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. Speaker labels are followed by one space.	(1) <S00> 去学习嘛 <N01> 去了江西 (2) <L09> 挺难的 <S00> 他们都说汉字很难
Pinyin	Use Pinyin to represent backchannels, such as *eng*, *en*, e.g., (1), and non-standard pronunciation, such as *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 (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 (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.	
Overlapping speech	Do not mark overlaps.	
Backchannels	Backchannels *eng* and *en* are marked with *Pinyin*.	<S00> eng eng eng
Minimal response tokens	Use standard Chinese characters.	(1) <S00> 对啊 (2) <S00> erm 对对是 (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 (2) App
Repetition	Use standard Chinese characters.	<L11> 啊太热太热了对太热了

Feature	Transcription guideline	Example
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
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 <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