step 5:** Click a folder destination on your computer for the zip file that you want to download from GitHub.

2. **Review the Data Online**

   If you want to review the data online, there are some steps that you can follow:

   **Step 1:** Open the link https://github.com/blculyn

   **Step 2:** Click on 'The-Spoken-Chinese-Corpus-of-L1-L1-Informal-Conversation', or 'The-Spoken-Chinese-Corpus-of-L1-L2-Informal-Conversation'

   **Step 3:** Click on the ID of the text that you want to review, for example, 

then you can see the content of the transcribed text.

## Appendix N The Discourse Marker 就是 *Jiushi* in the Spoken Chinese Corpus

This appendix shows the distributions of the participants' use of the marker 就是 *jiushi* in the conversations.

**Table N-1**

*Distribution of the Marker 就是 Jiushi Used by the L1 Participants in Each Conversation*

| Text | Raw frequency | Relative frequency |
|---|---|---|
| N01-C01 | 34 | 2.623862 |
| N01-C02 | 6 | 0.824969 |
| N02-C01 | 4 | 0.590406 |
| N02-C02 | 8 | 1.098901 |
| N02-C03 | 15 | 1.591174 |
| N03-C01 | 2 | 0.437445 |
| N04-C01 | 29 | 2.995558 |
| N04-C02 | 28 | 2.664636 |
| N05-C01 | 2 | 0.30888 |
| N06-C01 | 8 | 1.356852 |
| N07-C01 | 12 | 1.70068 |
| N08-C01 | 0 | 0 |
| N09-C01 | 2 | 0.694203 |
| N10-C01 | 6 | 1.150527 |
| N11-C01 | 5 | 0.593261 |
| N12-C01 | 3 | 0.656168 |
| N13-C01 | 5 | 1.28866 |
| N14-C01 | 25 | 3.244225 |
| N15-C01 | 5 | 0.6141 |
| N16-C01 | 7 | 1.52207 |
| N17-C01 | 3 | 0.558243 |
| N18-C01 | 6 | 1.025466 |
| N19-C01 | 2 | 0.297929 |
| N20-C01 | 19 | 3.051719 |
| N21-C01 | 1 | 0.147689 |
| N22-C01 | 8 | 1.696713 |
| **Total** | **245** | **1.073121** |

*Note*. The relative frequency (RF) of the reformulation marker 就是 *jiushi* in the L1 corpus was calculated as follows:

$$RF = \frac{\text{absolute frequency of each L1 participant's use}}{\text{number of tokens in L1 conversation}} \times basis\ for\ normalisation$$

Accordingly, the relative frequency of the L1 use of 就是 *jiushi* in conversation was calculated as follows:

$$RF\ (jiushi) = \frac{245}{228{,}306} \times 1{,}000 = 1.073121$$

The relative frequency of the L1 use of 就是 *jiushi* in the text N01-C01 was calculated as follows:

$$RF\ (jiushi) = \frac{34}{12{,}958} \times 1{,}000 = 2.623862$$

It means that, on average, there are about three instances of the marker 就是 *jiushi* for every 1,000 tokens in the speech of speaker N01.

**Table N-2**

*Computation of DP of 就是 Jiushi in the L1 Corpus*

| Text | Expected % | Observed % | Abs. difference |
|---|---|---|---|
| N01-C01 | 0.074467 | 0.138776 | 0.064309 |
| N01-C02 | 0.041797 | 0.02449 | 0.01731 |
| N02-C01 | 0.038935 | 0.016327 | 0.02261 |
| N02-C02 | 0.041837 | 0.032653 | 0.00918 |
| N02-C03 | 0.054175 | 0.061224 | 0.007049 |
| N03-C01 | 0.026275 | 0.008163 | 0.01811 |
| N04-C01 | 0.055635 | 0.118367 | 0.062732 |
| N04-C02 | 0.060388 | 0.114286 | 0.053898 |
| N05-C01 | 0.037211 | 0.008163 | 0.02905 |
| N06-C01 | 0.033883 | 0.032653 | 0.00123 |
| N07-C01 | 0.04055 | 0.04898 | 0.00843 |
| N08-C01 | 0.028935 | 0 | 0.02894 |
| N09-C01 | 0.016557 | 0.008163 | 0.00839 |
| N10-C01 | 0.02997 | 0.02449 | 0.00548 |
| N11-C01 | 0.048434 | 0.020408 | 0.02803 |
| N12-C01 | 0.026275 | 0.012245 | 0.01403 |
| N13-C01 | 0.022298 | 0.020408 | 0.00189 |
| N14-C01 | 0.044285 | 0.102041 | 0.057756 |
| N15-C01 | 0.046791 | 0.020408 | 0.02638 |
| N16-C01 | 0.02643 | 0.028571 | 0.002141 |
| N17-C01 | 0.030883 | 0.012245 | 0.01864 |
| N18-C01 | 0.033625 | 0.02449 | 0.00913 |
| N19-C01 | 0.038578 | 0.008163 | 0.03042 |
| N20-C01 | 0.03578 | 0.077551 | 0.041771 |
| N21-C01 | 0.038912 | 0.004082 | 0.03483 |
| N22-C01 | 0.027096 | 0.032653 | 0.005557 |

*Note.* According to the calculation steps given by Gries (2008), the first step which results in the second column from the left was calculated as follows:

$$expected\ proportion = \frac{\text{the size of each text}}{\text{the total number of tokens in L1 speech}}$$

For example, the expected proportion in the second column from the left in terms of the text labelled as N01-C01 was calculated as follows:

$$expected\ proportion = \frac{12{,}958}{174{,}009} = 0.074467$$

The second step results in the third column from the left, i.e., the observed proportion of 就是 *jiushi* as a reformulation marker under examination, was calculated as follows:

$$observed\ proportion$$

$$= \frac{\text{the absolute frequency of } jiushi \text{ as a reformulation marker in each text}}{\text{the total number of tokens of } jiushi \text{ as a reformulation marker in L1 speech}}$$

For example, in the text N01-C01, the observed proportion in the third column (on the basis of the information offered in Table M-1) was calculated as follows:

$$observed\ proportion = \frac{34}{245} = 0.138776$$

The third step results in the rightmost column, i.e., the absolute values of the difference between the observed and expected proportions, were calculated as follows:

$$absolute\ difference = \text{the observed proportion} - the\ expected\ proportion$$

For example, the absolute difference (between 0 and 1) in terms of 就是 *jiushi* in the text N01-C01 was calculated as follows:

$$absolute\ difference = 0.138776 - 0.074467 = 0.064309$$

Accordingly, the value of DP for the marker 就是 *jiushi* used by the L1 participants was calculated as follows:

$$DP = \frac{\text{the sum of the absolute differences}}{2} = \frac{0.064309 + \ldots + 0.005557}{2} = \frac{0.607293}{2}$$

$$= 0.303647$$

**Table N-3**

*Distribution of the Discourse Marker 就是 Jiushi Used in the L2 Corpus*

| Text | Raw frequency | Relative frequency |
|---|---|---|
| L01-C01 | 11 | 1.305638 |
| L01-C02 | 18 | 1.397733 |
| L01-C03 | 7 | 0.436001 |
| L01-C04 | 3 | 0.3245 |
| L02-C01 | 1 | 0.252589 |
| L02-C02 | 3 | 0.598683 |
| L03-C01 | 0 | 0 |
| L03-C02 | 0 | 0 |
| L04-C01 | 0 | 0 |
| L04-C02 | 0 | 0 |
| L04-C03 | 2 | 0.266738 |
| L05-C01 | 10 | 1.116819 |
| L05-C02 | 6 | 0.43088 |
| L05-C03 | 4 | 0.705343 |
| L06-C01 | 6 | 1.017639 |
| L06-C02 | 5 | 1.534684 |
| L07-C01 | 1 | 0.223214 |
| L08-C01 | 2 | 0.31506 |
| L08-C02 | 2 | 0.727802 |
| L08-C03 | 2 | 0.238124 |
| L09-C01 | 1 | 0.185805 |
| L09-C02 | 1 | 0.150989 |
| L09-C03 | 1 | 0.13132 |
| L10-C01 | 0 | 0 |
| L10-C02 | 2 | 0.251288 |
| L11-C01 | 5 | 1.049759 |
| L11-C02 | 5 | 0.709019 |
| L11-C03 | 2 | 0.315507 |
| L12-C01 | 3 | 0.810592 |
| L12-C02 | 4 | 0.894454 |
| L12-C03 | 1 | 0.415282 |
| L13-C01 | 0 | 0 |
| L13-C02 | 2 | 0.62422 |
| L14-C01 | 2 | 0.301659 |
| **Total** | **112** | **0.507265** |

*Note*. The relative frequency (RF) of the reformulation marker 就是 *jiushi* in the L2 corpus was calculated as follows:

$$RF = \frac{\text{absolute frequency of each L2 participant's use}}{\text{number of tokens in L2 conversation}} \times basis\ for\ normalisation$$

Accordingly, the relative frequency of the L2 use of 就是 *jiushi* in conversation was calculated as follows:

$$RF\ (jiushi) = \frac{112}{220{,}792} \times 1{,}000 = 0.507265$$

The relative frequency of the L2 use of 就是 *jiushi* in the text L01-C01 was calculated as follows:

$$RF\ (jiushi) = \frac{11}{8{,}425} \times 1{,}000 = 1.305638$$

It means that, on average, there is about one instance of the marker 就是 *jiushi* for every 1,000 tokens in the speech of speaker L01.

**Table N-4**

*Computation of DP of 就是 Jiushi in the L2 Corpus*

| Text | Expected % | Observed % | Abs. difference |
|---|---|---|---|
| L01-C01 | 0.048548 | 0.098214 | 0.049666 |
| L01-C02 | 0.066706 | 0.160714 | 0.094008 |
| L01-C03 | 0.060485 | 0.0625 | 0.002015 |
| L01-C04 | 0.048077 | 0.026786 | 0.02129 |
| L02-C01 | 0.01854 | 0.008929 | 0.00961 |
| L02-C02 | 0.021815 | 0.026786 | 0.004971 |
| L03-C01 | 0.014018 | 0 | 0.01402 |
| L03-C02 | 0.027374 | 0 | 0.02737 |
| L04-C01 | 0.018936 | 0 | 0.01894 |
| L04-C02 | 0.023138 | 0 | 0.02314 |
| L04-C03 | 0.03028 | 0.017857 | 0.01242 |
| L05-C01 | 0.036242 | 0.089286 | 0.053044 |
| L05-C02 | 0.045894 | 0.053571 | 0.007677 |
| L05-C03 | 0.023329 | 0.035714 | 0.012385 |
| L06-C01 | 0.03382 | 0.053571 | 0.019751 |
| L06-C02 | 0.019966 | 0.044643 | 0.024677 |
| L07-C01 | 0.023404 | 0.008929 | 0.01448 |
| L08-C01 | 0.028486 | 0.017857 | 0.01063 |
| L08-C02 | 0.013888 | 0.017857 | 0.003969 |
| L08-C03 | 0.034264 | 0.017857 | 0.01641 |
| L09-C01 | 0.017415 | 0.008929 | 0.00849 |
| L09-C02 | 0.024857 | 0.008929 | 0.01593 |
| L09-C03 | 0.024359 | 0.008929 | 0.01543 |
| L10-C01 | 0.03674 | 0 | 0.03674 |
| L10-C02 | 0.041058 | 0.017857 | 0.0232 |
| L11-C01 | 0.026794 | 0.044643 | 0.017849 |
| L11-C02 | 0.037879 | 0.044643 | 0.006764 |
| L11-C03 | 0.034339 | 0.017857 | 0.01648 |
| L12-C01 | 0.01661 | 0.026786 | 0.010176 |
| L12-C02 | 0.02427 | 0.035714 | 0.011444 |
| L12-C03 | 0.011781 | 0.008929 | 0.00285 |
| L13-C01 | 0.011985 | 0 | 0.01199 |
| L13-C02 | 0.014352 | 0.017857 | 0.003505 |
| L14-C01 | 0.040348 | 0.017857 | 0.02249 |

*Note.* According to the calculation steps given by Gries (2008), the first step which results in the second column from the left was calculated as follows:

$$expected\ proportion = \frac{\text{the size of each text}}{\text{the total number of tokens in L2 speech}}$$

For example, the expected proportion in the second column from the left in terms of the text labelled as L01-C01 was calculated as follows:

$$expected\ proportion = \frac{7{,}117}{146{,}598} = 0.048548$$

The second step results in the third column from the left, i.e., the observed proportion of 就是 *jiushi* as a reformulation marker under examination, was calculated as follows:

$$observed\ proportion = \frac{\text{the absolute frequency of } jiushi \text{ as a reformulation marker in each text}}{\text{the total number of tokens of } jiushi \text{ as a reformulation marker in L2 speech}}$$

For example, in the text L01-C01, the observed proportion in the third column (on the basis of the information offered in Table N-3) was calculated as follows:

$$observed\ proportion = \frac{11}{112} = 0.098214$$

The third step results in the rightmost column, i.e., the absolute values of the difference between the observed and expected proportions, were calculated as follows:

$$absolute\ difference = \text{the observed proportion} - the\ expected\ proportion$$

For example, the absolute difference (between 0 and 1) in terms of 就是 *jiushi* in the text L01-C01 was calculated as follows:

$$absolute\ difference = 0.048548 - 0.098214 = 0.049666$$

Accordingly, the value of DP for the marker 就是 *jiushi* used by the L2 participants was calculated as follows:

$$DP = \frac{\text{the sum of the absolute differences}}{2} = \frac{0.049666 + \ldots + 0.02249}{2} = \frac{0.643811}{2}$$

$$= 0.321906$$