



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

**DOTTORATO DI RICERCA IN
TRADUZIONE, INTERPRETAZIONE E INTERCULTURALITÀ**

Ciclo XXXVIII

Settore Concorsuale: 10/G1 - GLOTTOLOGIA E LINGUISTICA

Settore Scientifico Disciplinare: L-LIN/01 GLOTTOLOGIA E LINGUISTICA

**CONSTRUCTION AND APPLICATION OF A MULTIMODAL
INTERLANGUAGE CORPUS FOR TEACHING CHINESE AS A FOREIGN
LANGUAGE**

Presentata da: Siyuan Liu

Coordinatore Dottorato

Chiara Elefante

Supervisore

Maja Miličević Petrović

Co-supervisori

Adriano Ferraresi

Serena Zuccheri

Esame finale anno 2026

Abstract

This thesis focuses on the construction and application of a pedagogically oriented multimodal Chinese interlanguage corpus, MICICL (Multimodal Interlanguage Corpus for Italian Chinese Learners). The corpus targets the language production of Italian L1 learners of Chinese and systematically integrates handwritten and reading-aloud data. Through a unified data processing and annotation framework, learner outputs from different modalities are made directly comparable within a single analytical structure.

MICICL comprises language data from 39 learners collected across multiple tasks, including 155 handwritten samples and 153 corresponding reading-aloud samples. For each data unit, the original media files, transcriptions, error annotations, and minimally edited corrected versions. This design results in a structured, searchable, and reusable dataset. Given the formal and procedural differences between handwritten and reading-aloud data, separate transcription, annotation, and correction protocols were developed for each modality, with an emphasis on operational feasibility and pedagogical interpretability.

To illustrate the application of the corpus, the thesis presents a character-level analysis as an example of how MICICL can be used in interlanguage research and pedagogical analysis. The illustrative analyses include quantitative descriptions of character-level phenomena in handwriting, phonological performance in reading-aloud data, writing-reading correspondences, and comparisons across learners at different proficiency levels. These analyses are intended to demonstrate the corpus's capacity to support within-modality and cross-modality analysis and to provide data-based support for teaching-related research and practice.

Overall, through the construction of MICICL and its application examples, this thesis presents a workflow for building and using a multimodal Chinese interlanguage corpus, from data collection and processing to analytical implementation. It offers a practical reference for the development of similar corpus resources and their use in both research and pedagogical contexts.

Table of Contents

List of Tables.....	VII
List of Figures.....	XI
Introduction.....	1
Chapter 1 Chinese as a Foreign Language (CFL)	3
<i>1.1 Foreign language as interlanguage.....</i>	<i>3</i>
1.1.1 Definition and properties of interlanguage	3
1.1.2 Interlanguage and native language transfer	5
<i>1.2 Core features of the Chinese language</i>	<i>6</i>
1.2.1 Chinese orthography	7
1.2.2 Chinese phonology, Hanyu Pinyin and the orthography-phonology interface	10
1.2.3 The central elements of Chinese morphosyntax	13
<i>1.3 Learning challenges related to Chinese characters</i>	<i>15</i>
1.3.1 Chinese character as an analysis unit.....	15
1.3.2 Challenges in character writing	16
1.3.3 Challenges in character use and recognition.....	18
1.3.4 Challenges in character reading and pronunciation.....	20
<i>1.4 Italian and Chinese: Cross-linguistic contrast and specific learning difficulties</i>	<i>22</i>
1.4.1 Italian and Chinese in contrast.....	22
1.4.2 L1-specific character writing difficulties.....	23
1.4.3 L1-specific character use difficulties	25
1.4.4 Pinyin as a transitional strategy under alphabetic influence.....	28
1.4.5 Reading-related difficulties and phonological transfer.....	29
<i>1.5 Teaching Chinese as a foreign language in Italy.....</i>	<i>32</i>
Chapter 2 Chinese Interlanguage Corpora and Pedagogical Orientation	35
<i>2.1 Interlanguage corpora.....</i>	<i>35</i>
2.1.1 Definition of Interlanguage Corpora.....	35
2.1.2 Multimodality in interlanguage corpora	36
2.1.3 Transcription and annotation of interlanguage corpora	39
2.1.4 The pedagogical orientation in interlanguage corpora.....	40
<i>2.3 An overview of Chinese interlanguage corpora</i>	<i>43</i>
2.3.1 Chinese interlanguage corpora.....	43
2.3.2 Comparative summary of the reviewed corpora.....	49
2.3.3 An evaluative look on existing Chinese interlanguage corpora.....	54
<i>2.4 Data processing of Chinese interlanguage corpora</i>	<i>58</i>

2.4.1 Handwritten data processing in existing interlanguage corpora	58
2.4.2 Spoken data processing in existing interlanguage corpora	64
Chapter 3 Corpus Design and Data Collection	75
3.1 Chapter overview	75
3.2 Corpus design	75
3.2.1 L1 background	76
3.2.2 Data collection setting	77
3.2.3 Data collecting tasks	78
3.2.4 The modalities and media of the corpus	80
3.2.5 Longitudinal and quasi-longitudinal	81
3.3 Ethics and privacy considerations	83
3.3.1 Review documents and approval outcome	83
3.3.2 Compliance during data collection, processing, and application	84
3.3.3 Data encoding, storage, and retention period	86
3.4 Pilot study	87
3.4.1 Objectives of the pilot study	87
3.4.3 Pilot study data collection procedure	89
3.4.3 Pilot study data collection content	91
3.4.4 Findings, problems, and adjustment strategies	94
3.5 Main data collection	97
3.5.1 Students recruitment and data collection timeline	97
3.5.2 Student metadata collection results	99
3.5.3 Language data collection results	102
Chapter 4 Data Processing and Corpus Construction	109
4.1 Chapter overview	109
4.2 MICICL data processing framework and tools	110
4.2.1 MICICL data processing framework	110
4.2.2 MICICL data processing principles	114
4.2.3 ELAN as the transcription tool	117
4.2.4 Limitations	119
4.3 MICICL handwritten data processing rules and outcomes	120
4.3.1 Complexity of handwritten data	120
4.3.2 Transcription of handwritten data	126
4.3.2 Annotation of writing data	135
4.3.4 Character level annotation of writing data	137
4.3.5 Other levels of annotation of writing data	140
4.3.6 Correction of writing data	146
4.3.6 Writing data transcription and annotation overview	147

4.4 MICICL reading-aloud data processing rules and outcomes	149
4.4.1 Complexity of reading-aloud data	149
4.4.2 Accuracy and reliability of reading-aloud data transcription.....	151
4.4.3 Transcription of MICICL reading-aloud data	153
4.4.4 Annotation of reading-aloud data.....	157
4.4.5 Correction of reading-aloud data	162
Chapter 5. Character Level Handwritten Data Analysis	165
5.1 Introduction.....	165
5.2 Character writing: incorrectly written characters	166
5.2.1 Research questions.....	166
5.2.2 Distribution of incorrectly written characters among learners	168
5.2.3 Distribution of incorrectly written characters by character type	171
5.2.4 Incorrectly written characters by category.....	176
5.2.5 Categorical analysis of incorrectly written characters	182
5.3 Character use: Incorrectly used characters	195
5.3.1 Research questions.....	195
5.3.2 Distribution of incorrectly used characters among learners	197
5.3.3 Distribution of incorrectly used characters by character type	199
5.3.4 Category of incorrectly used characters.....	202
5.3.5 Incorrectly used characters by category.....	203
5.4 Character replacement: Pinyin use	210
5.4.1 Research questions.....	210
5.4.2 Distribution of Pinyin use among learners.....	212
5.4.3 Distribution of Pinyin use on different character type.....	214
5.4.4 Analysis by category	216
5.5 Foreign words	219
5.6 Summary and discussion of handwritten data analysis	220
Chapter 6 Character Level Reading-aloud data Analysis.....	225
6.1 Overview of this chapter	225
6.2 Tone in reading-aloud	226
6.2.1 Research questions.....	226
6.2.2 Distribution of incorrect tone types	228
6.2.3 Distribution of tone confusion patterns.....	231
6.3 Initials and finals reading.....	236
6.3.1 Research questions.....	236
6.3.2 Distribution of incorrect initials.....	237
6.3.3 Distribution of incorrect finals.....	241
6.3.4 Compound errors involving initials/finals and tones	246

6.4 <i>Incorrectly written/used characters reading-aloud: Cross modal analysis</i>	250
6.4.1 Research questions.....	250
6.4.2 Incorrectly written characters reading-aloud	251
6.4.3 Incorrectly used characters reading-aloud	254
6.5 <i>Summary and discussion of reading-aloud data analysis</i>	257
Chapter 7. Comparison Across Proficiency Levels	263
7.1 <i>Chapter overview</i>	263
7.2 <i>Writing data comparison</i>	264
7.2.1 Research questions.....	264
7.2.2 Incorrectly written character comparison	265
7.2.3 Incorrectly used character comparison	271
7.2.4 Pinyin use.....	276
7.2.5 Summary and discussion of writing data comparison	283
7.3 <i>Reading-aloud data comparison</i>	289
7.3.1 Research questions.....	289
7.3.2 Incorrect tone	290
7.3.3 Segmental errors (SC, YC and QC)	299
7.3.4 Summary and discussion of reading-aloud data comparison.....	303
7.4 <i>Cross modal comparison</i>	307
7.4.1 Incorrectly written characters reading-aloud	307
7.4.2 Incorrectly used character reading-aloud.....	312
7.4.3 Summary and discussion of cross-modal comparison	317
7.5 <i>Summary of cross-proficiency analysis</i>	320
Chapter 8. Conclusions and Future Work	323
8.1 <i>Synthesis of findings from MICICL</i>	323
8.1.1 MICICL as a pedagogically oriented interlanguage corpus.....	323
8.1.2 Character-level interlanguage patterns.....	326
8.2 <i>Pedagogical implications of MICICL findings</i>	328
8.2.1 Teaching implications for Chinese characters	328
8.2.2 Teaching priorities across proficiency levels	333
8.2.3 Using MICICL for diagnostic and formative assessment	336
8.3 <i>Future research directions and extended uses of MICICL</i>	339
8.3.1 Lexical and syntactic levels of interlanguage	339
8.3.2 Phonological development beyond citation forms.....	341
8.3.3 Cross-linguistic extensions and reuse of MICICL.....	343
8.4 <i>Conclusion</i>	346
8.4.1 Overall conclusion and main contributions	346

8.4.2 Limitations and Future Research Directions.....	348
References.....	350
Appendices.....	383
<i>Appendix A Metadata collection form used in MICICL main study</i>	<i>383</i>
<i>Appendix B Chinese proficiency level of each student in MICICL.....</i>	<i>385</i>
<i>Appendix C Number of samples provided by each student.....</i>	<i>386</i>
<i>Appendix D 39 students' incorrectly written character ratio</i>	<i>387</i>
<i>Appendix E Total incorrectly written target character types in MICICL and their frequency</i>	<i>389</i>
<i>Appendix F 39 Students' incorrectly used character ratio</i>	<i>392</i>
<i>Appendix G Total incorrectly used target character types in MICICL and their frequency</i>	<i>394</i>

List of Tables

Table 2.1	<i>Design approach in the 17 analysed corpora</i>
Table 2.2	<i>Data types in the 17 analysed corpora</i>
Table 2.3	<i>Data processing in the 17 analysed corpora</i>
Table 3.1	<i>Task topics of each session</i>
Table 3.2	<i>Number of students and collected samples of each session</i>
Table 3.3	<i>Number of language samples across different proficiency levels</i>
Table 3.4	<i>Distribution of learners' participation frequency and continuity in data collection</i>
Table 4.1	<i>Character level (including punctuation marks) annotation labels and definitions</i>
Table 4.2	<i>Word level annotation labels and definitions</i>
Table 4.3	<i>Phrase level annotation labels and definitions</i>
Table 4.4	<i>Sentence level annotation labels and definitions</i>
Table 4.5	<i>Writing data transcription overview</i>
Table 4.6	<i>Writing data annotation label numbers</i>
Table 4.7	<i>Syllable level annotation labels and definitions</i>
Table 4.8	<i>Segment level annotation labels and definitions</i>
Table 4.9	<i>Phrase level annotation labels and definitions</i>
Table 4.10	<i>Reading-aloud data transcription overview</i>
Table 4.11	<i>Reading-aloud data annotation overview</i>
Table 5.1	<i>Top 10 target characters most frequently written incorrectly by students</i>
Table 5.2	<i>The 10 character types with the highest frequency of incorrectly written character</i>
Table 5.3	<i>Top 5 SCs target character types</i>
Table 5.4	<i>Top 5 CCs target character types</i>
Table 5.5	<i>Top 5 STCs and MCs target character types</i>
Table 5.6	<i>Statistics of four incorrectly written characters type</i>
Table 5.7	<i>Top 9 Target characters most frequently incorrectly used by students</i>

Table 5.8	<i>The 10 character types with the highest frequency of incorrectly used characters</i>
Table 5.9	<i>Classification results of incorrectly used characters in MICICL</i>
Table 5.10	<i>Statistics of four incorrectly used characters category</i>
Table 5.11	<i>Top 6 PSCs character types</i>
Table 5.12	<i>Top 5 GSCs character types</i>
Table 5.13	<i>Top 5 PGCs character types</i>
Table 5.14	<i>Top 10 target character type of Pinyin (including student No. 13 and No. 17)</i>
Table 5.15	<i>Top 10 Pinyin corresponding Chinese character types</i>
Table 5.16	<i>The 10 target character types with the highest frequency of Pinyin</i>
Table 5.17	<i>Distribution of incorrectly used character types in PZ+B</i>
Table 5.18	<i>Distribution of incorrectly used character types in PZ+C</i>
Table 5.19	<i>Distribution of different categories of PC</i>
Table 5.20	<i>The target tones for tone errors in PC</i>
Table 5.21.	<i>Overall distribution of incorrectly written characters, incorrectly used characters, and Pinyin use in MICICL</i>
Table 6.1	<i>Total tone error type frequency</i>
Table 6.2	<i>Distribution of tone substitutions when the target tone is D0</i>
Table 6.3	<i>Distribution of tone substitutions when the target tone is D1</i>
Table 6.4	<i>Distribution of tone substitutions when the target tone is D2</i>
Table 6.5	<i>Distribution of tone substitutions when the target tone is D3</i>
Table 6.6	<i>Distribution of tone substitutions when the target tone is D4</i>
Table 6.7	<i>Overall distribution of correct and incorrect productions for five tones</i>
Table 6.8	<i>Top five students with the highest number of incorrect initial</i>
Table 6.9	<i>Top 6 target initials with the widest error distribution</i>
Table 6.10	<i>Substitution patterns of top 6 target initials</i>
Table 6.11	<i>Top five students with the highest number of final errors</i>
Table 6.12	<i>Top 6 target finals with the widest error distribution</i>
Table 6.13	<i>Substitution patterns of top 6 target finals</i>

Table 6.14	<i>Top five students with the highest number of multiple errors</i>
Table 6.15	<i>Top 6 target characters with the widest error distribution</i>
Table 6.16	<i>Substitution patterns of top 4 target characters</i>
Table 6.17	<i>Pronunciation accuracy of incorrectly written characters across error types</i>
Table 6.18	<i>Proportion of tone, initial, final, and multiple errors in the pronunciation of incorrectly written characters</i>
Table 6.19	<i>Pronunciation accuracy of incorrectly used characters (GSCs and OCs) based on target characters</i>
Table 6.20	<i>Pronunciation accuracy of incorrectly used characters (GSCs and OCs) when tone errors are excluded</i>
Table 6.21	<i>Pronunciation accuracy of incorrectly written and incorrectly used characters across error types</i>
Table 7.1	<i>Proportion of learners with incorrectly written characters across proficiency levels</i>
Table 7.2	<i>Incorrectly written character ratios across proficiency level</i>
Table 7.3	<i>Number of four categories in different proficiency levels</i>
Table 7.4	<i>Proportion of student with incorrectly used characters across proficiency levels</i>
Table 7.5	<i>Number of incorrectly used characters across proficiency level</i>
Table 7.6	<i>Categorical statistics of incorrectly used characters in different proficiency level</i>
Table 7.7	<i>Proportion of student with Pinyin use across proficiency levels</i>
Table 7.8	<i>Proportion of student with Pinyin use across proficiency levels</i>
Table 7.9	<i>Number of Pinyin use across proficiency level (excluding No.13 and No. 17)</i>
Table 7.10	<i>Proportion of student with three errors across proficiency levels</i>
Table 7.11	<i>Proportion of student with five tone errors across proficiency levels</i>
Table 7.12	<i>Number of tone errors across proficiency level</i>
Table 7.13	<i>Total frequency of each tone across different proficiency levels</i>
Table 7.14	<i>Frequency of tone substitutions across proficiency levels</i>
Table 7.15	<i>Proportion of student with segmental errors across proficiency levels</i>
Table 7.16	<i>Segmental error ratios across proficiency level</i>

Table 7.17 *Errors ratios in reading-aloud data across proficiency level*

Table 7.18 *Number of students who read incorrectly written characters correctly*

Table 7.19 *Number of correctly read incorrectly written characters across proficiency levels*

Table 7.20 *Number of students who read incorrectly used characters correctly*

List of Figures

- Figure 1.1** *Strokes and stroke order of the Chinese character “每 (měi)”*
- Figure 1.2** *Strokes and stroke order of the Chinese character “海 (hǎi)”*
- Figure 1.3** *Strokes and stroke order of the Chinese character “莓 (méi)”*
- Figure 1.4** *Five-point tone marking system*
- Figure 2.1** *Data Format in NUCISC (Screenshot)*
- Figure 3.1** *Metadata collection form used in pilot study of MICICL*
- Figure 3.2** *Distribution of students by number of participated data collection sessions*
- Figure 3.3** *Biographical information collection form*
- Figure 3.4** *Language proficiency information collection form*
- Figure 3.5** *Distribution of students’ metadata score*
- Figure 3.6** *An example of a task description for students*
- Figure 3.7** *Structure of data codes used in MICICL*
- Figure 4.1** *Screenshot of the ELAN Interface during the MICICL data processing*
- Figure 4.2** *An Example of Incorrectly written Character*
- Figure 4.3** *An Example of Incorrectly used Character*
- Figure 4.4** *Example 1 of code mixing*
- Figure 4.5** *Example 2 of code mixing*
- Figure 4.6** *Example 3 of code mixing*
- Figure 4.7** *Example 4 of code mixing*
- Figure 4.8** *Example 1 of correction*
- Figure 4.9** *Example 2 of correction*
- Figure 4.10** *An example of incorrect punctuation*
- Figure 4.11** *An example of correctly written character 饭 (fàn) and 馆 (guǎn)*
- Figure 4.12** *An example of incorrectly used character*
- Figure 4.13** *An example of Pinyin written in capital letters*
- Figure 4.14** *An example of coexistence of characters and Latin letters*
- Figure 4.15** *An example of multi-format writing in parentheses*

- Figure 4.16** *An example of incomplete Pinyin*
- Figure 4.17** *Example 1 of Pinyin/foreign word with character blank*
- Figure 4.18** *Example 2 of Pinyin/foreign word with character blank*
- Figure 4.19** *An example of using Arabic numerals*
- Figure 4.20** *An example of handwriting correction (text addition)*
- Figure 4.21** *An example of handwriting correction (text addition and scribbling)*
- Figure 4.22** *An example of incorrect punctuation*
- Figure 5.1** *Number of students by incorrectly written character ratio*
- Figure 5.2** *Example 1 of SCs*
- Figure 5.3** *Example 2 of SCs*
- Figure 5.4** *Example 3 of SCs*
- Figure 5.5** *Example 1 of CCs*
- Figure 5.6** *Example 2 of CCs*
- Figure 5.7** *Example 3 of CCs*
- Figure 5.8** *Example 1 of STCS*
- Figure 5.9** *Example 2 of STCs*
- Figure 5.10** *Example 3 of STCs*
- Figure 5.11** *Example of MCs*
- Figure 5.12** *Classification results of incorrectly written characters in MICICL*
- Figure 5.13** *Example 1 of character 喜 (xǐ) in SCs*
- Figure 5.14** *Example 2 of character 喜 (xǐ) in SCs*
- Figure 5.15** *Example 3 of character 喜 (xǐ) in SCs*
- Figure 5.16** *Example of character 京 (jīng) in SCs*
- Figure 5.17** *Example 1 of character 看 (kàn) in SCs*
- Figure 5.18** *Example 2 of character 看 (kàn) in SCs*
- Figure 5.19** *Example 1 of character 书 (shū) in SCs*
- Figure 5.20** *Example 2 of character 书 (shū) in SCs*
- Figure 5.21** *Example 1 of character 早 (zǎo) in SCs*
- Figure 5.22** *Example 2 of character 早 (zǎo) in SCs*

- Figure 5.23** *Example 3 of character 早 (zǎo) in SCs*
- Figure 5.24** *Example 1 of character 候 (hòu) in CCs*
- Figure 5.25** *Example 2 of character 候 (hòu) in CCs*
- Figure 5.26** *Example 1 of character 春 (chūn) in CCs*
- Figure 5.27** *Example 2 of character 春 (chūn) in CCs*
- Figure 5.28** *Example 1 of character 喜 (xǐ) in CCs*
- Figure 5.29** *Example 2 of character 喜 (xǐ) in CCs*
- Figure 5.30** *Example 1 of character 夏 (xià) in CCs*
- Figure 5.31** *Example 2 of character 夏 (xià) in CCs*
- Figure 5.32** *Example 1 of character 幫 (bāng) in CCs*
- Figure 5.33** *Example 2 of character 幫 (bāng) in CCs*
- Figure 5.34** *Example 1 of character 或 (huò) in STCs*
- Figure 5.35** *Example 2 of character 或 (huò) in STCs*
- Figure 5.36** *Distribution of students by incorrectly used character ratio*
- Figure 5.37** *Character type number of four category*
- Figure 5.38** *Pinyin use ratio by learner (without students No.13 and No.17)*
- Figure 6.1** *Distribution of students' tone error ratios*
- Figure 6.2** *Distribution of tone error types among learners*
- Figure 6.3** *Distribution of tone substitution patterns*
- Figure 6.4** *Distribution of correct and incorrect productions across five tones*
- Figure 6.5** *Distribution of incorrect initial ratios among learners*
- Figure 6.6** *Distribution of incorrect final ratios among learners*
- Figure 6.7** *Distribution of multiple error ratios among learners*
- Figure 6.8** *Correct pronunciation ratio of incorrectly written characters by error type*
- Figure 6.9** *Boxplot of the distribution of segmental and tonal errors (SC, YC, QC, DC)*
- Figure 6.10** *Stacked bar chart of tone error distribution across five tone categories*
- Figure 6.11** *Comparison of correct and incorrect pronunciations across error types of incorrectly written and incorrectly used (part) characters*
- Figure 7.1** *Trend of incorrectly written character ratios across proficiency levels*

- Figure 7.2** *Boxplot of incorrectly written character ratios across proficiency levels*
- Figure 7.3** *Categorical ratio of incorrectly written characters in different proficiency level*
- Figure 7.4** *Incorrectly used character ratio across three proficiency level*
- Figure 7.5** *Boxplot of incorrectly used character ratios across proficiency levels*
- Figure 7.6** *Categorical ratio of incorrectly used characters in different proficiency level*
- Figure 7.7** *Pinyin use ratio across three proficiency level*
- Figure 7.8** *Boxplot of Pinyin use ratios across proficiency levels*
- Figure 7.9** *Pinyin use ratio across three proficiency level (excluding Student. No.13 and No.17)*
- Figure 7.10** *Boxplot of Pinyin use ratios across proficiency levels (excluding Student No.13 and No.17)*
- Figure 7.11** *Proportion of student with three errors across proficiency levels*
- Figure 7.12** *Comparative ratios of three errors across three proficiency levels*
- Figure 7.13** *Boxplot of three errors ratios across proficiency levels*
- Figure 7.14** *Tone errors ratios across proficiency level*
- Figure 7.15** *Boxplot of tone error ratios across proficiency levels*
- Figure 7.16** *Accuracy ratio of five tones across different proficiency levels*
- Figure 7.17** *Error ratio of five tones at each proficiency level*
- Figure 7.18** *Comparative ratios of three errors across proficiency levels*
- Figure 7.19** *Boxplot of three segmental errors ratios across proficiency levels*
- Figure 7.20** *Errors ratios in reading-aloud data across proficiency levels*
- Figure 7.21** *Ratio of incorrectly written characters correctly read across proficiency levels*
- Figure 7.22** *Boxplot of incorrectly written characters correctly read across proficiency levels*
- Figure 7.23** *Ratio of incorrectly written characters correctly read across proficiency levels*
- Figure 7.24** *Boxplot of incorrectly used character correctly read across proficiency levels*
- Figure 7.25** *Ratio of IWC and IUC correctly read across proficiency levels*

Figure 7.26 *Ratio of IWC and IUC reading accuracy across proficiency levels*

Introduction

In recent years, the construction of Chinese interlanguage corpora has expanded steadily, providing a large amount of learner data for research on second language acquisition of Chinese. However, despite this rapid development at the research level, the use of such corpora in actual Chinese language teaching remains relatively limited. Many existing Chinese interlanguage corpora were designed primarily to meet research-oriented needs, and their data structures, annotation schemes, and analytical orientations often fail to directly address concrete pedagogical questions. As a result, a noticeable gap persists between corpus-based research findings and classroom practice. How pedagogical needs can be systematically incorporated at the corpus construction stage therefore remains an issue that warrants further investigation in Chinese interlanguage corpus research.

Against this background, the present study is motivated by the question of how to construct an interlanguage corpus that can better serve the teaching of Chinese. Specifically, the study adopts a pedagogical perspective to explore how key information from learners' authentic language production can be retained during corpus design, data collection, and data processing, and how such information can subsequently be transformed into observations and interpretations that are directly meaningful for teaching.

To this end, the study first undertakes the construction of a pedagogically oriented Chinese interlanguage corpus, MICICL. During data collection, learners' language production is gathered across multiple modalities, while the data processing stage places particular emphasis on the interpretability and operational feasibility of transcription, annotation, and correction. This approach aims to ensure that the corpus is not only suitable for research analysis but also accessible and usable for pedagogical purposes. The construction of the corpus itself therefore constitutes a central component of the present study.

Following the completion of the corpus, the study further presents a concrete application example focusing on Chinese characters to illustrate how the corpus can be used in interlanguage research and Chinese language teaching. This example is not intended to

systematically uncover acquisition patterns of a specific linguistic phenomenon; rather, it serves to demonstrate how, within a pedagogically oriented corpus framework, researchers and teachers can analyze learner data and extract information that is directly relevant to teaching.

By adopting a research pathway that combines corpus construction with the presentation of an application example, this thesis examines the design principles and practical value of pedagogically oriented Chinese interlanguage corpora. The study aims to propose a feasible approach to corpus construction and use that enables Chinese interlanguage corpora to more effectively support language teaching, rather than to advance general theoretical claims about language acquisition.

The thesis is structured as follows. Chapter 1 introduces the theoretical background of Chinese as a foreign language teaching and interlanguage research, and discusses core linguistic features of Chinese and their main pedagogical challenges, with particular attention to learners with an Italian language background. Chapter 2 reviews the development of learner corpus research, focusing on pedagogical orientations and multimodal trends, and provides a systematic overview and evaluation of existing Chinese interlanguage corpora. Chapters 3 and 4 describe the overall design of MICICL, including data collection procedures, ethical considerations, and the processing, transcription, and annotation of handwritten and reading-aloud data. Chapter 5 focuses on the analysis of handwritten data, examining character-level phenomena such as incorrectly written characters, character misuse, and pinyin use. Chapter 6 analyzes character-level patterns in the reading-aloud data and, in combination with the handwriting data, examines Chinese character learning from a cross-modal perspective. Building on these analyses, Chapter 7 presents a systematic comparison across learners at different proficiency levels to explore developmental patterns among different learner groups. Finally, Chapter 8 summarizes the main findings of the study, discusses the pedagogical implications of using MICICL for Chinese language teaching, and outlines future directions and potential extensions for research and pedagogical applications of the corpus.

Chapter 1 Chinese as a Foreign Language (CFL)

1.1 Foreign language as interlanguage

1.1.1 Definition and properties of interlanguage

Foreign language learning is fundamentally a creative process, during which learners unconsciously construct a linguistic system of their own (Corder, 1975). This system emerges from their attempts to approximate the norms of the target language (TL) (Selinker, 1972; Richards et al., 1996) and constitutes the learner’s evolving “hypothesis” about how the TL works. In this sense, the learner is the only “native speaker” of this system (Corder, 1975; Brown, 1994; Fauziati, 2011). Although this learner-generated system differs from both the learners’ first language (L1) and the TL—and contains what count as errors from the perspective of the TL (Selinker, 1969; Al-Khresheh, 2015)—it nevertheless functions as a usable communicative code, much like any natural language (Corder, 1975; Tarone, 1994).

Various terms have been proposed to describe this phenomenon. Corder (1967) introduced “transitional competence” to emphasize its status as a product of language transfer. Nemser (1971) used the term “approximative system” to highlight its developmental nature. Selinker (1972) coined the term “interlanguage”, foregrounding its position between the learner’s L1 and TL. Among these, the term “interlanguage” has gained the widest acceptance (Mahmood & Murad, 2018; Aljumah, 2020), and thus in this study I use it to refer to the learner’s language system.

The interlanguage system is inherently goal-directed (Corder, 1975): learners progress from their L1 toward the TL through successive stages of interlanguage (Selinker, 1972). Its significance lies in its dynamic and provisional nature; interlanguage is not merely a bridge between L1 and TL but a representation of the learning process itself (Barone, 2024). It is widely regarded as an autonomous linguistic system in its own right. Examining learners’ interlanguage allows researchers to uncover the strategies and processes they employ in acquiring a new language (Osman et al., 2022).

Interlanguage has three main observable characteristics: systematicity, permeability, and fossilization (Adjemian, 1976; Saville-Troike, 2006; Fauziati, 2011). First, interlanguage is a coherent linguistic system constructed by learners and governed by its own underlying principles. Although these principles may differ from those of the TL, and learner productions may superficially deviate from target norms, the system still exhibits identifiable regularities rather than constituting a set of arbitrary or unrelated forms (Al-Khresheh, 2015). Even when interlanguage varies across contexts or tasks, its rule-governed nature remains observable, as patterns of use can still be traced and interpreted (Kasper & Dahl, 1991; Kasper, 2001). As Yip (1995, p7) claimed, “with respect to the constructions under investigation, we claim that there is considerable consistency across speakers”. Interlanguage output can be considered systematic when learners show consistent preferences for particular forms at a given point in their developmental trajectory (Tarone et al., 1976; Tarone, 1994). Like any other natural language, it has an internally coherent structure and is therefore amenable to systematic linguistic analysis.

Second, permeability refers to the openness of the interlanguage system to the influence of both L1 and TL rules or forms, making it susceptible to infiltration from multiple linguistic sources (Yip, 1995; Fauziati, 2017). This characteristic is reflected in learners’ reliance on L1 structures to compensate for gaps in their TL knowledge, as well as in instances of overgeneralization when applying emerging TL rules (Afiana et al., 2018; Barone, 2024). Within this framework, learner errors are often classified into interlingual and intralingual interference (Corder, 1974; Scovel, 2001; Amara, 2015). Interlingual errors arise from L1 transfer or other external influences, while intralingual errors result from TL features, incomplete mastery of TL rules, or internal processing mechanisms (Puspita, 2019). The permeability of interlanguage therefore highlights that learners’ developing linguistic systems are shaped not only by L1 transfer but also by continuous restructuring driven by ongoing exposure to and interaction with the TL.

Last, fossilization refers to a range of learner behaviors associated with persistent failure to fully acquire certain linguistic features, including backsliding, stabilized errors, and errors

that continue to occur even in advanced learners (Wei, 2008). Even after many years of exposure to the TL, such incorrect linguistic forms may remain in a learner's production (Towell & Hawkins, 1994; Han, 2005), ultimately becoming a permanent part of the way an individual speaks or writes the language (Richards & Schmidt, 2003). Selinker (1972) provides a classic definition of fossilizable linguistic phenomena:

fossilisable linguistic phenomena are linguistic items, rules, and subsystems which speakers of a particular L1 will tend to keep in their interlanguage relative to a particular TL, no matter what the age of the learner or amount of explanation or instruction he/she receives in the TL (p. 215).

All language items have the tendency to stabilization, however, learners can counteract this process through increased attention, enhanced awareness, and the development of self-monitoring skills, which may help destabilize erroneous elements in their interlanguage system (Zhang & Xie, 2014).

1.1.2 Interlanguage and native language transfer

Native language transfer is one of the core processes underlying the formation of interlanguage (Selinker, 1972) and an important factor contributing to fossilization (Al-Khresheh, 2015). It refers to the transfer of formal properties of the learner's L1—an overt manifestation of L1 forms (Gass, 1988)—including linguistic and cultural knowledge, into the processing and organization of TL input (Guo, 2022). Odlin's (1989) widely cited definition captures this broadly:

transfer is the influence resulting from similarities and differences between the TL and any other language that has been previously (and perhaps imperfectly) acquired (p. 27).

Language transfer is generally divided into positive and negative (Selinker, 1983). Positive transfer occurs when L1 knowledge facilitates the acquisition of the TL, whereas negative transfer arises when L1 features interfere with the processing and production of the TL, thereby impeding learning (Selinker, 1983; Karim & Nassaji, 2013; Bardovi-Harlig & Sprouse, 2018;

Guo, 2022). Transfer from the L1 is considered one of the most prominent mechanisms shaping interlanguage development (Osman et al., 2022).

Language transfer is also influenced by developmental factors (Yi, 2012). Research shows that learners are more susceptible to L1 influence at beginning stages, and such influence tends to diminish as proficiency increases (Wang & Wen, 2002; Beare & Bourdages, 2007; Javadi-Safa, 2018). This pattern aligns with the dynamic nature of interlanguage (Corder, 1978), which is a fluid system that changes with continued exposure to the TL.

Among the mechanisms of transfer, language distance is another widely discussed variable. Scholars have long noted that linguistic differences between the L1 and the TL can create difficulties in second language acquisition (Yi, 2012). Learners do not enter the TL classroom with an “empty mind”, but rather bring extensive experience of how language is used for communication (Lado, 1957; Wolfersberger, 2003; Skoog, 2006), and they inevitably transfer this experience when using the TL. Several studies have shown that when L1 and TL are typologically related or closely aligned, language learning tends to be easier (Schachter, 1974; Chuang et al., 2012; Pasquarella et al., 2015). Therefore, L1-TL comparison constitutes an important prerequisite for understanding language transfer. Against this backdrop, a closer examination of the core structural features of the Chinese language is necessary, as these features constitute the typological basis on which cross-linguistic distance from learners’ L1 is manifested and experienced in learning.

1.2 Core features of the Chinese language

Strictly speaking, *Chinese* refers to several major Sinitic varieties (Gil, 2011). However, in the field of Chinese as a Foreign Language (CFL), the term *Chinese* generally denotes Mandarin, or Standard Chinese (Chan et al., 2022). Accordingly, *Chinese* in the present study also refers specifically to Mandarin. There are multiple characteristics of Chinese that tend to make interlanguage development complex in combination with typologically distant first languages, such as the Indo-European ones. A major hurdle from the early stages onwards is given by the non-alphabetic writing system and phonological features such as tone.

1.2.1 Chinese orthography

The notion of Chinese character (字, *zì*) can refer to the spoken syllable, to its written graphical form, or to both simultaneously. Chinese characters originate from pictorial representations and constitute a logographic writing system. Their earliest forms—such as oracle bone inscriptions and bronze inscriptions—expressed meaning through iconic representation (Lu, 2014). Unlike alphabetic scripts, whose basic units are phonemes or syllables, the fundamental unit of the Chinese writing system is the character. Language writing systems can be classified into phonemic, syllabic, and morphographic types, and Chinese is the only full representative of a morphographic script, in which form, sound, and meaning are integrated into a single unit (Su, 2017).

Chinese is organized around character as its fundamental structural unit: each character consists of a syllable paired with its written form, and functions as the basic locus of sound–meaning association (Xu, 1998; Wang, 2005a; Zhang & Lu, 2010; Wang, 2013), a feature that underscores the structural independence of the Chinese character within the language system. As a writing system, Chinese characters are a spatially organized square-shaped script with two-dimensional structural properties (Zhang, 1990). In terms of compositional hierarchy, characters can be divided into three levels: strokes, components, and whole characters (Fei, 1996; Cui, 1997).

Strokes are the most basic writing units of Chinese characters, and they do not independently convey the semantic or structural meaning of a complete character. According to the Standard Forms of Commonly Used Chinese Characters in Print (Character Reform Press, 1986), the fundamental strokes include five types: horizontal (一), vertical (丨), left-falling (丿), dot (丶), and right-falling (㇏). Based on these categories, more than twenty derived stroke forms—such as 乙, ㇀—have developed, which combine according to prescribed stroke orders to form components (Zhang, 1990), see Figure 1.1 as an example.

Components constitute an essential intermediate layer of character construction and provide crucial visual features that distinguish characters from one another (Liang, 2004). In compound characters, components are smaller than the whole character, whereas in simple

characters they coincide with the entire graph (Su, 1994, 1995). For example, “每(měi)” in Figure 1.1 can be a simple character which means “every”, but it can also be used as a component in compound characters like “海 (hǎi, ‘sea’)", see Figure 1.2; or “莓 (méi, ‘berry’)", see Figure 1.3.

Figure 1.1

Strokes and stroke order of the Chinese character “每 (měi)”¹



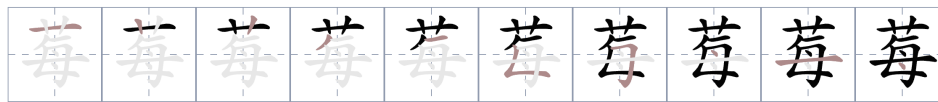
Figure 1.2

Strokes and stroke order of the Chinese character “海 (hǎi)”²



Figure 1.3

Strokes and stroke order of the Chinese character “莓 (méi)”³



In terms of structural arrangement, Chinese characters exhibit centripetal, cohesive, balanced, and stable visual patterns, forming neatly squared shapes (Fei, 1997). When strokes combine into components, they follow three main relational types—separated, connected, and intersecting (Fu, 1992)—and must be written sequentially according to specific stroke orders. Stroke order is crucial not only for writing speed and legibility but also for the use of stroke-based dictionary lookup methods⁴ (Fei, 1997). As components are further combined to form complete characters, spatial equilibrium must be maintained within a limited square frame.

¹ From <https://www.hanziwu.com/> (last accessed on 29/12/2025)

² From <https://www.hanziwu.com/> (last accessed on 29/12/2025)

³ From <https://www.hanziwu.com/> (last accessed on 29/12/2025)

⁴ Stroke-based dictionary lookup refers to non-phonetic character retrieval methods commonly used in Chinese dictionaries, where users identify a character by selecting its initial stroke type and/or total number of strokes. Successful use of such methods presupposes accurate stroke segmentation, correct identification of the first stroke, and reliable stroke counting, all of which depend on knowledge of standard stroke order.

Components typically adopt positional relationships such as top–bottom (e.g. 变, *biàn*), left–right (e.g. 海, *hǎi*), and (semi-)enclosing (e.g. 国, *guó*) structures, and the size and orientation of strokes often vary depending on the component’s position within the character (Shi, 1998). Components and their compositional patterns—such as positional arrangement, orientation, and connection methods—constitute key attributes for distinguishing one character from another (Wang, 1997).

Functionally, the informational content of a Chinese character is primarily conveyed through its components (Wang, 1997, 2002). Chinese characters are often classified as “phonosemantic characters” or “sound–meaning script” (Zhou, 1957), as they incorporate semantic components (意符, *yìfú*), phonetic components (音符, *yīnfú*), and additional symbolic elements (Qiu, 1988; Wan, 2004). Based on the relationship between a component and the corresponding whole character, components can be classified into three types: those providing semantic information (semantic components), those providing phonological information (phonetic components), and those unrelated to either meaning or pronunciation (Su, 1994; Wan, 1999). Wan’s (1999) statistical analysis of common characters identified approximately 246 semantic components and 1325 phonetic components, with the overall phonetic transparency of phonetic components averaging around 66% (Li & Kang, 1993).

Traditional Chinese character formation is based on Six Principles (六书, *liùshū*): indicative (指事, *zhǐshì*), pictographic (象形, *xiàngxíng*), phono-semantic compound (形声, *xíngshēng*), associative compound (会意, *huìyì*), derivative cognate (转注, *zhuǎnzhù*), and phonetic loan (假借, *jiǎjiè*). Among these, the phono-semantic compound characters constitute the vast majority—approximately 80% of all characters (Li et al., 1992; Li & Kang, 1995; Wan, 2000). Such characters combine a semantic component, which provides categorical meaning through abstract representation, and a phonetic component, which offers phonological cues (Song & Zhou, 2010). It is important to note, however, that the phonetic component does not always represent pronunciation with full accuracy in terms of onset, rhyme, or tone; overall phonetic consistency is estimated at about 66% (Li & Kang, 1993). For example, in the character 河 (*hé*, “river”), the semantic component 氵 indicates a water-related meaning,

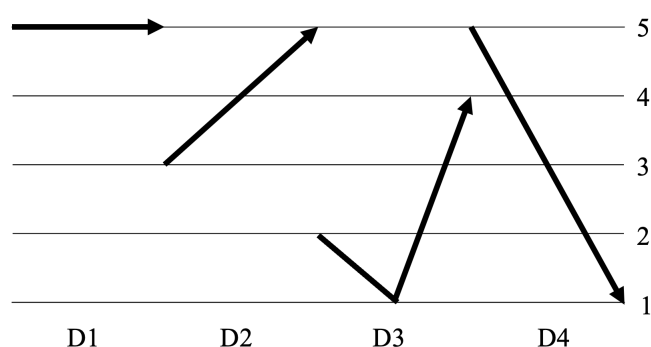
while the phonetic component 可 is pronounced kě. Although 可 provides a rough phonological cue, its pronunciation does not fully match 诃.

1.2.2 Chinese phonology, Hanyu Pinyin and the orthography-phonology interface

Chinese is a tonal language (Lin, 2004), in which every syllable carries a lexical tone, i.e. a specific pitch contour, and syllables that share the same segmental structure may differ in meaning when their tones differ. Chinese has five tones, conventionally referred to as the first, second, third, and fourth tones, as well as the neutral tone (hereafter abbreviated as D1, D2, D3, D4, and D0). Chao (1930) proposed the five-point tone marking system to describe relative pitch height during tonal production. This system uses a scale from 1 to 5, where 5 represents the highest pitch and 1 the lowest, and tonal contours are represented by indicating the starting and ending pitch levels. Emphasizing relative rather than absolute pitch height, the five-point system allows for cross-speaker comparability. The four lexical tones of Chinese can be stably represented within this framework, with tone contours of 55 for D1, 35 for D2, 214 for D3, and 51 for D4 (see Figure 1.4).

Figure 1.4

Five-point tone marking system



The neutral tone does not have a fixed tonal contour; its surface pitch realization depends on the tone of the preceding syllable and the prosodic context. It is therefore commonly treated as a dependent or toneless category. For this reason, the neutral tone is not assigned a specific pitch value in the five-point tone marking system and is typically excluded from standard tone contour diagrams.

In connected speech, Chinese tones are subject to tone sandhi, whereby the realized tonal

form differs from the citation tone (Cao, 1995). The most well-known sandhi rule in Chinese is the third-tone sandhi: when two third tones occur consecutively, the first is realized as a second tone (e.g., 你好 *nǐhǎo*). In addition, the third tone frequently surfaces as a half third tone in non-final positions, retaining only the low-falling portion without completing the rising contour. Beyond third-tone sandhi, phenomena such as neutralization, tone sandhi in connected speech, and intonational effects further influence tonal realization, highlighting the dynamic and context-dependent nature of Chinese tones in natural speech.

Against this background, the Chinese phonological system is commonly described through the Hanyu Pinyin (汉语拼音) system. Pinyin is a Romanization system officially promulgated in 1958 that has since been widely used in Chinese language teaching, lexicography, and linguistic research (Ke, 2003). It represents Chinese phonology at the syllabic level and serves as an important tool for phonetic analysis, transcription, and research on second language phonological acquisition. By encoding segmental structure through letter combinations and indicating pitch movement through tone marks, it provides a relatively complete representation of Chinese syllable structure. Beyond its function as a pronunciation aid, Pinyin also plays a crucial role in helping learners acquire the Chinese phonological system and establish form–sound correspondences (Wang, 2005b).

Within the Pinyin framework, a complete Chinese syllable typically consists of three components: an initial, a final, and a tone. Initials occur at the onset of the syllable and primarily correspond to consonantal phonemes, while finals occupy the syllabic nucleus and are composed of vowels or vowels followed by nasal codas, carrying the primary phonetic substance of the syllable. According to classifications commonly adopted in Chinese language pedagogy, Chinese has 21 initials and 39 finals, whose combinations form the basic syllabic inventory of the language. Compared with Indo-European languages such as English, Chinese is characterized by the absence of consonant clusters and a strong dominance of vowels: syllables do not permit sequences of two or more consonants, and each syllable must contain at least one vowel (Shi, 2010).

In the Pinyin system, tones are marked by diacritics placed above the vowel of the final,

corresponding to the four lexical tones and the neutral tone. The four lexical tones are indicated by diacritics (ˉ ˊ ˋ ˋˊ), whereas the neutral tone is left unmarked, reflecting its context-dependent pitch realization and its dependent status. Through the integration of tonal diacritics, the Pinyin system simultaneously encodes both segmental and suprasegmental features of Mandarin syllables.

It is important to distinguish Hanyu Pinyin from the International Phonetic Alphabet (IPA). Pinyin is a language-specific orthographic system designed primarily for Mandarin, serving pedagogical, normative, and lexicographic purposes. Its letter–sound correspondences do not necessarily align with universal phonetic values but rather reflect phonemic contrasts within the Chinese sound system. For example, the Pinyin initial *x* represents a voiceless alveolo-palatal fricative /ç/, and the Pinyin vowel *ü* represents a front rounded vowel /y/, both of which are category-based representations rather than direct phonetic transcriptions. In contrast, the IPA is a cross-linguistic system intended to provide a one-to-one representation of speech sounds based on articulatory and acoustic properties, allowing for fine-grained phonetic description. Thus, the same sound can be represented differently in Pinyin and IPA: for instance, the Mandarin syllable *xue* is written as *xue* in Pinyin, but transcribed as /çyɛ/ in IPA, where the symbols explicitly encode place of articulation and vowel quality. Consequently, the same Roman letter may represent different sounds in Pinyin and IPA; for example, *x* corresponds to /ç/ in Pinyin but to /ks/ in IPA-based conventions for languages such as English. In Chinese linguistics and CFL research, Pinyin is therefore more suitable for representing syllable structure, initial–final combinations, and tonal categories, whereas IPA is better suited for detailed phonetic analysis and the investigation of pronunciation errors, such as segmental substitution, coarticulatory variation, or fine-grained tonal realization.

Building on this distinction, it is necessary to consider the orthography–phonology interface in Chinese. Unlike alphabetic writing systems, Chinese characters constitute a morphosyllabic script in which there is no one-to-one correspondence between orthographic form and pronunciation. At the phonological level, a large number of characters converge onto the same pronunciation, giving rise to extensive homophony (e.g., 书 / 输 / 舒 all

pronounced as *shū*, 是 / 事 / 市 all pronounced as *shì*) (Chen & Ning, 2005). It has been estimated that only 16 Chinese characters lack homophones (Ma, 2005). At the same time, many characters are polyphonic, with a single orthographic form corresponding to multiple pronunciations depending on lexical or contextual factors (e.g., 行, *xíng/háng*; 重 *zhòng/chóng*) (Yang, 1995). This opacity in sound–form correspondence results in pronounced many-to-one and one-to-many mappings between orthography and phonology.

More specifically, homophony implies that phonological information alone is insufficient to uniquely identify a target character, whereas polyphonic characters require readers to integrate orthographic, lexical, semantic, and contextual information in order to select the appropriate pronunciation. Consequently, sound–form mapping in Chinese relies not only on segmental and tonal cues but also on higher-level lexical and semantic processing. This complex orthography–phonology interface constitutes a core mechanism underlying Chinese character recognition and reading aloud, and it imposes a distinct cognitive load on learners that differs fundamentally from that found in alphabetic writing systems.

1.2.3 The central elements of Chinese morphosyntax

Chinese is an analytic language, characterized by the absence of inflectional morphology and grammaticalized tense marking (Liu, 2015; Cao & Xu, 2017). Within the landscape of Chinese morphology, two domains are particularly central: the aspectual system and the classifier system.

Although Chinese is often described as a “tenseless” language, it exhibits well-developed and systematic aspectual distinctions (Li & Thompson, 1981; Xiao & McEnery, 2004). As an aspect-prominent language, Chinese expresses temporal relations primarily through temporal adverbials rather than inflectional marking on the verb (Jin & Hendriks, 2005). The core aspectual system consists of four functional particles: two perfective markers, 了 (*le*, post-verbal) and 过 (*guò*, post-verbal), and two imperfective markers, 在 (*zài*, pre-verbal) and 着 (*zhe*, post-verbal) (Klein et al., 2000). These aspect markers fulfill functions that parallel those expressed by inflectional morphemes in other languages, although they do not phonologically attach to the verb stem (Wang, 2012). The literature documents extensive analyses of their

semantic and grammatical properties, including their interaction with verb types and the encoding of event structure (Sun, 1993; Zhao, 1996; Wen, 1995; Jin & Hendriks, 2005; Yang, 2016). For example, the perfective marker *le* marks the completion of an event, as in 他吃了饭 (*tā chī-le fàn*, “he has eaten”), while *guo* encodes prior experience, as in 我去过北京 (*wǒ qù-guo Běijīng*, “I have been to Beijing”). By contrast, the imperfective marker *zai* precedes the verb to indicate an ongoing action, as in 她在看书 (*tā zài kàn shū*, “she is reading”), whereas *zhe* follows the verb to express a durative or resultant state, as in 门开着 (*mén kāi-zhe*, “the door is open”).

A second core component of Chinese morphology is the classifier system. Classifiers are grammatical morphemes that categorize nouns semantically and are obligatory in noun phrases containing demonstratives or numerals (Gao, 2010; Lau & Grüter, 2015). Mandarin is estimated to have around 70–80 commonly used classifiers (Erbaugh, 2004). Classifier selection is based on inherent noun attributes—such as shape, animacy, natural category, or functional features (Croft, 1994; Schmitt & Zhang, 1998; Chao, 1968; Matthews & Yip, 2011). Research provides extensive descriptions of the semantic principles underlying classifier selection and the distribution patterns associated with different classifier categories (Gao, 2009). In addition, the presence of a general classifier *ge*—which can apply to a wide range of nouns—has drawn scholarly attention regarding its grammatical status and semantic breadth (Liang, 2008; Polio, 1994). For example, the classifier *zhī* is typically used for long, thin, and animate entities, as in 一只猫 (*yì zhī māo*, “one cat”); *běn* is used for bound items such as books, as in 两本书 (*liǎng běn shū*, “two books”); and *tiáo* commonly categorizes long, flexible objects, as in 一条河 (*yì tiáo hé*, “one river”). By contrast, the general classifier *ge* can be used across a wide range of nouns, for instance 一个人 (*yí gè rén*, “one person”) or 一个问题 (*yí gè wèntí*, “one question”), often replacing more specific classifiers in both learner language and colloquial use.

Taken together, the aspectual and classifier systems reflect key features of Chinese morphology, involving intricate interactions among form, semantics, and grammatical distribution. These systems illustrate the typological characteristics of Chinese as a highly

analytic, non-inflectional language and form essential components of its grammatical architecture.

1.3 Learning challenges related to Chinese characters

1.3.1 Chinese character as an analysis unit

Learners from different linguistic backgrounds may perceive the difficulty of learning Chinese differently. For example, based on their language experience, CFL learners are often classified into “Sinosphere” learners and “non-Sinosphere” learners (Zhou, 1989). The Sinosphere refers to East Asian regions that currently or historically use Chinese characters as an important medium for transmitting language and culture, and whose languages have incorporated a large number of Sino-lexical items; these regions primarily include China, Japan, and Korea (Feng, 2004; Lin, 2006). In contrast, non-Sinosphere is typically referred to learners whose native languages are alphabetic and who have had no prior exposure to Chinese characters (You, 2003; Du, 2007). Learners from the Sinosphere are likely to experience fewer difficulties in acquiring Chinese characters (Liu, 2017). This is because Chinese differs fundamentally from alphabetic languages in its phonology, orthography, and morphology (Shen & Xu, 2015; Ma et al., 2017).

As mentioned in Section 1.2.1, Chinese character as the smallest distinguishable unit of the Chinese writing system possesses an identifiable visual configuration (Ren, 2024) while simultaneously encoding orthographic, phonological, and semantic information (Zhu, 2002). This multidimensional nature grants the Chinese character a central role in cognitive processing, making it an important entry point for understanding the development of interlanguage in learners. As discussed in Section 1.2, the major phonological and orthographic challenges faced by CFL learners are all rooted in the character-based nature of Chinese, and many of the typological features of the language can ultimately be traced to this *zi*-centered linguistic structure (Zhang, 2025). Consequently, the character is one of the major sources of difficulty in learning Chinese as a foreign language.

From the perspective of interlanguage development, the Chinese character is often the earliest acquired, most central, and most error-prone processing unit in the learning process. It

thus provides a clear window into the transitional, systematic, and unstable nature of learners' interlanguage. Factors such as visual complexity, morpho-semantic motivation, the presence of homographic and visually similar characters, and the opacity of sound–form mappings all contribute to the challenges learners encounter (Zhang & Xing, 2023). As emphasized by Lü (1987), Chinese characters and alphabetic scripts differ fundamentally in the cognitive mechanisms they require: the former relies on the spatial organization of strokes and components in a two-dimensional structure, whereas the latter depends on linear phoneme-based spelling principles. This fundamental contrast means that learners must confront the structural complexity of the character from the very beginning of their exposure to Chinese, with the single character serving as the essential entry point into the writing system. For this reason, focusing on the character level holds particular value in research on Chinese interlanguage.

1.3.2 Challenges in character writing

Chinese orthography has long been recognized as one of the greatest challenges in learning Chinese as a foreign language, posing substantial difficulties in writing, character recognition, and character use. Characters are often perceived as the most difficult component of the language: in Razali's (2015) survey, 40.9% of learners identified Chinese characters as the single hardest part of learning Chinese. For learners whose native languages employ alphabetic writing systems, mastering Chinese orthography is an essential first step toward developing literacy (Zhu & Halina, 1999). Existing studies likewise indicate that learners from non–Sinosphere backgrounds almost unanimously describe Chinese characters as “difficult” or “very difficult” (Shi & Wan, 1999). Ma (2003) estimates that acquiring the number of characters and vocabulary items required by standard CFL curricula demands approximately 3000 instructional hours, whereas an alphabetic writing system of comparable lexical size may require only about 750 hours. Xu (2003) further argues that for learners whose linguistic systems and cultural backgrounds differ substantially from Chinese, familiar cognitive frameworks often collapse when confronted with the Chinese writing system. This sudden breakdown not only causes processing difficulties but also triggers psychological discomfort,

anxiety, or even resistance, thereby exacerbating the challenge of learning characters.

The difficulty of writing Chinese characters stems first from their structural complexity. As outlined in 1.2.1, characters are two-dimensional graphic units composed of strokes, components, and formation rules (Liu, 2005), and they carry high visual complexity. Chinese characters contain more than twenty types of strokes, many of which differ only subtly and are therefore difficult to distinguish. The average number of strokes per modern character is 10.75, with the most common range being 9–11; the simplest characters consist of only a single stroke, whereas the most complex may contain as many as 36 (Liu, 2005). A second difficulty lies in the sheer quantity of characters: approximately 3000 are considered common-use characters (Fei, 1998), and at least 2000 are needed to read a typical newspaper (Alshammari, 2020). In addition, the iconic and semantically transparent features that characterized early Chinese characters have been significantly weakened through centuries of evolution and simplification (Xu, 2009). For learners accustomed to alphabetic writing systems—who are naturally less sensitive to form–meaning connections—the reduced degree of semantic motivation in modern characters further limits their ability to infer or memorize meaning from visual form (Jin, 2015).

Incorrectly written characters refer to written forms that are intended to represent a specific target character but deviate from the standard orthographic form, typically as a result of errors in strokes, components, or structural configuration. In such cases, learners select the correct lexical item at the semantic level, yet fail to produce its conventional written form accurately. These deviations may involve missing or extra strokes, incorrect stroke shapes, misplacement or substitution of components, or violations of character structure rules, while the intended meaning remains identifiable to varying degrees.

Although handwriting has traditionally been regarded as central to the acquisition of Chinese orthographic knowledge, its necessity has increasingly been questioned in the digital age (Allen, 2008; Zhang & Lu, 2014). In contemporary learning contexts, the most commonly used Chinese input method is Pinyin-based typing, whereby learners enter syllables without tone marks and select target characters from automatically generated candidate lists. Many input systems further incorporate “fuzzy matching” functions, which allow correct character

selection even when the Pinyin input is incomplete or inaccurate. As a result, learners are often able to produce orthographically correct output without fully encoding or memorizing character structure, thereby weakening the development of stable orthographic representations (Zhu et al., 2009).

Empirical studies have shown that, under comparable task conditions, typed texts typically exhibit higher surface-level accuracy than handwritten texts (e.g., Kang, 2011; Zhang & Min, 2019). However, this apparent accuracy relies heavily on technological assistance rather than on learners' internalized orthographic knowledge. Consequently, typing-based performance may mask underlying weaknesses in learners' control of character structure and form–meaning mapping, making handwriting a more sensitive indicator of learners' true orthographic competence.

From a cognitive perspective, incorrectly written characters reflect learners' incomplete mastery of Chinese orthographic representations. Incorrectly written characters arise primarily from difficulties in encoding, storing, and retrieving detailed visual–orthographic information. The high visual complexity of characters, the large inventory of forms to be memorized, and the weak form–meaning transparency of modern Chinese characters jointly increase the likelihood of such errors. As a result, incorrectly written characters constitute a core manifestation of orthographic difficulty in CFL learning and provide a crucial window into learners' developing interlanguage writing system.

1.3.3 Challenges in character use and recognition

Challenges also arise at the level of character use. Because Chinese characters encode syllables and modern Chinese has only around 1300 distinguishable syllables (Lu, 2001), the number of characters far exceeds the number of available syllables. This inevitably results in large sets of homophones, near-homophones, and visually similar characters (Zhang, 2010). Learners lacking a background in Chinese orthography often have unstable or incomplete mental representations of components and similar characters, leading to weak form–meaning connections and frequent character confusions or misselections (Feng, 2015). Moreover, the phonetic transparency of Chinese characters is low—character form does not reliably indicate

pronunciation—and although semantic transparency exists to some degree, it is not always explicit in modern forms. As a result, learners often find that characters are difficult both to “read for sound” and to “read for meaning,” which stands in sharp contrast to the processing routines used for alphabetic writing systems.

At the level of character use, a particularly common error type is that of incorrectly used characters. In such cases, learners produce a well-formed and standard character, but select an inappropriate one for the intended meaning or context, often due to homophony, near-homophony, or visual similarity among characters. For example, characters such as 是 (*shì*, “be”) / 事 (*shì*, “thing”) / 市 (*shì*, “city”) share the same pronunciation but differ in meaning and use, increasing the likelihood of misselection. Importantly, incorrectly used characters differ fundamentally from incorrectly written characters: whereas the latter involve deviations from the conventional orthographic form (e.g., missing strokes or malformed components), errors of character use arise from failures in lexical or semantic selection despite accurate orthographic production. As such, incorrectly used characters primarily reflect weaknesses in form–meaning mapping and lexical differentiation rather than deficits in orthographic knowledge itself.

Character recognition presents an additional layer of difficulty. Alphabetic-language learners rely heavily on phonological pathways during reading-aloud, but the low phoneticity of Chinese characters prevents them from quickly accessing pronunciation through orthographic form, substantially increasing cognitive load (Jin, 2015). Xu (2000) observes that learners from non–Sinosphere backgrounds typically progress through three stages before achieving stable character recognition: initially linking sound and meaning, then integrating character form with meaning, and finally establishing firm associations between form and sound. Only after completing this multi-step process do learners develop stable and reliable orthographic representations.

These cumulative difficulties align with Lü’s (1995) well-known summary of the “five difficulties” of Chinese characters—difficult to write, large in number, complex interactions among form, sound, and meaning, difficult to recognize, and difficult to look up—which

remain accurate descriptors of the challenges faced by today's CFL learners. In light of these obstacles, Zhang (2010) argues that effective character instruction must integrate form, sound, and meaning. By using structural analysis of characters as the foundation and linking orthographic form with pronunciation and semantic information, instructors can help learners build a more systematic understanding of the Chinese writing system and alleviate the overall learning burden.

Character recognition difficulties may also manifest in cases where characters are written correctly but read incorrectly. In this type of error, learners are able to retrieve and reproduce the correct orthographic form, yet fail to access or produce the appropriate pronunciation during reading-aloud. This dissociation indicates that orthographic representations can be relatively stable while the mapping between character form and phonology remains fragile. Such errors are especially common in the presence of homophones and polyphonic characters, where orthographic cues alone are insufficient to determine pronunciation. Errors of this kind therefore reflect problems at the orthography–phonology interface, rather than at the level of writing accuracy or lexical meaning selection.

1.3.4 Challenges in character reading and pronunciation

The difficulty in reading Chinese characters stems largely from the opaque nature of their orthography (Yan & Lin, 2023; Wang et al., 2020). Unlike highly transparent alphabetic systems such as Italian and German, the mapping between graphic forms and phonological representations in Chinese is far from regular. For instance, a single character can carry multiple pronunciations, and the phonetic components in phonograms do not consistently provide reliable cues to pronunciation. As a result, the orthography-to-phonology mapping in Chinese is best characterized as a mixture of transparent and arbitrary correspondences.

However, the challenge does not stop at resolving this opaque mapping; even when the pronunciation of a character is known, producing it accurately adds another layer of difficulty. For most CFL learners, the perception and production of tones constitute a major and persistent challenge (Hao, 2012). Inaccurate tone control often results in a noticeable “foreign accent” in learners' spoken Chinese (Lin, 1996; Shi & Wen, 2012).

The difficulty of tone acquisition is closely related to the neural mechanisms involved in tone processing. Research indicates that Chinese native speakers primarily rely on the left brain hemisphere to process tones, treating them as integral linguistic units, whereas learners whose L1 is a non-tonal language (e.g., English) tend to engage both hemispheres when processing Chinese tones (Wang et al., 2001; Guo & Tao, 2008). As Wang et al. (2003) point out:

Learning a tone-based second language involves both expansion of pre-existing language-related areas and recruitment of additional cortical regions specialized for functions similar to the new language functions. (p. 1019)

This suggests that tone acquisition requires not only auditory and phonetic learning but also substantial neural reorganization. Tone learning difficulty is also influenced by language transfer. Studies on English-speaking learners (White, 1981; Chen, 1997) show that tone errors partly stem from learners' transfer of English intonation patterns into Chinese. Furthermore, Li (2023) demonstrates that even learners whose L1s are tonal languages (e.g., Bengali, Thai) may incorrectly assimilate Chinese tonal contours into their native tonal categories, resulting in negative transfer and making tone acquisition more challenging. Similarly, Hao's (2012) study on Cantonese speakers shows that the presence of Cantonese tones does not necessarily facilitate the learning of Chinese tones.

At the segmental level, as mentioned in 1.2.2, Chinese syllables consist of initials (consonants) and finals (vowels). Learners often struggle with contrasts such as retroflex vs. non-retroflex initials, aspirated vs. unaspirated stops, as well as nasal and diphthongal finals. Nevertheless, research indicates that with systematic training, most learners can approximate native-like segmental pronunciation within a relatively short period (Xu, 2003). Cross-linguistic comparisons consistently show that error rates for initials and finals are far lower than those for tones (Ran & Yu, 2019).

At the prosodic level, tone sandhi rules pose additional challenges. Tone sandhi refers to context-dependent tonal alternations at the synchronic level. Learners' tone-sandhi errors typically manifest as insufficiently precise or inaccurate tonal contours, which may be attributed to underdeveloped phonological representation of allophonic variants (Chen et al.,

2019). Moreover, the acquisition of tone sandhi in connected speech differs from the acquisition of citation tones; mastery of the former does not necessarily guarantee mastery of the latter (Jin, 2019).

In sum, Chinese phonological system presents multi-layered challenges at the segmental, tonal, and prosodic levels. These difficulties arise from typological differences, neural processing patterns, cross-linguistic transfer effects, and the complexity of tone-sandhi rules, making tones one of the central bottlenecks in learning Chinese as a foreign language.

1.4 Italian and Chinese: Cross-linguistic contrast and specific learning difficulties

1.4.1 Italian and Chinese in contrast

Italian and Chinese exhibit significant differences in their phonological systems, writing systems, grammatical structures, and semantic organization. Existing research indicates that the interlanguage formed by learners during Second Language Acquisition (SLA) is systematically influenced by their L1. This cross-linguistic transfer effect directly reflects how learners reconstruct the rules of the target language (Benson, 2002; Ellis, 2015; Jarvis, 2015).

Italian employs an alphabetic writing system, whereas Chinese uses a morphosyllabic script. This difference entails distinct cognitive demands in literacy acquisition: while Italian literacy is built upon phoneme-level decoding, Chinese literacy requires the acquisition of complex visual forms and their associations with sound and meaning.

In addition, as discussed in Section 1.3.4, the orthography of Chinese is highly opaque, and the mapping between graphic forms and phonological representations is irregular. Psycholinguistic studies have shown that native speakers of Chinese tend to rely more on visual processing pathways, taking character form as the primary cue in recognition and memory tasks, while the activation of phonological information is often secondary or delayed (Guo, Tang, & Li, 2007; Perfetti & Tan, 1998). In other words, the cognitive processing of Chinese generally follows a form-to-sound trajectory, in which individuals first access visual character information and then activate the corresponding phonological representation. In contrast, the

Italian orthographic system is highly transparent, with an almost one-to-one correspondence between sound and spelling. Italian native speakers are accustomed to a sound-to-form mapping in their native language processing, deriving written forms directly from phonological cues. This long-term orthographic experience may transfer to their Chinese learning, leading them to rely more heavily on phonological routes rather than orthographic ones when processing Chinese characters.

Finally, Italian is a non-tonal language in which pitch variation serves primarily intonational or pragmatic functions, whereas Chinese is a tonal language in which pitch contrasts carry lexical meaning. For Italian learners of Chinese, the absence of lexical tone in their native phonological system means that tone must be acquired as a new contrastive dimension, rather than mapped onto existing categories. This typological mismatch further complicates the acquisition of Chinese pronunciation and the establishment of stable sound–form correspondences.

1.4.2 L1-specific character writing difficulties

Existing studies on incorrectly written Chinese characters can be broadly divided into two main types. The first type includes general investigations that do not distinguish learners by their L1 background. These studies focus on the universal characteristics of writing errors among learners of different linguistic backgrounds and explain them from the perspective of cognitive mechanisms (Shi, 1991; Gao, 2001; Xiao, 2002; Liu, 2008; Guo, 2008; Wang, 2010; Liang, 2019). The second type targets learners with specific L1 backgrounds, exploring how their native orthographic systems shape the manifestation of Chinese character errors. However, most of this research has concentrated on learners from East and Southeast Asian language groups, such as Thai (Liu, 2008; Chen, Liu & Zhu, 2009; Chen, 2014), Korean (Xu & Sun, 2014; Park, 2011; Ma, 2016), Japanese (Han, 2014; Liu, 2011; Ding, 2013; Wang, 2015), and Cambodian (Lv, 2014; Duan, 2019; Hao, 2019).

In contrast, studies specifically addressing L1 Italian learners remain remarkably limited. For instance, Centamore (2020) analyzed Italian learners' Chinese writing errors using a small-

scale dataset (183 instances) and partial HSK Dynamic Essay Corpus⁵ data, which was insufficient for reliable generalization. Qian (2016) investigated beginner-level Italian learners only, without covering intermediate or advanced learners. Sun (2021) examined Italian–Chinese heritage students, whose bilingual background, however, makes their writing behavior less representative of true L1 Italian learners. Liu (2017) conducted a survey at the University of Venice focusing primarily on learning strategies and motivation; the Chinese character writing data were collected via a 20-item fill-in-the-blank test, which could not adequately capture learners’ natural writing performance. Overall, these findings indicate that systematic research on incorrectly written Chinese characters by L1 Italian learners remains largely underrepresented.

As discussed in Section 1.2, the constitution of a Chinese character is a multi-layered, nested system. Chinese characters can be analyzed at three interrelated levels: strokes, components, and overall structure, the classification of incorrectly written characters is necessarily closely linked to the analysis of these three constitutive elements. However, existing research still exhibits disagreement on how to define these levels and their interrelationships; the scope and defining criteria for the three elements—stroke, component, and structure—in the classification of incorrectly written characters are not entirely consistent among different scholars.

Some scholars, when studying Chinese character writing errors, did not distinguish between the incorrectly written character and the incorrectly used character, instead treating them collectively as Chinese character writing errors and classifying them uniformly. For example, Hu (2013), Tian (2022), and Kang (2023) all adopted a “three-way classification”, dividing Chinese character writing errors into stroke errors, component errors, and whole-character errors. The “whole-character error” in these systems is separately categorized as an incorrectly used character in this study. Thus, for the incorrectly written character as defined here, these scholars effectively utilized a “dichotomous approach”, classifying errors along the two dimensions of stroke errors and component errors. Similarly, Gan (2020) also divided

⁵ Further details about this corpus can be found in Section 2.3.1.

learners' Chinese character writing errors into two categories—stroke errors and component errors—but did not distinguish between incorrectly written and incorrectly used concepts. Instead, Gan uniformly included all formal writing errors into these two categories based on the graphic form.

In contrast, Deng & Hu (2022) offered a more refined classification system, distinguishing learners' writing errors into two broad categories: incorrectly written and incorrectly used. The incorrectly written category was further detailed into three types: incorrectly written strokes, incorrectly written component, and incorrectly written characters (no detailed explanation is given, but based on the examples provided it can be understood that there are multiple errors). Wei (2017) proposed a four-way classification based on this, dividing errors into stroke errors, component errors, structure errors, and misuse errors. If the “misuse error” is excluded as a use-level error, Wei's division of the incorrectly written character essentially added the structure error level to the “dichotomous approach”. Feng (2019) further expanded Wei's (2017) framework by adding the stroke order error type, making the classification system even more detailed. In addition, Fan (1993) proposed a more complex eight-category division: stroke form errors, stroke addition/omission errors, stroke arrangement errors, component addition/omission errors, component arrangement errors, comprehensive stroke and component errors, and incorrectly used characters, among others.

1.4.3 L1-specific character use difficulties

In existing research on learners' Chinese character writing, some scholars have broadly categorized incorrectly used character as a type of glyph writing error, discussing them alongside incorrectly written characters without clearly distinguishing between the execution-level error of “failing to form the correct character” and the use-level error of “writing a legitimate but incorrect character”. For instance, Xiao (2002), Li (2014), Gong (2014), Liu (2016), Wei (2017), Gan (2020), Tian (2022), Kang (2023) all integrate incorrectly used character into a broader category of glyph writing errors.

Other scholars have treated incorrectly used character as an independent object of study, with a significant portion of this research relying on existing Chinese interlanguage corpora.

For example, Zhao (2013) and Tian (2018) both conducted analyses using the HSK Dynamic Composition Corpus. However, this corpus contains a total of 585 compositions, with only 14 contributed by L1 Italian learners, an extremely low proportion. Zhang & Lei (2021) focused their research on high-level English-speaking learners, and while Deng & Hu (2022) included eight countries across three language families, they did not specify the data distribution, only noting that L1 Italian learners represented approximately one-eighth of the Romance language group. In general, focused investigations specifically targeting L1 Italian learners remain limited. Moreover, with the exceptions of Zhang & Lei (2021)—who incorporated grouping by proficiency level—the majority of other studies have adopted a holistic approach that overlooks differentiation by learner proficiency.

Some scholars (e.g. Gao & Meng, 2000) have adopted a dichotomous approach, dividing incorrectly used characters according to error source into phonologically identical characters and graphically similar characters. However, this classification is overly general and fails to account for situations where both the sound and the form of the Chinese character influence the learner. Some other scholars (e.g. Jiang & Liu, 2004) have employed a four-category classification, referencing both sound and form to divide incorrectly used characters into: phonologically similar characters, graphically similar characters, phonologically and graphically similar characters, and other errors. This system is considered more comprehensive as it accounts for various potential scenarios. Some scholars have further subdivided these four categories. For instance, Wu et al. (2006) refined phonologically similar characters into “phonologically identical” and “phonologically similar.” However, this refinement implies a prerequisite assumption: that the learner has clearly mastered the correct pronunciation and tone of the Chinese character. For second language learners, especially those at lower-to-intermediate proficiency levels, this assumption is not always reasonable, which may limit the explanatory power of the classification.

Other scholars have added new categories to the four-part system: Tian (2018) proposed “in-word substitution,” such as 传染 [B 污] 病 (*chuánrǎn[wū]bìng*, “infection,” where *wū* “dirty” is used instead of *rǎn* “infect”), while Zhao (2013) adopted a three-category

classification, dividing incorrectly used characters into phonologically similar characters, graphically similar characters, and homophonous heterographic words (a group of words with identical pronunciation but unrelated meanings, such as 反映 (*fǎnyìng*, reflect) vs. 反应 (*fǎnyìng*, reaction), where 映 and 应 are the misused characters). Deng and Hu (2022) proposed “Misused word/phrases”, such as 自行车 (*zìxíngchē*, bike) vs. 己行车 (*jǐxíngchē*). All three of these classification methods involve misused character discrimination within the higher-level unit of the “word”, which exceeds the scope of single-character writing errors and can thus introduce ambiguity into the analysis of learner corpora.

Beyond these, some researchers have also included character meaning in their classification criteria. Tian (2022) established four categories under incorrectly used characters: phonologically similar characters, graphically similar characters, phono-graphically similar characters, and phono-semantically similar characters; Peng (2001) proposed four categories: graphically similar characters, phonologically and graphically similar characters, phonologically and semantically similar characters, and phono-graphic-semantically similar characters; and Zhang & Lei (2021) even proposed eight categories, with causes detailed down to: phonologically identical/similar, graphically similar, semantically similar, phono-graphically similar, graphic-semantically similar, phonol-semantically similar, phono-graphic-semantically similar, and phonol-graphic-semantically unrelated.

Deng & Hu (2022) conducted their study based on a dynamic interlanguage corpus of international students established at a Chinese university. That corpus, totaling approximately 100,000 characters primarily sourced from students’ written exam essays, was divided into three levels—Beginner, Intermediate, and Advanced—based on assessments using the unified written and spoken proficiency tests. Their sample covered learners from eight countries across diverse L1 backgrounds, including Germanic (English, German, Swedish), Romance (Italian, French, Spanish), and Slavic (Russian) language families. Their results indicated that the overall frequency of incorrectly used characters consistently declined as the proficiency level increased, with a significant drop from the Beginner to Intermediate stage, and a continued, though slowed, decline from Intermediate to Advanced.

In contrast, Zhang's (2025) research was based on an interlanguage character corpus from international students at Sun Yat-sen University, which collected routine essay samples written at different stages by students from over 70 L1 backgrounds. The results also showed a trend of increasing frequency of character misuse from Beginner to Intermediate and then decreasing from Intermediate to Advanced, consistent with the findings here. Zhang's (2025) study was also based on non-examination materials, with data derived from classroom and natural writing tasks, and it distinguished between *quasibiezi* (near-misused character) and *biezi* (misused character) errors in the statistics. Even though this study did not make that distinction, combining the two error types similarly yielded a trend of increasing from Beginner to Intermediate and then decreasing from Intermediate to Advanced.

The research by Wu et al. (2006) and Gao & Meng (2000) showed that graphically similar characters consistently remained the dominant proportion across all stages of Chinese character proficiency and did not exhibit a noticeable downward trend. The primary reason for this discrepancy likely lies in the differences in research task type and corpus nature. Gao & Meng (2000) employed a proofreading paradigm, which required participants to identify and correct errors in a text. This task primarily emphasizes visual discrimination and attentional control, thereby tending to amplify the proportion of graphically similar characters confusion. Wu et al. (2006) utilized a dictation task, which required learners to write Chinese characters based on auditory input. This task inherently relies on the immediate conversion from sound to character form, which increases the potential for both graphically similar characters and phonologically similar characters interference.

1.4.4 Pinyin as a transitional strategy under alphabetic influence

For learners whose native language is Italian, the frequent use of Pinyin in place of Chinese characters can be understood as a natural consequence of cross-linguistic transfer. As a result, Italian learners tend to rely strongly on sound-based representations when accessing lexical meaning. When this phonology-driven strategy is transferred to the learning of Chinese, learners may resort to Pinyin as an accessible and reliable means of representing meaning when the retrieval of character form is uncertain.

The frequent substitution of Pinyin for characters has often been excluded from analyses of character writing on the grounds that it represents an avoidance strategy rather than a genuine orthographic error (Jiang & Liu, 2004). From a developmental perspective, however, Pinyin substitution provides valuable insight into learners' reliance on phonological representations and their uncertainty in character retrieval during the writing process.

The relative scarcity of research on Pinyin use also reflects divergent views on how Pinyin substitution for Chinese characters should be conceptualized. Jiang and Liu (2004), for example, explicitly excluded Pinyin from their analysis of character writing, arguing that substituting Pinyin for characters one cannot write does not reflect learners' orthographic knowledge. From the perspective of language development, however, Pinyin use—while not directly indicative of orthographic competence—can still reveal the depth of learners' character knowledge, the nature of sound–form connections, and learners' cognitive orientation during character production.

Rather than being treated as a conventional “error”, Pinyin substitution should therefore be understood as a transitional writing strategy. It reflects learners' uncertainty in character retrieval and provides indirect evidence of the cues—phonological or visual—that they rely on during recall. As the preceding analysis of incorrectly used characters suggests, Italian learners of Chinese tend to rely more heavily on phonological cues than on visual–orthographic information. Previous studies similarly indicate that learners whose native languages have transparent orthographies often develop phonology-oriented processing strategies, which, when transferred to Chinese character learning, result in meaning retrieval that is strongly mediated by pronunciation (Everson, 1998; Jiang, 2003). This phonological orientation is further manifested in learners' use of Pinyin.

1.4.5 Reading-related difficulties and phonological transfer

Chinese is a tonal language in which pitch functions contrastively to distinguish lexical meaning. For learners whose first language is non-tonal, the perception and production of tones constitute one of the greatest challenges in Chinese phonetics acquisition. Previous studies have pointed out that tone errors among non-native learners primarily result from their failure to

establish stable tonal categories in both perception and production, leading to tonal confusion during recognition and imitation (Chen, 2011). Such difficulties are not limited to beginners but may persist even after extended learning (Mao, 2007). Therefore, a systematic analysis of the distribution and substitution patterns of tone errors is crucial for understanding the structure of learners' interlanguage phonological system and for improving pedagogical approaches.

As discussed in Section 1.2.2, Modern Standard Chinese has a syllable-based structure consisting of an initial and a final, with no consonant clusters and a phonological system dominated by voiceless consonants. Compared with Italian, Chinese differs markedly in aspiration contrasts, vowel nasalization, and the integration of tones within the syllable, and these structural differences often lead to systematic segmental and syllabic-level deviations in learners' production.

Previous research on tone acquisition among learners from various L1 backgrounds has produced extensive but inconsistent findings. Teng (2025) found that Uyghur learners' errors were mainly concentrated in Tone 2 (rising tone) and Tone 3 (dipping tone). For convenience, tones are hereafter referred to as D1–D4 (corresponding to Tone 1–Tone 4). Among English-speaking learners, some studies argued that D1 (high-level tone) and D4 (falling tone) were harder to master, while others found D2 and D3 to be more error-prone, suggesting variability in the perception of level versus contour tones (Shen & Gui, 2000). Thai learners, despite having a tonal L1, showed a pattern in which D1 and D4 were more difficult than D2 and D3 (Yi & Liang, 2010), indicating that high-level and falling tones posed greater challenges. Japanese and Korean learners (Zhang et al., 2006; Dong et al., 2013) commonly confused D2 and D4, suggesting that speakers of pitch-accent languages often experience perceptual ambiguity between rising and falling tones. A mixed-group study (Li & Zhao, 2012) found that tone difficulty generally followed the order from D1 to D4 to D2 and finally D3, while Guo and Tao (2008) reported that D1 and D4 had the highest accuracy rates, with D3 being the most error-prone. Mao (2007) and Zhang (2012) both found that, among non-tonal learners, tone error frequency was highest for D3, followed by D2, then D4, and lowest for D1, highlighting D3 and D2 as the most challenging tones. Wang (2006) identified two common confusion

patterns from multi-syllabic dictation and reading tasks: (1) D2–D3 confusion, widely observed across L1 groups; and (2) D1–D4 confusion, which was more influenced by syllable position and L1 transfer. In general, D3 is consistently regarded as the most difficult tone to acquire, followed by D2, while D1 tends to be the easiest to master. Compared with tone acquisition, research on the acquisition of initials and finals among Chinese language learners is relatively limited, and the existing studies have primarily focused on learners from Asian language backgrounds—for example, Korean (Ma, 2004; Wang, 2000; Wang & Ju, 2014) and Japanese (Wang & Shangguan, 2004; Li, 2011). Among non-Asian languages, relevant studies have examined learners with English (Tan, 2007; Yang, 2015) and French (Huang, 2022) as their L1s, while systematic investigations on L1 Italian learners are still lacking.

Moreover, previous research shows a clear imbalance in focus: most studies have concentrated on the acquisition and errors of initials (e.g., Liu, 2013; Shi, 2014), while investigations into the acquisition, error types, and systematic patterns of finals remain scarce. This imbalance limits our understanding of how learners develop an integrated phonological system in Chinese as a second language.

Previous studies on the acquisition of Chinese initials by non-native Chinese learners have shown that learners commonly experience difficulties in controlling the place of articulation and aspiration. Huang (2022) found that English-speaking learners had the lowest accuracy rates in producing *j*, *q*, and *x* (/tɕ/, /tɕʰ/, and /ɕ/ respectively), three alveolo-palatal sounds that do not exist in English. These learners often demonstrated insufficient fronting of the tongue or improper airflow control. Shen and Yu (2009) observed that learners frequently made errors in distinguishing between aspirated and unaspirated sounds, as well as between voiceless and voiced sounds, often exhibiting phenomena such as voicing of voiceless consonants or insufficient aspiration. Bi (2001) reported that among the 21 Chinese initials, American students found the most difficult ones to be *zh*, *ch*, *sh* and *j*, *q*, *x* (/tʃ/, /tʃʰ/, /ʃ/ and /tɕ/, /tɕʰ/, /ɕ/ respectively), which are absent from English phonology. The next most challenging contrasts were those between aspirated and unaspirated consonants, as well as certain sounds similar to English phonemes but differing in articulation, such as *r*, *h*, and *t*.

Deng (2020) further noted that the three consonant groups *j, q, x*; *z, c, s*; and *zh, ch, sh* have consistently been the most difficult for foreign learners, with pronunciation errors primarily stemming from inaccurate articulation and unstable tongue positioning.

1.5 Teaching Chinese as a foreign language in Italy

The history of Chinese language education in Italy dates back to the first half of the 18th century (Jin & Zhuang, 2019), making it “probably one of the first European countries where the Chinese language was taught” (Masini, 2009, p.157). Early Chinese classes were mostly established by missionaries, and the purpose of learning Chinese at that time was primarily related to religious activities (Masini, 2009; Yang, 2014). In contemporary times, however, the number of Chinese learners in Italy has increased exponentially, and the scale of Chinese learning has ranked among the top in Europe for many consecutive years (Jin & Shi, 2019). Italy currently has more than 90 comprehensive universities, over 40 of which offer Chinese courses or Chinese-related degree programs (Li & Li, 2021).

Regarding teaching personnel, in addition to Italian local teachers, a considerable number of Chinese teachers are active in Italian schools. Some of them have lived in Italy for many years and have a good command of the Italian language; others are exchange teachers sent jointly by the Chinese and Italian governments, and they usually teach in English or Chinese (Masini, 2009; Yang, 2014). While this group of teachers can provide students with a relatively immersive learning environment, many of them lack knowledge of Italian, and some even adopt Chinese-language teaching methods drawn from domestic language arts classes. As a result, their preparation in international Chinese education and cross-cultural pedagogy is sometimes insufficient (Jin & Shi, 2019). Moreover, this group of teachers tends to circulate frequently and thus presents a relatively low level of stability. At the University of Bologna, for example, some courses in the Department of Interpreting and Translation are taught by Chinese teachers appointed by the Confucius Institute. These teachers are often graduate students in Teaching Chinese to Speakers of Other Languages in China, and their teaching terms usually last for one semester or one academic year.

In terms of learning resources, as of 2023, a total of 155 Chinese textbooks appropriate

for learners of different age groups had been published in Italy (Ma, 2025). At the higher education level, the most widely used textbook is *Il cinese per gli italiani* (Zhu & Sun, 2014; Jin & Shi, 2019; Gong, 2022), which is often supplemented by self-developed teaching materials created by instructors (Gong, 2022). According to Ma (2025), more than 100,000 copies of the *Il cinese per gli italiani* textbook series were distributed between its initial publication in 2006 and 2021, and its use rate exceeds 80% in Chinese majors at universities in Italy and San Marino. The textbook is considered comprehensive or “cross-domain” (Sun, 2012; Ma, 2025), covering Chinese characters, grammar, phonetics, and cultural components. While a range of teaching materials for Chinese as a foreign language is available in the Italian context, including resources specifically designed for character writing, *Il cinese per gli italiani* is selected here as a representative example due to its widespread adoption and its use in the teaching context from which the present data were collected.

Regarding Chinese character instruction, *Il cinese per gli italiani* adopts the “character learning within texts” approach (Jin, 2015; Xu & Yao, 2014), in which new characters are introduced while learning the accompanying texts. This method has the advantage of presenting characters within contextualized discourse, thereby facilitating the integrated development of listening, speaking, reading, and writing skills (Jin, 2015). However, this also results in beginners being exposed to relatively complex characters at a very early stage. For instance, the textbook begins character instruction from Lesson 1 (Sun, 2012), and characters such as 谢 (*xiè*) / 学 (*xué*) / 现 (*xiàn*), which have relatively complex structures, already appear in Lesson 3 of the first volume.

In actual classroom practice, based on classroom observations conducted at the University of Bologna, I observed that most teachers adopt what can be described as a largely “text-driven” approach to character instruction. By “text-driven,” I refer to a reactive instructional pattern in which character teaching is triggered by the appearance of new items in the teaching text, rather than being planned as a structured or systematic component of instruction. These observations were drawn from auditing several Chinese language courses, including introductory Chinese classes, elective Chinese courses, and Chinese–Italian

translation courses. When new characters appear in the course materials, teachers typically provide immediate explanations to support comprehension, but rarely deliver character-focused lessons as an independent instructional component. This approach differs from the commonly discussed “character learning within text” approach, which integrates character learning into textual contexts in a more deliberate and pedagogically structured way. In contrast, the text-driven approach observed here tends to be incidental and explanation-oriented, focusing on short-term understanding rather than the systematic development of learners’ character knowledge. In addition, most teachers prefer to use PPT slides or projection screens to present characters, and seldom demonstrate character structure or decomposition through traditional blackboard writing.

However, both the “character learning within texts” and “text-driven” approaches have been considered unsuitable for Chinese character instruction in CFL contexts (Liu, 2009). Early exposure to complex characters may impose considerable cognitive pressure on beginners. Xu and Yao’s (2014) study found that learners’ integrative motivation significantly decreases under this model, while the “segregated instruction model” (语文分流, *yǔwén fēnliú*) leads to a better affective learning experience. Furthermore, these approaches overlook the semantic motivations and structural principles underlying Chinese characters (Tian, 1979), and they do not provide systematic instruction on radicals, components, or structural patterns. Existing research indicates that understanding the motivated nature of Chinese characters and mastering high-frequency semantic and phonetic components are essential for learners to develop strong form–sound–meaning connections (Fan, 2015). Li and Ye (2013) also found that students receiving segregated instruction showed enhanced strength in linking orthographic, phonological, and semantic information, as well as improved ability to analyze and combine morphemic meanings.

Chapter 2 Chinese Interlanguage Corpora and Pedagogical Orientation

2.1 Interlanguage corpora

2.1.1 Definition of Interlanguage Corpora

From the 1970s onwards, with the emergence of Second Language Acquisition (SLA) as a distinct field, the concept of interlanguage was introduced to theorize learner language as a systematic linguistic system in its own right (e.g. Selinker, 1972). Subsequent studies by researchers such as Fillmore (1976, 1979) and Schumann (1978) drew on learner language data to investigate developmental patterns and variability in second language acquisition. However, sample sizes in terms of participants and datasets were often limited (Gilquin & Granger, 2015), leading to criticism of SLA for its weak empirical foundation, raising doubts about the generalizability of the findings (Gass & Selinker, 2001; Ellis, 2008). The expansion of learner corpora, particularly those containing interlanguage data, has opened up new possibilities for SLA by providing a substantially broader empirical basis than was available in earlier experimental and small-scale observational studies, and has consequently supported a growing body of empirical research on language acquisition (Granger, 2002; Gilquin & Granger, 2015; Wang et al., 2015). A corpus composed of interlanguage data is widely referred to as a learner corpus. However, in this thesis, I adopt the term interlanguage corpus. This choice is motivated by a conceptual difference between the two terms. The notion of a “learner corpus” tends to take the target language as its point of reference and treats learner language primarily in terms of deviation from target norms. In contrast, the term “interlanguage corpus” emphasizes that the data consist of interlanguage, which can be understood as an independent and dynamic linguistic system (Corder, 1975). It highlights the unique characteristics of this system and its developmental changes over time. Although interlanguage exists between the learner’s native language and the target language, and may contain forms that deviate from target language norms, it can nevertheless be regarded as an independent object of linguistic inquiry.

Interlanguage corpora are generally defined as collections of machine-readable authentic

learner texts (including transcripts of spoken data) “which are sampled to be representative of a particular language or language variety” (McEnery et al., 2006, p. 5). Such corpora provide large-scale, systematically organised interlanguage data in a digital format, making it possible to query, extract, and analyse learner production in ways that were previously unattainable (Callies, 2024). Building on this definition, two additional characteristics are frequently highlighted in the literature. The first is systematicity. Nesselhauf (2004) emphasizes that learner corpus data should be selected on the basis of explicit criteria to guarantee representativeness and balance. This methodological requirement places corpus design among the central concerns of learner corpus research (Gilquin, 2015). The second characteristic is (near-)naturalness. As emphasised by Granger (2008), learner data should ideally be produced in relatively natural contexts so that it can reflect learners’ actual linguistic competence, rather than highly controlled or artificially elicited performance.

Since the late twentieth century, the expansion of interlanguage corpora has provided SLA with a substantially broader empirical basis than was previously available, thereby supporting a sustained growth of empirical research on language acquisition (Gilquin, 2015; Wang et al., 2015). In addition to contributing to SLA theory by offering a more accurate description of interlanguage, interlanguage corpora can also be used in practical teaching and learning contexts. The application of interlanguage corpora in the field of foreign language teaching/learning began with Johns (1991). Interlanguage corpora can help language teachers to reduce the gap between classroom teaching content and daily language use, help them understand students’ problems in learning and their own problems in teaching, can provide scientific reference for teachers to develop teaching syllabi (Mukherjee & Pohrbach, 2006; Tono, 2009; Wang & Pan, 2004); interlanguage corpora can also provide valuable language input for learners, help them form the way of inquiry-based learning and autonomous learning (Gilquin & Granger, 2010).

2.1.2 Multimodality in interlanguage corpora

Modality refers to the channels through which communication is conducted. It may denote sensory modalities—such as vision, hearing, and touch—through which humans interact with

the external environment (Allwood, 2008; Huang, 2015), or production modalities, including gestures and vocal output (Allwood, 2008). Multimodal discourse thus involves the use of multiple perceptual or expressive channels in communication. A “multimodal corpus” is a systematical collection that integrates multiple communicative channels and constitutes a computer-based repository of linguistic and interaction-related materials (Foster & Oberlander, 2007). For instance, a corpus that contains both written and spoken data qualifies as multimodal. Such corpora treat communicative events as the primary unit of analysis and aim to extract information relating to language, sound, and image from the raw data (Zhang, 2007; Knight, 2011; Gu, 2013). Typically, multimodal corpora include aligned transcripts and annotations for audio-visual materials to support analyses of lexical patterns, prosody, gesture use, and their interaction in authentic communication (Allwood, 2008; Gilquin & Granger, 2015; Huang, 2015; Knight, 2009).

It is important to distinguish between *multimodal* and *multimedia* corpora. Multimedia focuses on the diversity of media formats (Allwood, 2008). A corpus that, in addition to machine-readable text or transcription, also includes images, audio recordings, and video files qualifies as a multimedia corpus (Knight et al., 2009; Liang, 2023). Such corpora incorporate visual, auditory, and textual resources into linguistic analysis (Wang & Zhang, 2008). By contrast, multimodal corpora emphasise the co-occurrence and interaction of semiotic modalities in meaning-making. As widely noted, while multimedia corpora are characterised by the inclusion of multiple media formats (e.g., audio, video, images), multimodal corpora focus on the presence of multiple semiotic modalities involved in meaning-making. Multimodality does not simply refer to technical media diversity; rather, it concerns the integration of different communicative resources. Therefore, multimodal corpora are often multimedia, but multimedia corpora are not necessarily multimodal. When corpora include audio or video files, aligning transcripts with the corresponding media is a common practice to enable integrated querying and interpretation (Gilquin & Granger, 2015).

Multimodal interlanguage corpora provide valuable support for research on language acquisition and assessment. By examining both written production and spoken production,

studies adopting a multimodal perspective broaden the empirical scope of interlanguage research. Written data primarily provide access to learners' orthographic, lexical, and grammatical choices, while spoken data make it possible to observe not only verbal aspects such as pronunciation and lexical selection, but also non-verbal communicative behaviour, including gesture, facial expression, and body posture (Kotani & Yoshimi, 2023). Studies have shown that such corpora facilitate a deeper understanding of learners' communicative strategies and enhance learners' awareness of their language production, thereby contributing to more effective pedagogy (Franceschi, 2018; Bonsignori & Cappelli, 2022; Ishikawa, 2022). Furthermore, multimodal interlanguage corpora play a significant role in language proficiency assessment (Hu & Xu, 2020; Ishikawa, 2022).

In the context of Chinese interlanguage research, multimodal corpora—especially those integrating multimedia resources such as audio, handwritten images, and video—are of particular importance. Chinese learners must establish links across multiple levels, including stroke and radical processing, initial-final and tone production, and the coordination between orthographic recognition and semantic activation. Single-modality data cannot adequately capture these processes. Multimodal and multimedia-integrated corpora make it possible to document handwriting traces, visual processing, phonological production, textual choices, and cross-modal inconsistencies, thereby revealing more fine-grained cognitive mechanisms involved in Chinese learning (Huang, 2015).

For Chinese language teaching, such corpora provide equally significant benefits. They offer a more accurate representation of learners' multimodal performance in authentic tasks—for instance, the links between unstable orthographic memory and errors in pronunciation. They also help instructors identify learners' weaknesses in integrating form, sound, and meaning, which can inform targeted pedagogical interventions such as enhancing radical awareness, improving tone sensitivity, and strengthening form-sound mapping. In sum, multimodal and multimedia-based Chinese learner corpora not only deepen our understanding of Chinese interlanguage systems but also provide essential empirical evidence for evidence-based pedagogy (Yang et al., 2006).

2.1.3 Transcription and annotation of interlanguage corpora

A prerequisite for the effective analysis of learners' handwritten and oral data is the transcription of raw records into machine-readable textual form (Cameron, 2001). Transcription enables researchers to conduct statistical processing and computer-assisted analysis of linguistic data, and is therefore a fundamental and indispensable component of corpus construction. As Gablasova et al. (2019, p. 134) observe, "transcription quality has direct implications for the reliability and usability of the corpus as a research resource." In other words, the quality of transcription largely determines the reliability of interlanguage data and, by extension, the quality of the corpus as a whole (Liu, 2022).

In general, research on transcription has focused primarily on spoken data. One reason is that, for most alphabetic languages, the "what you see is what you get" nature of written output means that learners' writing errors do not typically lead to characters that fall outside the digital encoding range; even when spelling errors occur, they can simply be typed as they appear. In contrast, Chinese handwritten data pose additional challenges: learners may produce characters with stroke, component, or structural errors that cannot be directly entered into a computer, thus requiring more complex transcription procedures.

For spoken data, researchers may adopt different levels of detail depending on their analytical goals, ranging from basic orthographic transcription to highly detailed phonological or phonetic transcription (Gilquin & Granger, 2015). However, transcription and annotation of speech are generally the most time-consuming and labour-intensive components of corpus construction (Wei et al., 2007). Consequently, due to constraints of time and resources, most spoken learner corpora rely primarily on orthographic transcription, with additional conventions introduced when required by language-specific features, learner proficiency levels, or task design.

Once transcription is completed, annotation becomes essential for enhancing the corpus's searchability and research value. Annotation typically includes two main types: linguistic annotation (such as part-of-speech tagging, morphological and syntactic annotation, or lemmatisation) and error annotation (Pravec, 2002; Díaz-Negrillo et al., 2010; Hana et al.,

2014). Error annotation itself involves two core components: error editing (or correction), which provides one or more hypothesised target forms representing the learner's intended meaning, and error tagging, which applies labels from a predefined classification system to each error (Rosen et al., 2014).

While the methodological foundations for linguistic annotation are well established—supported by extensive work in native-language corpora—error annotation is unique to interlanguage corpora. Given that error analysis has long been a central tradition in interlanguage research, the identification and classification of learner errors have been regarded as key concerns in the development of learner corpora (Díaz-Negrillo et al., 2010). As a result, error annotation is not only inherent to the construction of interlanguage corpora but also forms a central part of their analytical methodology (Dagneaux et al., 1998; Díaz-Negrillo & Domínguez, 2006). Corpora enriched with systematic error annotation are particularly valuable for research purposes (Granger, 2003), although the specific annotation schemes and implementation procedures must typically be tailored to the aims and data types of each individual corpus.

2.1.4 The pedagogical orientation in interlanguage corpora

Learner corpus research is fundamentally connected to the study and practice of language teaching and learning (Mukherjee, 2006). As Götz and Granger (2024) emphasise, interlanguage corpora and language pedagogy share fundamental commonalities: they address the same object of inquiry (interlanguage), ask similar types of research questions, and ultimately pursue overlapping educational goals. In this sense, interlanguage corpora “open up exciting pedagogical perspectives in a wide range of areas of language teaching pedagogy” (Granger, 2003, p. 542). Despite this potential, the field still faces a well-known paradox: pedagogical “implications” remain far more numerous than actual “applications” (Granger, 2015, p. 487). A persistent gap continues to separate interlanguage corpus research from classroom practice (Mendikoetxea et al., 2010), and “there is not a real systematic collaboration” between the two communities (Myles, 2015; 2021).

Interlanguage corpora fulfil at least two major functions. On the one hand, they serve as

research tools for the empirical analysis of interlanguage, an area in which substantial progress has already been made (Mukherjee, 2006; Staples et al., 2024). On the other hand, they also have considerable pedagogical potential. For example, they can inform the design of teaching materials by providing authentic learner data, support the development of assessment tasks based on real learner performance, and offer resources for data-driven learning activities in the classroom (Mendikoetxea et al., 2010; Götz & Granger, 2024). Corpus tools further accelerate research processes (Myles, 2015) and allow SLA specialists to access large amounts of authentic learner data for different research purposes (Mendikoetxea et al., 2010; Iurato, 2022). There have been promising attempts in the first two domains—materials design and assessment (e.g., Barker et al., 2015; Callies & Götz, 2015a, 2015b; Granger, 2015; Dumont, 2017). However, applications in real classroom practice remain relatively limited, largely due to the practical and methodological constraints discussed above, including the complexity of corpus tools and the gap between research outputs and classroom needs (Staples et al., 2024).

This limited classroom impact can be traced to two broad sets of factors. The first concerns the current state of interlanguage corpora as tools. Despite rapid development, learner corpora still suffer from quantitative imbalances: L1 groups are unevenly represented, meaning that corpora “cannot guarantee a bounty of evidence for any research question that might be put to them” (McEnery et al., 2019, p. 80). Written corpora remain more numerous than spoken corpora, and synchronic data overwhelmingly dominate over longitudinal data, leaving many questions without empirical support. Issues of naturalness also arise. Corpora containing written or spoken data inevitably undergo transcription, yet transcription standards are inconsistent, and the degree of analyst intervention varies; the extent to which naturalness is compromised remains unclear. Moreover, multimedia learner corpora are constrained by technological limitations, restricting the fidelity with which original communicative events can be represented.

Annotation presents another major challenge. Most interlanguage corpora still rely heavily on manual annotation, which raises issues of both accuracy and consistency (McEnery et al., 2019). In addition, approaches to annotation vary widely across studies, with no clear

consensus on what should be annotated, how it should be annotated, or to what level of detail (Tono, 2003). This variation leads to incompatibilities across corpora and limits the possibility of large-scale data integration. More importantly, interlanguage corpora tend to function primarily as research instruments rather than as practical classroom tools. Their effective use requires specialised knowledge and technical skills, which creates a substantial barrier for teachers. For instructors with demanding workloads, acquiring corpus literacy represents an additional burden (Leńko-Szymańska, 2017). Furthermore, once applied linguists are no longer directly involved, non-research-active teachers may find it difficult to sustain or replicate corpus-based practices (Chambers, 2019). These constraints help explain why, despite their frequently claimed pedagogical value, the actual implementation of interlanguage corpora in everyday classroom settings remains limited.

The second factor relates to the theoretical development of learner corpus research. The field has often prioritised exploratory and pre-theoretical work, sometimes focusing disproportionately on “basic theoretical constructs” (McEnery et al., 2019, p. 83). As a consequence, learner corpus research is frequently criticised for being overly descriptive rather than theoretically grounded or explanatory (Granger, 2021; Myles, 2021; Iurato, 2022), and in pedagogical contexts—especially classroom instruction—it is often perceived as less robust than psycholinguistic experimentation. The field has not yet provided teachers with a strong, operationalisable pedagogical model that integrates corpus findings into everyday practice. This has led to a persistent gap between the enthusiasm of corpus linguists for pedagogical applications and the reluctance of many teachers to adopt corpus-based approaches in their own classrooms. It also highlights the continuing need for greater teacher involvement in corpus-informed practices and for more accessible ways of translating corpus research into pedagogical use.

Even in the domain most closely connected to corpus linguistics—Data-Driven Learning (DDL)—many studies rely on small sample sizes and short-term interventions (Boulton & Vyatkina, 2021; Crosthwaite & Boulton, 2022). Consequently, the effectiveness of DDL has not been consistently validated in real classrooms (Pérez-Paredes, 2022; O’Keeffe, 2020), and

its benefits often remain intuitive rather than empirically demonstrated. The absence of longitudinal classroom-based research, together with the lack of pedagogically oriented theorisation from corpus linguists, continues to limit the uptake of interlanguage corpora in teaching. As Hunston (2002) notes, teachers must retain a sense of agency and control in their classrooms; without frameworks that respect this need, sustained interaction between corpus research and pedagogy is difficult to achieve.

2.3 An overview of Chinese interlanguage corpora

This section provides an overview of major Chinese interlanguage corpora. It first offers brief descriptions of representative corpora, outlining their design features, data types, and research focus. This is followed by a comparative summary and critical evaluation, highlighting shared strengths, limitations, and unresolved issues in existing corpus construction practices.

2.3.1 Chinese interlanguage corpora

Since the 1990s, the development of Chinese interlanguage corpora has gradually taken shape (Zhang, 2010; Xu, 2019), and these corpora have significantly advanced research in the field of teaching Chinese as a foreign language (TCFL). However, like all emerging fields, this area remains at a critical developmental stage and continues to face a range of challenges. Many existing corpora exhibit various limitations (Iurato, 2022a), and they do not yet fully satisfy the needs of Chinese language teaching and relevant research. Moreover, theoretical discussions on corpus construction remain underdeveloped, with numerous unresolved methodological and conceptual issues (Zhang & Cui, 2013). Although new corpora continue to be created, the field still lacks systematic evaluation.

This subsection surveys a selection of representative Chinese interlanguage corpora that reflect the current state of corpus construction in this field. A total of seventeen corpora are reviewed, including widely cited resources as well as more recent or specialized datasets, covering both written and spoken learner data and a range of design orientations.

To ensure clarity and avoid unnecessary repetition, this subsection adopts a two-step organizational strategy. First, each corpus is briefly introduced in turn, with concise

descriptions focusing on its institutional background, data type, processing approach, and primary research orientation. These descriptions are intended to provide readers with a basic understanding of what each corpus is and how it is positioned, rather than to offer detailed comparisons. Second, a systematic comparison of the reviewed corpora is presented in Section 2.3.2 through a set of summary tables. These tables reorganize key design features, data characteristics, and processing practices into a unified analytical framework, allowing similarities and differences across corpora to be examined more transparently.

(1) HSK Dynamic Essay Corpus [HSK 动态作文语料库]⁶

The HSK Dynamic Essay Corpus (HSKDEC), developed by Beijing Language and Culture University, is one of the earliest large-scale Chinese interlanguage corpora. It consists of handwritten compositions produced in HSK examinations by learners from diverse L1 backgrounds and proficiency levels. Both scanned images and transcribed texts are provided, with learner errors preserved and manually corrected, while annotation is limited to error tagging. As a test-based writing corpus, HSKDEC has played a foundational role in written Chinese interlanguage research.

(2) Global Chinese Interlanguage Corpus [全球汉语中介语语料库]⁷

The Global Chinese Interlanguage Corpus (GCIC), led by Beijing Language and Culture University, is a large-scale, publicly accessible learner corpus collecting data from international students studying in China. It integrates written, spoken, and video-based learner data, presented in transcribed textual form with error annotation and multi-level linguistic annotation. Learner output is segmented at the sentence level rather than preserved as full texts, reflecting a balance between research and pedagogical usability (Cui & Zhang, 2011; Zhang, 2019).

(3) Yet Another Chinese Learner Corpus⁸

Yet Another Chinese Learner Corpus (YACLC) is a text-based learner corpus constructed from online learner compositions. Its distinctive feature lies in its annotation methodology: each sentence is evaluated and revised by multiple trained annotators, with a focus on fluency

⁶ <http://hsk.blcu.edu.cn/>

⁷ <http://qqk.blcu.edu.cn/#/search/survey>

⁸ <https://github.com/blcuicall/YACLC>

judgment and acceptability. The corpus prioritizes post-edited output, making it more suitable for fluency assessment than for the analysis of raw interlanguage forms (Wang et al, 2021).

(4) Chinese Continuity Mediators for Chinese Character Error Annotation Corpus [汉字偏误标注的汉语连续性中介语语料库]⁹

The Chinese Continuity Mediators for Chinese Character Error Annotation Corpus (CCMCCEAC), developed by Sun Yat-sen University, focuses on written Chinese interlanguage data produced by international students. It combines scanned handwritten samples with transcribed texts, preserving learner errors and providing error annotation. The corpus is primarily oriented toward character- and word-level error analysis, with limited annotation beyond error tagging (Zhang, 2012).

(5) The Jinan Chinese Learner Corpus [暨南大学学习者语料库]¹⁰

The Jinan Chinese Learner Corpus (JCLC), constructed by Jinan University, consists of written texts collected from foreign students studying in Chinese universities. Learner texts are manually transcribed with original errors preserved, but no systematic error annotation is available. JCLC is not publicly available at the time of writing, from the published information I can only know that it does not contain error annotation at this stage (Wang et al, 2015). It is worth mentioning that JCLC also includes other languages that participants have mastered/are learning in their metadata.

(6) The L2 Component of the Spoken Chinese Corpus¹¹

The L2 Component of the Spoken Chinese Corpus (LCSCC) contains transcribed spoken interaction data produced by L2 learners of Chinese. The corpus adopts corrected transliteration, presenting target-like forms rather than original learner errors, and does not include annotation. The learners are L2 New Zealanders and 2 Australians, their first language is English, and their Chinese level is at an intermediate level (Li, 2021).

(7) The Multilingual Corpus of Second Language Speech¹²

The Multilingual Corpus of Second Language Speech (MuSSeL) is a developing multilingual

⁹ <https://cilc.sysu.edu.cn>

¹⁰ Currently not open to the public.

¹¹ <https://github.com/blculyn/The-spoken-L2-corpus>

¹² <https://l2trec.utah.edu/learner-corpora/mussel/>

learner speech corpus that includes Chinese among several target languages. It contains audio recordings and corresponding transcriptions collected through standardized oral proficiency assessments (Schnur et al, 2019). Most of its sample providers are native speakers of English. There is no detailed explanation about the transcription method and annotation method of its data.

(8) Nanjing University Chinese Interlanguage Spoken Corpus (Version 1.0 beta3)[南京大学汉语中介语口语语料库（1.0 beta3 版）]¹³

The Nanjing University Chinese Interlanguage Spoken Corpus (NUCISC) consists of spoken interaction data collected through longitudinal recordings of foreign students. The corpus provides unmodified transcriptions that preserve learner errors but does not include annotation. This corpus has not published detailed database information on the website, and no published documents can be obtained. The relevant information comes from the slide *Report on the Joint Construction of Confucius Institute between Nanjing University and Foreign University* [南京大学与国外大学共建孔子学院情况汇报] by Professor Aimin Cheng of Nanjing University).

(9) Guangwai-Lancaster Chinese Learner Corpus¹⁴

The Guangwai–Lancaster Chinese Learner Corpus (GLCLC), jointly developed by Guangdong University of Foreign Studies and Lancaster University, contains both written and spoken learner data accessible through Sketch Engine¹⁵. The corpus presents corrected and transliterated texts without preserving original learner errors and does not include annotation. Its design prioritizes comparability and accessibility over interlanguage form analysis (Xu et al, 2019).

(10) TOCFL learner corpus (NTU)¹⁶

The TOCFL learner corpus developed by National Taiwan University consists of written texts collected through computer-based proficiency tests. Multiple processed versions of learner output are provided, but original error forms are not systematically preserved. However, all the data is in traditional Chinese characters, and they were all collected in Taiwan. The Chinese

¹³ <http://yuliaoku.hanyu123.cn/nju?name=&c=&hsk=>

¹⁴ <https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fguangwai>

¹⁵ <https://www.sketchengine.eu/>

¹⁶ <http://tocfl.itc.ntnu.edu.tw:8080/>

used in Taiwan is still different from the Mandarin used in mainland China and widely taught in teaching TCFL in terms of vocabulary and other aspects (Zhang, 2013).

(11) TOCFL learner corpus (NCU) ¹⁷

The TOCFL learner corpus constructed by National Central University also contains computer-based written test data in traditional Chinese characters. Unlike the NTU version, it includes multi-level grammatical error annotation. Its regional focus and reliance on examination data limit its generalizability to other instructional contexts. However, like NTU, all the data is in traditional Chinese, and the collection site is in Taiwan. As mentioned earlier, there are certain differences between the Chinese used in Taiwan and mainland China (Lee et al, 2018).

(12) Chinese as a Second Language Oral Corpus [華語為第二語口語語料庫]¹⁸

The Chinese as a Second Language Oral Corpus (CSLOC), developed by National Taiwan Normal University, consists of spoken data collected through standardized oral proficiency tests. It provides audio recordings and corresponding transcriptions that preserve learner output without correction or annotation. The corpus primarily represents advanced learners but since the corpus is collected in Taiwan, the learners all have a strong Taiwanese accent, and the transliteration is also carried out in traditional Chinese characters.

(13) The COPA corpus¹⁹

The COPA corpus is a small spoken learner corpus consisting of dialogue recordings and uncorrected transcriptions produced by university students learning Chinese. While it preserves natural learner speech, it lacks annotation and search functionality, at present users can only browse one by one, it does not support retrieval (Zhang, 2009).

(14) The HKPU corpus²⁰

The HKPU corpus is a small multimodal corpus combining video recordings with transcribed learner speech. Data were collected through structured narrative retelling tasks, with no error correction or annotation applied. The corpus is primarily suited for qualitative analyses of

¹⁷ <https://l1p.ee.ncu.edu.tw/resource/tocfl.html>

¹⁸ <http://140.122.83.243/mp3c/>

¹⁹ <https://sla.talkbank.org/TBB/slabank/Mandarin/COPA>

²⁰ <https://sla.talkbank.org/TBB/slabank/Mandarin/HKPU>

learner speech production. This corpus is included in SLABank²¹ together with the COPA corpus.

(15) Chinese Character Writing Errors for Second Language Learners Corpus [二语学习者汉字书写错误语料库]²²

The Chinese Character Writing Errors for Second Language Learners Corpus focuses on handwritten Chinese character data collected through dictation tasks. This corpus is more like a small data set. Because of the Chinese character input method (non-existent typos cannot be typed directly), the corpus only contains picture samples, and the sample size is very limited.

(16) Bimodal Italian Learner Corpus of Chinese²³

The Bimodal Italian Learner Corpus of Chinese (BILCC) was developed to investigate Italian learners' acquisition of the Chinese cleft construction “是...的 (*shì...de*)”. It integrates written and spoken learner data alongside native-speaker control data. The corpus is relatively focused in scope and is designed to address a specific acquisitional research objective, namely the development and use of the “是...的” construction by Italian learners of Chinese. The corpus is accompanied by rich metadata, and the annotation methods include error annotation and pragmatic annotation, especially for learners' inappropriate use of the structure “是...的 (*shì...de*)” (Iurato, 2023, 2024).

(17) Chinese Interlanguage Speech Multimodal Corpus for Intelligent Speech Teaching [面向智能语音教学的汉语中介语语音多模态语料库]²⁴

The Chinese Interlanguage Speech Multimodal Corpus for Intelligent Speech Teaching (CISMCIIST) comprises several sub-corpora related to learner speech and pronunciation training. Although labeled “multimodal,” most components consist primarily of audio data, with limited natural learner output. Its primary orientation is toward speech technology and instructional applications rather than interlanguage development analysis. This corpus is not publicly available at this time.

²¹ <https://slabank.talkbank.org/>

²² <https://ccw.eduhk.hk/>

²³ Currently not open to the public, so no related links

²⁴ Currently not open to the public, so no related links

2.3.2 Comparative summary of the reviewed corpora

The evaluation and comparison of existing Chinese interlanguage corpora can be conducted from three complementary perspectives. The first one is corpus design (see Table 2.1), which examines the overall configuration of each corpus, including its modality, longitudinal or cross-sectional nature, accessibility, and search interface; the second is corpus data (see Table 2.2), which summarizes the types of learner data (written or spoken), corpus size, task topics, and information on learners' backgrounds such as native language and proficiency level; and, finally, corpus processing and annotation (see Table 2.3), which focuses on transcription practices—mainly for spoken data—as well as annotation methods and standards. The following section provides a comparative summary of the seventeen reviewed corpora based on these three analytical dimensions.

Table 2.1

Design approach in the 17 analysed corpora

Corpus	Media	Cross-sectional/longitudinal	Availability	Retrieval method
HSK DEC	multimedia	Cross-sectional	Open	Online
CIC	unimedia	Cross-sectional	Open	Online
YACLC	unimedia	Cross-sectional	Open	Available for browsing
CCMCCEA C	multimedia	Cross-sectional	Open	Online
JCLC	unimedia	Cross-sectional	No	-
LCSCC	unimedia	Cross-sectional	Open	Available for browsing
MuSSeL	multimedia	Cross-sectional	No	-
NUCISC	unimedia	Longitudinal	Open	Online
GWLCCLC	unimedia	Cross-sectional	Open	By using Sketch Engine
TOCFL (NTU)	unimedia	Cross-sectional	Open	Online
TOCFL	unimedia	Cross-sectional	Open	Available for

(NCU)				browsing
CSLOC	multimedia	Cross-sectional	Open	Online
COPA	multimedia	Cross-sectional	Open	Available for browsing
HKPU	multimedia	Cross-sectional	Open	Available for browsing
CCWESLL C	unimedia	Cross-sectional	Open	Available for browsing
BILCC	multimedia	Cross-sectional	No	-
CISMCIST	multimedia	one sub-corpus is longitudinal	No	-

The comparison of corpus design features in Table 2.1 reveals a structural imbalance in current Chinese interlanguage corpus construction when multimodality and multimedia are distinguished as separate dimensions. From a linguistic perspective, most corpora remain unimodal, focusing exclusively on either written or spoken learner production. Corpora that integrate both written and spoken data—that is, genuinely multimodal corpora—are relatively rare.

At the technical level, however, the use of multimedia formats is more widespread. Several corpora incorporate audio recordings, video materials, or scanned handwritten images alongside textual representations. Yet in most cases, these multimedia resources serve as alternative representations of a single modality rather than as parallel manifestations of different types of learner production. As a result, the presence of multimedia data does not necessarily entail multimodal linguistic coverage.

Another prominent design feature is the dominance of cross-sectional data collection. Only a very small number of corpora incorporate longitudinal components, while the majority provide static snapshots of learner performance. This tendency limits the extent to which existing corpora can support analyses of interlanguage development over time, particularly across different modalities.

Finally, although many corpora are nominally open to the public, access conditions and retrieval functionality vary substantially. Fully searchable online interfaces remain limited, and

several corpora are accessible only through browsing or download, which constrains their usability for systematic research and pedagogically oriented applications.

Table 2.2

Data types in the 17 analysed corpora

Corpus	Modality	Size²⁵	Data source	Learner's L1	Learner's proficiency level
HSK DEC	written	4.42 million words	test composition	diverse	diverse
CIC	multimodal	23.67 million words	daily homework and exam text	diverse	diverse
YACL C	written	32,124 sentences	computer-based composition	diverse	diverse
CCMC CEAC	written	5 million tokens	daily composition and classroom writing tasks	diverse	diverse
JCLC	written	6 million tokens	exam texts and daily homework	diverse	diverse
LCSCC	spoken	220792 words	L1-L2 conversation	English	intermediate
MuSSe L	spoken	111267 words	oral interview test	English	diverse
CCNUIC	written	3 million words	compositions	no information	no information
NUCIS	spoken	2 million words	natural	diverse	diverse

²⁵ original data size, without annotation

C			conversion		
GWLC LC	multimodal	1.2 million tokens	examinations and tutorials	diverse	diverse
TOCFL (NTU)	written	1,5 million tokens	computer- based test	diverse	diverse
TOCFL (NCU)	written	1 million characters	computer- based test	diverse	diverse
CSLO C	spoken	773000 characters	oral proficiency test	English, Japanese and Korean	advanced
COPA	spoken	no information	dialogue recording	diverse	diverse
HKPU	spoken	no information	story retelling	English	diverse
CCWE SLLC	written	no information	dictation test	Vietnames e	diverse
BILCC	multimodal	57600 characters	open-ended tasks	Italian	diverse
CISMC IST	spoken	no information	no information	diverse	diverse

Table 2.2 highlights several general tendencies in the composition of learner data across the reviewed corpora. With respect to modality, written-only and spoken-only corpora are equally represented, whereas corpora integrating both written and spoken learner data remain relatively limited. This separation of modalities constrains the extent to which relationships between different aspects of learner performance can be examined.

Regarding data sources, examination-based data constitute a substantial proportion of existing corpora, particularly in writing datasets. Such data are produced under time-constrained and evaluative conditions, which may shape learner language production in both writing and speaking. While these settings ensure standardization and comparability, they differ from language use in more relaxed instructional or communicative contexts. Interlanguage data collected outside testing conditions therefore remain relatively limited.

Learner coverage also shows an uneven distribution. Most corpora include learners from multiple L1 backgrounds and proficiency levels, reflecting an orientation toward broad representativeness. However, corpora focusing on a single L1 group are relatively rare, this design tendency limits the scope for detailed, L1-specific analyses of interlanguage development.

Table 2.2

Data processing in the 17 analysed corpora

Corpus	Transcription	Annotation
HSK DEC	preserves errors	error annotation
CIC	preserves errors	language information annotation+error annotation
YACLC	no transcription ²⁶	Fluency annotation+error annotation
CCMCCEAC	preserves errors	error annotation
JCLC	no information	only know that there is no error annotation
LCSCC	corrected transliteration	No annotation
MuSSeL	unknown	unknown
CCNUIC	unknown	No annotation
NUCISC	preserves errors	No annotation
GWLCLC	corrected transliteration	No annotation
TOCFL (NTU)	no transcription ²⁷	No annotation (but with correction)
TOCFL (NCU)	no transcription ²⁸	error annotation
CSLOC	preserves errors	No annotation
COPA	preserves errors	No annotation
HKPU	corrected transliteration	No information
CCWESLLC	no transcritpion	No annotation
BILCC	preserves errors	Pragmatic annotation + error annotation+ syntactic annotation
CISMCIIST	no transcription	phonetic annotation

²⁶ data is machine-readable text

²⁷ data is machine-readable text

²⁸ data is machine-readable text

The comparison of transcription and annotation practices in Table 2.3 reveals substantial methodological diversity across Chinese interlanguage corpora. While several corpora preserve learner errors in transcription, others adopt corrected or target-like representations, and a number of corpora do not provide transcription at all. This lack of consistency complicates cross-corpus comparison and limits the reusability of learner data for interlanguage-focused analysis.

Annotation practices show an even greater degree of variation. Error annotation is the most commonly implemented annotation type, but it is far from universal, and only a small subset of corpora provide additional layers of linguistic or pragmatic annotation. Many corpora remain entirely unannotated or provide insufficient documentation of their annotation standards.

2.3.3 An evaluative look on existing Chinese interlanguage corpora

Although existing Chinese interlanguage corpora have substantially expanded the empirical base for TCFL and SLA research, a synthesis of their design choices and documentation practices reveals several recurring limitations that restrict broader research comparability and pedagogical usability. This section synthesizes these common issues to identify directions for improvement.

Media limitations and the need for richer data representation

Most existing Chinese interlanguage corpora rely primarily on text-based representations of learner output. While textual data constitute an indispensable foundation for corpus-based analysis, unimedia corpora tend to decontextualize learner language and cannot fully reflect the complexity of language learning processes (Chen & Liang, 2010). This limitation is particularly salient in Chinese language learning.

As discussed in Chapter 1, learners whose first language is alphabetic typically encounter major difficulties related to character complexity, tonal contrasts, and grammatical structures that differ substantially from those of their native language (Wei, 2018). From this perspective, corpora that provide only electronic text offer an incomplete representation of learners' developing competence. Ideally, Chinese interlanguage corpora should integrate picture data

that document handwritten performance as well as audio or video data that capture reading and spoken production, thereby presenting a more comprehensive picture of the learning process. The construction and analysis of multimodal Chinese interlanguage corpora therefore constitute not only a useful extension of corpus linguistics, but also an important contribution to research on Chinese teaching and acquisition. Direct observation of multimodal learner data can help identify modality-specific factors affecting Chinese acquisition and provide empirical support for multimedia teaching practices (Huang, 2015).

Temporal limitations: The scarcity of longitudinal data

A further limitation concerns the scarcity of corpora that track learner development over time. As noted in previous research, longitudinal interlanguage corpus construction is time-consuming and logistically demanding, and meaningful analysis can only begin after extended data collection has been completed (Meunier, 2015). Additional challenges include participant attrition, as learners may discontinue participation before the end of the collection period (Meunier, 2021). These factors partly explain why longitudinal designs remain relatively rare.

From the perspective of interlanguage theory, longitudinal data are essential. Interlanguage is inherently dynamic and continuously reorganized as learners progress, rather than a static system observed at isolated points in time. Without longitudinal evidence, it is difficult to observe developmental trajectories or to distinguish temporary fluctuations from stable change. From an applied perspective, longitudinal corpora also allow researchers and teachers to identify stage-specific patterns and to align teaching progression more closely with learners' actual developmental paths.

Limited L1-specific corpora and pedagogical relevance

Another common issue is the limited availability of L1-specific Chinese interlanguage corpora. While multi-L1 corpora provide broad coverage and support cross-linguistic comparison, they offer a less reliable basis for identifying difficulties characteristic of particular learner groups. L1-specific corpora can support the design of country-specific syllabi, teaching materials, and learner dictionaries by providing stable, multidimensional baseline data (Gan & Zhang, 2005). They also make it possible to develop annotation norms that are tailored to the recurrent error

patterns of a specific learner population, thereby avoiding the risk that pedagogically meaningful individual differences are obscured by overly general coding schemes (Hu, 2018).

As Nesselhauf (2004) emphasizes, it is important to understand not only what native speakers typically produce, but also the typical difficulties experienced by particular learner groups. For this reason, native or in-house learner corpora have been strongly recommended as tools for addressing the specific language problems of defined learner populations (Mukherjee & Rohrbach, 2006; Seidlhofer, 2002).

Learner proficiency classification and data distribution

Current Chinese interlanguage corpora also vary considerably in how learner proficiency is defined and operationalized. Some corpora adopt external frameworks such as the Common European Framework of Reference (CEFR). Others rely on institutional criteria such as years of study, which can only serve as a rough proxy for proficiency and may obscure substantial individual differences. Although official proficiency tests such as the HSK offer a more standardized basis for classification, they are not universally available or compulsory, particularly for learners studying outside China.

The lack of consistent proficiency criteria makes it difficult to align learner groups across corpora or to interpret observed differences as developmental rather than methodological. Moreover, existing corpora tend to contain a disproportionate amount of data from intermediate and advanced learners. Examination-based corpora, for example, often include only advanced-level writing samples, while learners who can produce coherent written texts in Chinese have typically already reached at least an intermediate level (Zhang, 2008; Ren, 2010). In addition, many corpora are constructed in China and draw on international students studying at Chinese universities, whose exposure to the language and learning environment differs substantially from that of learners in non-Chinese-speaking contexts. Greater attention to beginner learners and to learners studying outside China is therefore necessary to achieve a more balanced representation of Chinese interlanguage development.

Task design, data sources, and authenticity

Biases related to task design and data sources further constrain the interpretability of existing

corpora. In many cases, learner data are derived partly or exclusively from examinations. Although examination data represent genuine learner output, they are produced under time constraints and specific evaluative conditions, which may influence learners' language use (Yang & Hui, 2002). In addition, examination materials are often subject to restrictions on publication in order to preserve the integrity of teaching and assessment practices (Zhang, 2010). At the same time, in the current context of widespread access to machine translation and AI-assisted writing tools, examination settings remain one of the few contexts in which learner production can be reasonably assumed to be independent and minimally mediated. This creates a methodological trade-off: while examination data may lack naturalness, they offer a higher degree of control over the authenticity of learner output compared to data collected in less controlled environments.

Other corpora rely on computer-based input, where learners may use pinyin input methods or external aids, including dictionaries, machine translation, or AI-assisted writing tools, making it difficult to assess their actual character-writing ability. Task designs that restrict learners to a single topic, such as story retelling, may also lead to lexical and structural concentration, thereby limiting the range of observable language use. Homework-based data pose similar concerns, as learners may consult external resources or receive assistance. While such data are not inherently invalid, they require careful documentation and supervision to ensure that learner output reflects independent production rather than assisted performance. In practice, this can be achieved through a combination of measures, such as clearly specifying task conditions, limiting or monitoring access to external tools during data collection, recording the context and mode of production (e.g., timed vs. untimed, supervised vs. unsupervised), and, where possible, triangulating data from multiple tasks or modalities. At the same time, excessive control or pressure during data collection should be avoided so as not to distort learners' actual proficiency.

Inconsistencies in annotation and metadata practices

Annotation practices across Chinese interlanguage corpora remain highly uneven. A large proportion of corpora provide unannotated raw data, while others restrict annotation to error

tagging without incorporating broader linguistic information. Only a small number of corpora combine error annotation with more comprehensive linguistic or pragmatic annotation. As previous research has argued, focusing exclusively on errors risks overlooking learners' correct and developing language use. Extending analysis beyond error identification toward performance-oriented or language-use analysis makes it possible to trace learners' developmental trajectories more systematically (Sun, 1993; Zhang, 2010).

Metadata practices present additional challenges. Rich metadata are essential for interpreting interlanguage data, given the wide range of variables involved in second language acquisition (Myles, 2021; Callies, 2023). They also support comparability across studies and facilitate cumulative research (Paquot et al., 2024). According to the metadata framework proposed by Grange and Paquot (2017), interlanguage corpora should document not only learner background information but also corpus design, task conditions, annotation procedures, and transcription practices. However, many existing Chinese interlanguage corpora provide only limited learner metadata and omit other crucial categories. In some cases, insufficient anonymization of learner data further restricts corpus accessibility and reuse.

In sum, existing Chinese interlanguage corpora have laid an important empirical foundation for research and teaching. At the same time, shared limitations—such as reliance on text-only data, limited multimodal and longitudinal coverage, inconsistent proficiency classification, task-related biases, and uneven annotation and metadata practices—continue to constrain comparability and pedagogical applicability. Addressing these issues would enhance the interpretability, reusability, and educational value of future Chinese interlanguage corpora.

2.4 Data processing of Chinese interlanguage corpora

2.4.1 Handwritten data processing in existing interlanguage corpora

According to the overview of 17 Chinese Interlanguage Corpora (CIC) presented in Chapter 1, 10 corpora contain written or written/spoken data. Four of them (BILCC, YACLC, TOCFL-NTU, TOCFL-NCU) specify that their writing data are typed text obtained from examinations or standardized tasks and thus do not involve handwritten processing. JCLC contains

handwritten data; however, as the corpus is not currently publicly accessible, the available information is limited to a related study (Wang et al., 2015), which reports that all data were manually transcribed and that all errors were retained. The present section therefore focuses on five corpora that explicitly include handwritten data, analysing how they handle transcription, annotation, and correction, both in content and in formal representation.

CCICCCEA

Constructed by a team at Sun Yat-sen University, this corpus's writing data is sourced mainly from daily compositions and in-class writing tasks of international students at the university's School of International Chinese Studies and the Chinese Department. The core data includes two modalities: scanned images of the handwritten text and the corresponding transcribed text.

Transcription Strategy:

The three sub-corpora adopt a consistent approach to preserve learners' errors as much as possible. The transcription strategy for incorrectly written characters evolved in two phases²⁹: Before 2016 - A character creation method was used, where researchers manually traced the exact form of miswritten characters using TrueType software, embedding them as images into the transcribed text. After 2016: The approach shifted to directly clipping image segments of the erroneous characters from the scanned images and embedding them. This simplified the procedure while maximally retaining the learner's authentic handwriting.

Annotation System:

CCICCCEA features annotation schemes at the character, lexical, syntactic, discourse, and punctuation levels. The first two sub-corpora mainly focus on single-character errors, while the later one expanded to include lexical, grammatical, sentence pattern, discourse, and punctuation errors, also providing word segmentation and part-of-speech (POS) labeling at the syntactic level (though the specific algorithm is not disclosed, only the label set). In single-character annotation, the corpus expanded beyond incorrectly written character and incorrectly used character to include two intermediate types: 似别字(*sìbiézi*, "similar but mistaken character") and 不规范字(*bùguīfànzi*, "non-standard character") refers to a character that exists

²⁹ see: <https://cilc.sysu.edu.cn/static/manual.pdf>

in the system but is written mistakenly, sits between a correct character and a miswritten one, showing non-standard stroke or component proportions but remaining discernible.

Correction Presentation:

Correction information is partially provided, such as the correct character corresponding to incorrectly written or the missing target word. However, its presentation form is complex, see example 1:

(1)

今天/t <是>QS 我/r 的/u 生日/n
jīntiān/t <shì>QS wǒ/r de/u shēngrì/n

Source: CCICCCEA

The correct, supplemented word “是 (*shì*)” is enclosed in angle brackets, followed by the annotation code “QS”. While this embedded structure is informationally complete, it significantly increases the difficulty of both reading and automated parsing.

CIC-BLCU

Led by Professor Zhang Baolin at Beijing Language and Culture University (BLCU), this corpus includes written and spoken data from various institutions. The writing data primarily consists of examination essays and classroom writing tasks. As a unimodal corpus, CIC-BLCU provides only the text transcription results without retaining the original handwritten scans.

Transcription Strategy:

The corpus adheres to the principle of “preserving errors” but often presents only the corrected form, without simultaneously providing the original erroneous form. The text is entered sentence by sentence to minimize interference with pedagogical application.

Annotation System:

CIC-BLCU’s annotation system spans ten levels, including character, lexicon, phrase, syntactic component, discourse, style, rhetoric, punctuation, phonetics, and visual body language. The annotation uses an inline embedded structure, where errors, comments, and corrected text are mixed within the same document, see example 2 and 3.

(2)

我最不喜欢广告就是在看电视或者电影的时候，广告【是】[Cd]【打乱】 [Cb 打断] 剧情，浪费时间。

Wǒ zuì bù xǐhuān guǎnggào jiù shì zài kàn diànshì huòzhě diànyǐng de shíhou, guǎnggào 【shì】 [Cd] 【dǎluàn】 [Cb dǎduàn] jùqíng, làngfèi shíjiān.

Source: CIC-BLCU

Here, “是 (shì)” in brackets should be deleted; “打乱 (dǎluàn)” is the corrected form, and “打断 (dǎduàn)” was the original incorrect form.

(3)

我们在电视【】[Cq 上]看广告希望看一个最好的电视剧，看最好的故事。

Wǒmen zài diànshì 【】 [Cq shàng] kàn guǎnggào xīwàng kàn yīgè zuì hǎo de diànshìjù, kàn zuì hǎo de gùshi.

Source: CIC-BLCU

The empty brackets “【】” indicate a missing word in the original handwritten text, which should be inserted as “上” (in the content following the brackets). This approach ensures the text’s completeness and computational tractability. However, the mixed structure reduces readability and intuitiveness, requiring the reader to be highly familiar with the marking scheme. It is difficult for non-expert users to distinguish the original text from the corrections.

GLCLC

Jointly constructed by Guangdong University of Foreign Studies and Lancaster University, this corpus is searchable via Sketch Engine. Written samples come from examinations or routine writing tasks.

Transcription Strategy:

The principle is faithful recording of the learner’s original output. Sections with miswritten characters, or substitutions using Pinyin or foreign words, are fully preserved. If a character cannot be typed due to stroke errors, it is replaced by the correct character with an attached label.

(4)

有人爱<orth>看</orth>电影

Yǒurén ài <orth>kàn</orth> diànyǐng

Source: GLCLC

“看 (kàn)” is the correct character, and “<orth>...</orth>” is the error correction marker. However, the corpus’s definition of “erroneous characters” is vague, described only generally as “characters with too many or too few strokes,” without further considering finer-grained errors like structural faults or component misplacement. Furthermore, actual searches on Sketch Engine do not reveal instances of error annotation, suggesting potential inconsistency between documentation and practical implementation. While GLCLC prioritizes fidelity and structured presentation in design, the lack of implementation transparency makes it difficult for users to determine if errors are completely recorded.

HSK DEC

Organized by Cui Xiliang’s team at BLCU, the corpus uses HSK examination essays as its source. Its distinguishing feature is its multimodal integration, providing handwritten scanned images synchronously with the transcribed and annotated text.

Transcription Strategy:

HSK DEC strives for a faithful record of the original text, including traditional characters, variant characters, and punctuation errors, while attaching the correct form when the original cannot be typed. Nevertheless, the mixing of transcribed, annotated, and corrected text makes both reading and processing challenging, see example 5 and 6.

(5)

通过多聊天能力互相了解, 保证双方的自由{CD 的话}, 所有的代沟问题{CJ-zy 都}会逐渐消失。

Tōngguò duō liáotiān nénglì hùxiāng liǎojiě, bǎozhèng shuāngfāng de zìyóu {CD de huà}, suǒyǒu de dàigōu wèntí {CJ-zy dōu} huì zhújiàn xiāoshī.

Source: HSKDEC

Here, “的话” in the curly braces is the content intended for deletion.

(6)

要解决代沟{CQ 问题}的话就和父母或老人家找共同[B 通]话题说几句[BQ,]{CJ-zhuy
代沟}会消失。

*Yào jiějué dàigōu {CQ wèntí} de huà jiù hé fùmǔ huò lǎorénjiā zhǎo gòngtóng [B tōng] huàtí
shuō jǐ jù {BQ ,} { CJ-zhuy dàigōu}huì xiāoshī.*

Source: HSKDEC

Here, “问题 (wèntí)” in curly braces is the missing part added to the text. This multi-layered nested annotation structure is rich in information but visually burdensome. As noted by Ren Haibo (2010), if the symbol system is unfamiliar, the extracted text is difficult to restore to the original or to conform to standard grammar, complicating automated analysis.

Overall, existing corpora demonstrate diverse approaches to the transcription, annotation, and correction of handwritten data, reflecting different methodological priorities in corpus construction. However, a comparative analysis reveals several areas where improvements are needed, particularly when evaluated against established principles of corpus design, such as transparency, consistency, usability, and extensibility (Leech, 1993; McEnery & Hardie, 2011).

- Lack of publicly available and operationalised data processing guidelines. A number of existing corpora provide only limited documentation of their transcription and annotation procedures, sometimes offering only label inventories without clear procedural guidelines. This lack of explicit documentation reduces transparency and makes it difficult to assess the reliability and reproducibility of the data. In corpus linguistics, well-defined annotation protocols are considered essential for ensuring consistency both across annotators and within the same annotator over time (McEnery & Hardie, 2011). Furthermore, discrepancies between documented methods and actual implementation (e.g., GLCLC) raise additional concerns regarding methodological reliability. This issue is particularly critical for handwritten learner corpora, where transcription constitutes the primary representation of the original data. In unimodal corpora lacking access to original images, users must rely entirely on the transcription, making accuracy at this stage crucial for all subsequent analysis.

- Redundancy in transcription representation. Some corpora (e.g., CCICCCEA) adopt strategies that embed image fragments of erroneous characters directly within the transcribed text in order to preserve visual authenticity. While this approach aligns with the principle of fidelity to the original data, it also introduces redundancy when full scans of the handwritten texts are already available. From the perspective of corpus usability, such redundancy may increase cognitive load for users and complicate downstream processes such as annotation, searching, and statistical analysis (McEnery & Hardie, 2012). Effective corpus design requires balancing authenticity with clarity and efficiency, avoiding both excessive embedding and over-simplification.
- Limitations in readability and extensibility of text structure. As observed in corpora such as CIC-BLCU and HSK DEC, the integration of transcription, annotation, and correction within a single textual layer can reduce readability and hinder interpretability. The use of densely embedded symbolic systems, while informationally rich, may increase processing difficulty for users and reduce accessibility for non-expert researchers. From a design perspective, such practices conflict with the principle of layer separation, whereby primary data, annotation, and correction are maintained in independent tiers to enhance clarity, interoperability, and extensibility (Ide & Romary, 2004). Moreover, mixed-layer representations complicate the addition of new annotation levels (e.g., phonological or multimodal layers) and increase the complexity of computational parsing. A multi-layered architecture, in which different types of information are stored separately but remain aligned, provides a more scalable and flexible solution for corpus development..

2.4.2 Spoken data processing in existing interlanguage corpora

Audio data within an interlanguage corpus can generally be categorized into two types based on the task: reading-aloud data and spontaneous speech data. Currently, there is limited research investigating the distinct characteristics of these two types of data in the context of interlanguage. However, studies focusing on native speakers have revealed that different task types lead to the presentation of distinct linguistic features in the resulting data (Liu & Li, 2002). Therefore, the choice of audio type should be made strategically according to the specific goals

of the corpus construction: Spoken data tends to reveal richer communicative features, while reading-aloud data provides a more comprehensive reflection of the learner's mastery of the phonological and orthographic aspects of Chinese characters.

Based on the general review of Chinese Interlanguage Corpora in Chapter 1, ten corpora involve audio data. The audio data types in these corpora are primarily natural dialogue, monologue, and other forms of spoken data. Only CISMCIIST involves data types related to learners' pronunciation practice. Although this does not directly correspond to the reading-aloud data type in MICICL, its data processing methods are still valuable for reference.

CISMCIIST

CISMCIIST was developed by the Frontier Innovation Center for Language Resources at Beijing Language and Culture University. It consists of six subcorpora classified by content: monosyllabic, disyllabic, trisyllabic, quadrisyllabic, sentence, and discourse (Cao & Zhang, 2009, p.124). Among them, the Sentence Corpus and Discourse Corpus correspond to the data types used in MICICL, and this section focuses primarily on these two components.

As CISMCIIST is not yet publicly available, its data processing framework can only be inferred from the published description (Cao & Zhang, 2009). The transcription and annotation are performed synchronously within Praat, encompassing multiple tiers of information. For instance, the sentence corpus includes nine annotation layers: Word, TPY (normalized pinyin), I&F (initials and finals), Tone, C-PNT (critical points), Inton (intonation), Dura (duration), BI (breaks/intervals), and Phona (phonation) (Cao & Zhang, 2009, p.128). The transcription records the actual learner pronunciation (including errors), while annotations are inserted in parentheses following the segment, without correction information. The transcription combines Chinese characters and pinyin, where pinyin is rendered using segmental Latin letters with tone numbers instead of diacritic tone marks. Some annotation symbols are used to represent prosodic or phonetic phenomena, such as “!” indicating segmental lowering, and “.” also marking segmental lowering (Cao & Zhang, 2009, p.126).

COPA, HKPU, and MuSSeL

The COPA Corpus was jointly developed by the University of Nottingham Ningbo China and

the Chinese University of Hong Kong. It contains recordings of around 120 Mandarin learners in Hong Kong engaged in conversations with native speakers, comprising both audio and transcribed text. The HKPU Corpus, developed by the Hong Kong Polytechnic University, is a smaller spoken interlanguage corpus that includes narrative speech data from 14 adult English-speaking learners and 6 native speakers of Chinese. The MuSSeL Corpus was created at the University of Utah and covers six target languages, including Chinese, French, Portuguese, and Spanish. The data comprise recordings from bilingual immersion classrooms (children) and adult OPIc speaking tests. Both COPA and HKPU are part of the SLABank within the TalkBank project, and their data follow the CHAT transcription standard³⁰. Although MuSSeL is not part of TalkBank, it also adheres to CHAT conventions³¹. None of the three corpora include corrected speech, so they can be analyzed collectively.

All three corpora transcribe learner output using Chinese characters, while retaining errors and disfluencies present in the original speech. Pauses and repetitions are systematically annotated (e.g., “.” for a short pause, slashes for repetition). Fillers such as “um” or “嗯 (*en*)” are transcribed verbatim. Nonverbal sounds follow CHAT conventions (e.g., “&=laugh” for laughter). For mispronounced words, the transcription first records the learner’s actual pronunciation in numbered pinyin, followed by an annotation indicating the intended word. Repetitions and pauses are marked with precise frequency and position information, while fillers are transcribed in Chinese characters with a preceding marker indicating their non-lexical nature. Segmentation criteria are not explicitly defined, and sentence boundaries are inconsistently marked, sometimes with periods, sometimes with commas, and both using English punctuation marks.

CLDESRISBC

Constructed by National Taiwan Normal University, CLDESRISBC is a multimodal corpus containing both audio and transcribed text. The majority of learners in this corpus come from Singapore with a prior Chinese language background and exhibit a high level of Chinese proficiency. Consequently, they do not present many typical pronunciation issues with

³⁰ see: <https://talkbank.org/0info/manuals/CHAT.pdf>

³¹ see: <https://l2trec.utah.edu/learner-corpora/mussel/about-mussel.php>

individual words.

However, because the region of study is Taiwan, their pronunciation is not standard Mandarin Chinese, but rather Huayu (Taiwanese Mandarin) with a pronounced regional accent. In terms of transcription format, Traditional Chinese characters are used as the transcription symbols. Punctuation is used extensively to represent oral features such as pauses and repetitions. Overall, the corpus provides a comprehensive and authentic transcription of the data. However, as it only includes data from advanced learners, there is no corresponding information available regarding the handling of errors typically observed among non-advanced learners. The primary goal of constructing this corpus was to study pragmatic and communicative strategies in dialogue. Therefore, the annotation mainly focuses on pragmatic aspects, with labels including such as feedback, guidance, topic management³². No correction information is provided.

LCSCC

The LCSCC corpus was primarily constructed by Lin Li at Massey University (Li, 2021). It contains recordings of interview dialogues between 19 Chinese learners and 25 native speakers. This corpus currently only includes transcription information, without any annotation or correction data. It uses Simplified Chinese characters for transcription, but also incorporates Latin letters in certain situations. For instance, filler words are transcribed using Latin letters corresponding to their pronunciation. Since the pronunciation of “*tā*” (他/她/它) is identical in Chinese, and it is impossible to distinguish which character the learner intended based on the audio, the unified form “TA” is used as a substitute. Repetitions and foreign words in the audio are transcribed faithfully. The transcription process generally respects the authenticity of the audio. Punctuation is not used to represent oral features like pauses, and general pauses are not explicitly marked. However, the text is segmented between speaker turns, adopting the convention established by Spoken BNC2014³³ (Love et al., 2017). While the transcription is authentic, the absence of any marking means that only anonymization labels are used for sensitive information where necessary.

³² see: <https://wang-homer.idv.tw/language-data-exchange/%e5%88%86%e6%9e%90%e6%a8%99%e8%a8%98/>

³³ <https://cass.lancs.ac.uk/cass-projects/spoken-bnc2014/>

NUCISC

The NUCISC corpus is a large-scale spoken interlanguage corpus organized and constructed by Nanjing University. Currently, a 1.0 beta version is available for trial use. This corpus provides transcriptions of natural dialogues and monologues between international students and native speakers in a pure text format. The corpus lacks accompanying documentation detailing its specific transcription and annotation strategies, meaning rules must be inferred through browsing the data³⁴. NUCISC uses Simplified Chinese characters to transcribe all audio content. The transcription principle is similar to COPA and HKPU: the actual words spoken by the learner are recorded verbatim, linguistic errors are preserved, and no corrections are made. In the specific transcription process: Pauses in the learner's speech are usually not marked with special symbols, but punctuation marks are frequently used. Filler words such as 嗯 (*ēn*) and 呃 (*è*) appear frequently in the transcription, meticulously recording every instance of learner hesitation. Repetitions are presented verbatim. Non-linguistic sounds (e.g., laughter, coughing) are generally noted by the transcriber in parentheses with a description, such as “(笑)” or “(咳嗽)” to maintain text coherence. Foreign words (e.g., English) are also transcribed faithfully. The corpus contains some annotation labels, such as [Y?-c] (see Figure 2.10), but the meaning of these labels is unclear since the corpus has not provided corresponding documentation.

Figure 2.10

Data format in NUCISC (screenshot)

S: [咳嗽]呃, 我们家乡。因为我住在很多的地方。我住在[Y?-c]呃, 中国和苏格兰, 但是现在我说苏格兰, 不是在[Y?-c]。我住在格拉斯哥一年, 很漂亮, 很冰, 因为格拉斯哥很北苏格兰。呃, 你可以看很多[Y?-c], 有一个地铁, 就是这个地铁很小的, 因为不是新的地铁, 是1951年的地铁, 很小。我们不喜欢去地铁。不知道。格拉斯哥有非常好的公共汽车, 我觉得南京的公共汽车不好。在格拉斯哥, 不是不舒服, 很好, 很便宜。在苏格兰, 你可以去很多地方公共汽车, 是好, 是便宜, 但是你one pound, 一英镑去很多的地方。苏格兰的人口, 是, 格拉斯哥的人口是八百万人, 不是很大的, 苏格兰不是很大的地方, 南京有很多人, 在格拉斯哥, 很多人喜欢喝酒, drinking culture 是很important。呃, 很popular。呃, [咳嗽]。苏格兰的文化是不一样。在Edinburg, 是首都苏格兰。呃, 我觉得, 是很好, 我喜欢下雨。但是, 不是在呃, 从5月到10月, 不是一点冷, 在11月到4月, 很冰。在格拉斯哥, 我去画画的学校。[Y?-]

Source: NUCISC

³⁴ see: <https://yuliaoku.hanyu123.cn/nju>

GCIC

As previously mentioned, GCIC contains data sourced from both writing and speaking activities. Its spoken data is derived from oral examination content or everyday conversational dialogues. The corpus uses Chinese characters as the transcription symbols, supplemented with punctuation. Non-linguistic content, such as coughing, is also marked using parentheses. Due to the absence of specific transcription documentation, the methodology can only be inferred through observing the corpus samples. However, browsing the corpus data reveals inconsistent transcription standards. For instance, filler words are sometimes transcribed using Chinese characters (Example 7) and sometimes using Latin letters (example 8):

(7)

所以，我，嗯可以学习汉语。

suó yǐ , wǒ , ēn kěyǐ xuéxí2 hànǔ.

Source: CIC-BLNU

(8)

um，所以我，um，我真喜欢啊住在南京。

um, suóyǐ wǒ, um, wǒ zhēn xǐhuān a zhù zài Nánjīng.

Source: CIC-BLNU

Inconsistent handling is also evident in pauses:

(9)

所以我 um，我可以 um，我可以，我可以 um，我可以 um 看来南京和香港这两个城市。

suóyǐ wǒ um, wǒ kěyǐ um, wǒ kěyǐ, wǒ kěyǐ um, wǒ kěyǐ um kàn lái Nánjīng hé Xiānggǎng zhè liǎng gé chéngshì.

Source: CIC-BLNU

(10)

还有她的——嗯——所以我，啊——阿拉送是，阿拉送的很的城市。

háiyǒu tā de —— en ——suóyǐ wǒ, a —— ālā sòng shì, ālā sòng de hěn de chéngshì。

Source: CIC-BLNU

Although the original audio is unavailable for reference, both sentences clearly contain multiple pauses within the clause. However, the first sentence uses commas to mark pauses, while the second uses the em dash (“——”). Without transcription documentation, it is impossible to determine whether this difference is due to the categorization of distinct pause types (e.g., varying duration) marked by different symbols, or simply the result of differing transcription habits among transcribers. Furthermore, there is inconsistency in the treatment of numbers, which are sometimes transcribed using Chinese characters and sometimes using Arabic numerals. This lack of standardization inevitably affects subsequent corpus analysis and retrieval. Similar to CIC’s handling of writing data, the transcription, annotation, and correction are placed on the same layer, which leads to comprehension difficulties example 11).

(11)

嗯，在中国学习的【时】[Cd]时候，我想，在中国【】[Cq 有]很多人，走路在印尼也有【】[Cq 很]多的人走路，所以我我走路走得很慢，嗯，所以中国跟印尼差不多一样。

éng, zài Zhōngguó xuéxí de 【shí】[Cd] shíhou, wǒ xiǎng, zài Zhōngguó 【】[Cq yǒu] hěn duō rén, zǒulù zài Yìnní yě yǒu 【】[Cq hěn] duō de rén zǒulù, suǒyǐ wǒ wǒ zǒulù zǒu de hěn màn, éng, suǒyǐ Zhōngguó gēn Yìnní chà bu duō yīyàng.

Source: CIC-BLNU

In the example, the word “时” in the first bracket is a spoken filler that should be deleted. The second and third empty brackets indicate a missing word, but this can misleadingly suggest a pause to the reader.

GLCLC

The spoken transcription scheme for this corpus is similar to that of NUCISC. It uses Chinese characters as the transcription symbols and adds appropriate punctuation based on semantics. Repetitions and filler words are transcribed as they actually occur. It is unclear whether the absence of any signs of pronunciation hesitation or errors concerning individual words and

characters is because the included learners are all highly proficient. However, the corpus does not mark any pauses or pronunciation errors. Furthermore, it provides no annotation or correction information.

BILCC

The spoken subcorpus of BILCC, developed by Iurato (2024) at Ca' Foscari University of Venice, follows the transcription conventions of LCSCC and applies orthographic transcription to spoken Chinese data. It includes the Pinyin romanization of each Chinese character, followed by an English translation of each word and a full-sentence translation. For grammatical morphemes, the standard abbreviations provided by the *Leipzig Glossing Rules* were used, with some additions. Tags were also employed to mark foreign words or unclear items occurring in speech, and punctuation marks were retained in the transcription. Regarding annotation, since BILCC was designed to investigate Italian learners' mastery of the fixed Chinese structure *shi...de* (“是...的”), the annotation in the spoken subcorpus focuses specifically on certain linguistic targets, such as the overuse or underuse of grammatical elements (e.g., the aspect marker *-le*).

Similar to handwritten data, there is currently no unified standard for the transcription norms of Chinese interlanguage audio. In fact, the transcription process is often overlooked in existing corpus development, with researchers primarily focusing on subsequent annotation work. As pointed out by Li (2021), transcription is still not a core concern for most Chinese spoken corpora research. Furthermore, existing corpora show significant variation in the following aspects:

- **Form of Transcription Content (Pinyin vs. Chinese Characters):** The majority of corpora opt to record the learner's spoken output using Chinese characters to ensure readability and semantic clarity of the results. This approach aligns with the orthographic norms of Chinese as the target language, allowing readers to understand the corpus content without needing to master special symbols. However, the “normalized” nature of Chinese characters can mask pronunciation errors, thereby weakening the authenticity at the phonological level. Only a few corpora (e.g., CISMCIIST) partially utilize Pinyin for

transcription. While character transcription enhances content readability, Pinyin transcription offers a more detailed representation of phonetic features, making it especially suitable for capturing pronunciation errors. Generally, for corpora focusing on highly proficient learners (e.g., CIC), character transcription is reasonable due to the low frequency of pronunciation errors. However, for corpora covering multiple proficiency levels, Pinyin transcription may be more appropriate as it can more comprehensively document phonetic variations.

- **Use of Punctuation Marks:** Punctuation marks are fundamentally components of written language and do not exist in spontaneous spoken discourse. However, with the exception of the LCSCC corpus, most corpora artificially insert punctuation into the transcription to delineate clauses. This practice raises concerns with respect to the principle of authenticity in spoken language transcription, where prosodic and temporal features are ideally represented based on the speech signal rather than imposed orthographic conventions (McEnery & Hardie, 2011; Jones, 2021). While punctuation is often introduced to enhance readability, it may obscure the underlying structure of spoken discourse by projecting written language conventions onto speech data (Liu, 2022; Aggazzotti, Andrews, & Smith, 2024). As noted by Liu (2022, p. 27), transcription systems should aim to minimise the use of written-language symbols and preserve the original characteristics of the audio data as much as possible. From a corpus design perspective, an alternative approach is to adopt dedicated annotation systems (e.g., time-aligned markers or symbolic tiers) to represent pauses and prosodic features, thereby improving both the authenticity and analytical precision of spoken transcription (Schmidt & Wörner, 2014; Wang & Wang, 2024; Brill & Smith, 2024; Lukeš et al., 2024).
- **Marking of Pauses and Repetitions:** Existing corpora generally do not measure pause duration with precision, instead using simple symbols or punctuation to denote pauses. Furthermore, the symbolic systems are not consistent across different corpora, and inconsistencies can exist even within the same corpus. In addition to unnatural pauses, learner speech frequently features various repetitions. These repetitions differ from

rhetorical repetitions in native speech; they typically stem from factors of disfluency, such as unfamiliarity with pronunciation or difficulty in lexical retrieval. As Hu and Wang (2011, p. 4) noted, these phenomena reflect both the learner's thinking process and their speaking fluency, making them valuable for research and necessitating their retention in the transcription. Most corpora adopt a verbatim repetition transcription method, faithfully presenting the original audio content, which is a commendable practice.

- **Handling of Filler Words:** Filler words (e.g., 嗯 (*ēn*), 呃 (*è*)) are common non-lexical components in spoken language, and their method of recording directly affects the corpus's operability and visualization. Most corpora choose to record every instance of a filler word to fully reflect the frequency of learner hesitation and pausing, which is highly valuable for analyzing language proficiency. However, overly detailed rules for transcribing fillers can burden both transcribers and users. Functionally, whether the vocalization is “en” or “a,” the pragmatic function remains the same—reflecting learner disfluency. Love (2017) suggested that phonetic differences in fillers often lack substantive meaning, allowing for the adoption of simplified strategies. From an operational perspective, requiring transcribers to precisely distinguish every filler variant in spontaneous dialogue is time-consuming and difficult to standardize. Andersen (2016, p. 343) advised that to enhance consistency, the annotation of such features should be limited to a “reasonably minimal set,” such as using a unified symbol to represent all filler words.
- **Annotation of Non-Linguistic Content:** Spoken corpora often contain non-linguistic elements such as laughter and coughing. While some corpora choose to retain this content, its reflection of the learner's linguistic ability is limited. Corpus recording should balance authenticity and economy; excessive non-linguistic annotation adds to the burden of construction and use. However, the learner's real-time interaction with researchers during the speaking task (e.g., questions or explanations) should be retained. Although this content is not part of the interlanguage output, it reflects the difficulties the learner faces in comprehension or expression. For example, they may switch to English or another

language to seek clarification, and this information aids in the comprehensive assessment of their language proficiency.

Chapter 3 Corpus Design and Data Collection

3.1 Chapter overview

Building on the review of existing Chinese interlanguage corpora in Chapter 2, which identified a gap between corpus construction and pedagogical needs, this chapter presents the design and implementation of MICICL (Multimodal Interlanguage Corpus for Italian Chinese Learners), a multimodal interlanguage corpus developed for Italian native speakers learning Chinese. Designed with a clear pedagogical orientation, MICICL integrates multimodal learner data—including handwriting, reading-aloud, and aligned transcriptions—to systematically capture Italian learners' Chinese learning performance across multiple dimensions and to provide empirically grounded evidence for corpus-based teaching and learning.

Section 3.2 introduces the design principles and overall framework of MICICL, detailing its modalities, data types, and annotation structure. Section 3.3 reports the ethical review and approval process, outlining measures for participant consent, data anonymization, and secure storage. Section 3.4 describes the pilot study conducted to test the feasibility of data collection and annotation procedures and the subsequent refinements made to the corpus design. Section 3.5 presents the main study, including participant recruitment, metadata collection, and the scope and timeline of language data collection, which together form the basis for the data processing and analyses in the following chapters.

3.2 Corpus design

In learner corpus research, corpus design is typically guided by a set of core criteria, including learners' language background, data type and modality, and the temporal structure of data collection (Granger, 2008; Granger, Gilquin, & Meunier, 2015; Tracy-Ventura & Paquot, 2021). Following these widely adopted design principles, MICICL is characterized along the following dimensions.

With respect to learners' language background, MICICL exclusively includes data from L1 Italian speakers. In terms of data type and modality, the corpus comprises handwritten data

produced in natural learning contexts, together with corresponding reading-aloud data. These tasks yield three aligned data modalities—image, audio, and text—allowing for multimodal analysis of learners’ written and oral performance. Regarding its temporal structure, MICICL combines longitudinal and quasi-longitudinal characteristics, reflecting both repeated observations of the same learners and the practical constraints of classroom-based data collection. The following sections elaborate on each of these design features in detail.

3.2.1 L1 background

As said in Section 1.1.2, studying the interlanguage of a single native language group helps to reveal the specific mechanisms by which L1 transfer impacts Chinese language acquisition. However, most existing Chinese interlanguage corpora (see Section 2.3) are designed to be multi-L1, with mixed native language backgrounds. This complexity makes it difficult to reveal the systematic transfer patterns exhibited by a particular L1 group in their Chinese learning, both at the theoretical and empirical levels. Consequently, this study clearly defined the native language boundary during the corpus construction phase, recruiting only Italian native speakers to participate in data collection. All participants came from Chinese language programs at University of Bologna, covering multiple stages of learning duration and proficiency. This design ensures the homogeneity and comparability of the corpus, providing a solid foundation for analyzing the characteristics of Chinese acquisition by Italian native speakers.

The L1-specificity of this corpus is valuable for several reasons. It helps to systematically reveal the typical error patterns of Italian native speakers in Chinese learning. This will not only deepen the understanding of cross-linguistic transfer mechanisms but also provide empirical reference for teaching Chinese as a foreign language, allowing teachers to design more targeted instructional content and strategies. Furthermore, this choice offers a comparable reference framework for cross-linguistics contrastive studies. In the future, by contrasting this corpus with corpora from other L1 groups, researchers can further investigate the role of typological differences between native languages in Chinese acquisition.

3.2.2 Data collection setting

The overarching principle followed during the data collection phase of this study was to maximize the authenticity and naturalness of the learners' language production.

As discussed in Section 2.3, some interlanguage corpora rely on data produced during tests and examinations. However, such contexts may place additional pressure on learners (Gilquin & Granger, 2015), potentially influencing their language production and affecting the naturalness of the data. The present study sought to minimise, rather than eliminate, the tension, anxiety, or strategic behaviour that may arise in formal assessment settings. To this end, no standardized tests or explicitly test-like tasks were employed, and no strict time limits were imposed during data collection. Learners were allowed to complete the tasks at their own pace, while the researcher remained present. It is acknowledged that the data collection setting cannot be considered entirely free of pressure, as the presence of the researcher and the task-based nature of the activity may still influence learner performance. However, compared to formal testing conditions, a non-assessment and untimed approach can help reduce performance pressure and may yield data that more closely approximate learners' typical language use.

To achieve this, the main part of the corpus was collected in a one-on-one format between myself and the learner. Depending on the participant's location and schedule, the collection process included both in-person and online video formats. The one-on-one approach provided a degree of supervision, effectively preventing the use of external aids such as dictionaries or online translation software, thereby minimizing the impact of technological interference on language output. This setup not only ensured the authenticity of the corpus but also better reflected the learner's true mastery of character form memory and phonetic retrieval when lacking external support.

During the collection, I maintained a relaxed and informal atmosphere to further reduce the learners' language anxiety. Before formally starting the task, I typically engaged in a brief exchange with the learner, introducing the study's aims and procedures or casually chatting about everyday topics to help them relax. I explained to the learners that this was not an examination or grading activity, and the results would not be linked to their academic

performance. In the writing phase, participants completed the text independently on paper. In the reading phase, they were invited to read the text they had just finished writing. Throughout the entire process, I did not offer any form of guidance or correction on the learners' language production, providing only technical or operational instructions when necessary. My role was limited to task instruction and process supervision, not intervention in the linguistic content. This non-interventionist collection method ensured that the corpus maximally reflected the learners' natural and independent language processing, avoiding the influence caused by researcher involvement.

Furthermore, considering the balance between corpus coverage and quantity, this study also collected a small amount of data completed under unsupervised conditions, including some learners' classroom assignments and independently completed writing or reading texts. Although this type of data was not collected under researcher supervision, it still holds a certain degree of representativeness and helps supplement tasks that individual learners were unable to complete.

3.2.3 Data collecting tasks

To capture learners' interlanguage performance across writing and reading modalities, each data collection session in MICICL consisted of two tightly linked tasks: a handwritten task followed by a reading-aloud task based on the learner's own written output.

In each session, participants were first asked to complete a handwritten task, producing a short written text in Chinese in a classroom setting under supervised conditions, without access to external aids such as dictionaries or digital tools. Immediately after completing the writing task, they were asked to read aloud the text they had just written. Both tasks were conducted individually and recorded separately, yielding aligned handwritten images, reading-aloud audio, and corresponding transcriptions.

The choice to collect handwriting rather than typed data was made to more authentically reflect the learner's actual mastery during the Chinese language acquisition process. Chinese writing involves the interaction of orthography, phonology, and semantics, with relatively opaque form-sound correspondences compared with alphabetic writing systems (Zhang & Min,

2019). Therefore, in Chinese learning, the promotion of writing ability by reading is not as significant as it is in alphabetic languages. Conversely, the process of writing Chinese characters requires the learner to establish coordination between visual attention and sensory-motor activity. This process helps to form a visual-spatial memory with motor-memory traces, thereby providing additional support for the activation of visual information during character recognition. It is thus evident that writing holds a central position in the process of Chinese character learning and reading development (Cao et al., 2013; Guan et al., 2011).

Although some Chinese interlanguage corpora choose to include typed texts (see Section 1.2), I believe that the technical intervention in typing input masks the learner's true level of orthographic knowledge mastery (as discussed in Section 1.3.2). Therefore, the decision was made to collect handwritten data to more effectively present the learner's authentic interlanguage performance in Chinese character acquisition. Existing studies have shown that handwriting practice is more effective than typing in promoting the orthographic-semantic mapping of Chinese characters and improving long-term memory retention (Guan et al., 2011, 2015; Zhang & Li, 2010). The interview results of Zhang & Min (2019) also indicated that learners generally believe that handwriting aids in memorizing Chinese characters.

Building upon the collection of learners' handwritten data, this study further included a reading task to demonstrate the correspondence between the learner's character-sound and character-form knowledge.

If the handwritten data reveals the learner's mastery of character structure and orthographic knowledge during the writing process, the reading-aloud data is designed to examine their ability to reproduce these written units at the phonetic level. In other words, the reading corpus complements the aspects that the handwritten corpus cannot reflect, namely, the learner's true state of "can they read what they can write" .

Compared to a free oral conversation task, the reading task offers higher controllability and comparability. In oral communication, a learner's language output is often influenced by communication strategies, pragmatic goals, and other factors, making their phonetic performance highly context-dependent and variable. In contrast, the reading task requires the

learner to convert written symbols into speech output, which can more accurately reflect their cognitive processing in orthography–phonology mapping under controlled conditions.

Furthermore, the reading task in this study did not require learners to read a standardized text; instead, they were asked to read the text they had personally written. The innovation of this design lies in its ability to more effectively reveal the gap between a learner’s ability to write and their ability to read. Through this one-to-one writing-reading correspondence, this study can further explore learners’ phonetic decoding errors in self-produced materials, thereby observing their interlanguage performance during the cross-modal learning process.

3.2.4 The modalities and media of the corpus

Following the previous section’s introduction to the data collection methodology, this section clarifies the modal organization of the corpus by distinguishing learners’ language modalities from the data formats and processing procedures adopted during corpus construction. At the level of learner modality, MICICL includes two primary modalities: written modality (handwriting) and spoken modality (reading aloud). These constitute the core language data produced by learners. At the level of data formats (or media), the corpus represents these modalities using three formats: image data for handwritten texts, audio data for reading-aloud recordings, and text data derived from the transcription, annotation, and correction of both written and spoken modalities. This separation ensures a clear distinction between learner modalities, data representation, and subsequent processing within the corpus architecture.

In the corpus construction phase, all learners’ writing data and reading-aloud data underwent systematic transcription, annotation, and correction. The transcription process involved converting the learner’s handwritten text and reading content into machine-readable text formats to facilitate unified management and analysis. Subsequently, the transcribed text was subjected to multi-layered annotation, and the annotated data was then completed with error corrections. This process ensures the retrievability of the data, making the corpus both capable of retaining the characteristics of the original output and operable for analysis.

In the final presentation stage, the corpus will showcase data in three formats:

- i. Images: Scanned images of the learner's handwritten texts. These fully preserve handwritten features, correction traces, and spatial layout, providing an intuitive basis for writing analysis and orthography research.
- ii. Audios: Recording files of the learner reading the text they personally wrote. This is used for phonetic analysis and cross-modal comparison.
- iii. Texts: Text files generated from the transcription, annotation, and correction of the first two types of data. This is used for the analysis of linguistic structure and error patterns.

It should be noted that although the corpus design is multimodal, the reading-aloud data was collected using only audio recording, not video recording. While video corpus can present non-verbal information such as facial expressions, gestures, and gaze, which is valuable for studying pragmatic interaction or multimodal communication, the core focus of this study is the language acquisition process itself. Specifically, the focus is on the learner's mastery of the Chinese grapheme–phoneme mapping (sound-form correspondence) during the writing and reading tasks, not on non-verbal behavior. Audio data is sufficient to fully reflect the learner's phonetic features and pronunciation patterns, while avoiding the complexities associated with video recording in terms of privacy protection, storage volume, and technical processing.

3.2.5 Longitudinal and quasi-longitudinal

In terms of the temporal dimension design of the corpus, this study adopted a combined approach of quasi-longitudinal and longitudinal methods, aiming to address both group developmental trends and individual dynamic changes.

A quasi-longitudinal corpus is typically composed of data from learners at different proficiency levels, and can be used to analyze systematic differences and developmental patterns of the interlanguage across various proficiency stages (Granger, 2002). This study collected data from learners at different proficiency stages and with varying lengths of Chinese study, and categorized them into different proficiency groups for comparison. This design allows for the revelation of linguistic feature differences across various learner stages through

cross-sectional comparison, providing rich evidence for analyzing the stage-specific patterns of interlanguage development.

Concurrently, this study also adopted a longitudinal design, conducting data collection 11 times from the same group of learners over approximately 20 months. Although learner attrition and re-recruitment issues meant that not all learners participated in the full time series, a substantial portion of the longitudinal sample remains, capable of presenting individual language development trajectories over a certain time span (see later in Section 3.5.3).

Regarding the collection method for longitudinal data, comparability of the corpus and consistency of task design were maintained. The effectiveness of longitudinal corpus analysis largely depends on the consistency of data sources and collection methods (Gass & Mackey, 2011). If different tasks or topics are used to elicit learner output at different time points, it becomes difficult to distinguish whether language change is caused by the time factor or the task factor. Therefore, this study sought to maintain consistency in task requirements and procedures during each round of data collection, minimizing the interference of external variables. However, completely repeating the same tasks could also lead to new problems: early similar topics might affect later expression, and excessive repetition could diminish participant interest and reduce data naturalness. Addressing this issue, the study referenced Cao's (2013) suggestion, adopting a cyclic rather than repetitive task design. This means maintaining similarity in the task type but making moderate changes in the content, thereby achieving a balance between comparability and authenticity.

It must also be noted that no research can definitively establish the time frame over which a learner experiences a systematic qualitative leap. Language learning trajectories are highly individualized (Tasker, 2018); a learner's pace of progress is determined by their learning methods, time investment, and language contact environment. In other words, even over a period of 20 months, while learners may achieve development at different levels, this development does not necessarily constitute a proficiency level shift. Therefore, adopting a quasi-longitudinal comparison method during the data analysis stage remains rational—it can broadly compare the linguistic features of learners at different proficiency levels, revealing

group-level interlanguage development patterns, without being interfered with by individual learning progress differences.

The combination of longitudinal and quasi-longitudinal approaches also offers the potential for further investigation into the internal mechanisms of interlanguage development. The quasi-longitudinal corpus helps us map the overall developmental differences across proficiency stages at a macro level, while the longitudinal corpus allows us to observe individual language evolution processes over a limited time span at a micro level. The two methods complement each other: the former provides structured cross-sectional comparison, and the latter offers dynamic longitudinal evidence.

3.3 Ethics and privacy considerations

3.3.1 Review documents and approval outcome

The construction and use of MICICL involve the collection of student handwritten samples, audio recordings of their reading, personal information, and language learning background data. These types of data contain sensitive or potentially identifiable information which, if improperly handled, could lead to privacy breaches. Therefore, formal approval from the relevant ethics body was required prior to the commencement of the study.

The ethical review for this project was conducted by the Ethics Committee (Comitato di Bioetica³⁵) of the University of Bologna. Prior to data collection, the required documentation was submitted to the Comitato di Bioetica of the University of Bologna, including the Research Protocol (Richiesta parere Comitato Bioetica) and the Privacy and Informed Consent Form (Privacy Informativa Generica Ricerca), both based on official institutional templates³⁶. Data collection and processing procedures complied with the EU General Data Protection Regulation (Regulation (EU) 2016/679)³⁷ and the Italian Data Protection Code (D.Lgs. 196/2003, ss.mm.ii.), and informed consent was obtained from all participants. In addition, a

³⁵ <https://www.unibo.it/it/ricerca/organizzazione-e-infrastrutture-di-ricerca/comitati-etici-1/comitato-di-bioetica>

³⁶ Alma Mater Studiorum – Università di Bologna. (2023). *Privacy Informativa Generica Ricerca (ENG)* [Information and consent form]. Bologna: Comitato di Bioetica.

³⁷ <https://www.privacy-regulation.eu/en/article-13-information-to-be-provided-where-personal-data-are-collected-from-the-data-subject-GDPR.html>

Statement of Commitment (Dichiarazione di impegno) was signed by the supervisor and the author, and the study was formally approved by the Ethics Committee (Estratto del Comitato di Bioetica).

3.3.2 Compliance during data collection, processing, and application

During the data collection phase, two categories of data were gathered:

- i. Students' metadata, including personal information such as age and gender, and language background information such as learning experience.
- ii. Students' language data, including handwritten samples and audio recordings, but not personal images.

All data collection took place only after participants signed the Privacy and Informed Consent form. When distributing the form, I emphasized that participants could ask questions at any time to ensure they signed under conditions of full understanding. As the study follows a longitudinal design, some learners participated multiple times. The consent form was therefore signed only once, during the participant's first session, rather than at every occurrence. All signed documents were archived alongside the data as proof of compliance.

It is important to note that the data collection process consisted of two phases: a pilot study and a main study. The pilot study was conducted to test the feasibility of the task workflow and data collection procedures. Since the quality and applicability of the pilot data could not be guaranteed beforehand, I informed participants that pilot data would be used only for preliminary observation and not for constructing the final corpus. Nevertheless, all pilot participants signed the same informed consent form.

Pilot data were not automatically incorporated into MICICL for compliance reasons. However, some learners participated in both the pilot and the main study, and many of their pilot samples met the standards required for the main study. To balance ethical compliance with the goal of avoiding unnecessary data loss, I re-approached these learners during their first session of the main study, explicitly asking—via written confirmation—whether they agreed to include their pilot data in the final corpus and analyses. Only after receiving their renewed, written consent were pilot data integrated into the dataset. This procedure preserved valuable

data while ensuring full transparency, ethical legitimacy, and respect for participant rights.

During the data processing stage (including transcription, annotation, and correction), handwriting and audio data were manually transcribed into machine-readable text. Strict anonymization principles were followed. All real personal names in handwriting and audio data were replaced with the tag [NM]. Fictional names (e.g., “Harry Potter”) and names not pointing to identifiable persons (e.g., pet names) were preserved to maintain naturalness. This approach minimizes privacy risks while maintaining the integrity and authenticity of the corpus.

During data application and dissemination, MICICL data are used for research analysis, teaching material development, and academic publication. Given the inherent re-identification risks of multimodal data—e.g., handwritten characteristics or identifiable voice signatures—strict privacy protection measures apply across all scenarios. All data are subject to rigorous anonymization and minimization, such as deleting signatures or personalized markings in handwriting, muting or replacing spoken real names in audio, and replacing all names in transcription with [NM]. The teaching materials use only short excerpts to avoid exposing identifiable features. Future dissemination of the corpus will provide only strictly anonymized versions of transcription, annotation, and correction files.

Although the ethics application explicitly addressed the potential for sharing anonymised handwritten and audio data, the dissemination of multimodal data involving identifiable features (e.g., voice) raises additional privacy considerations. In the present study, voice data were not modified (e.g., through voice distortion), as all participants provided informed consent for the use of their recordings under controlled research conditions. However, to minimise potential privacy risks, a restricted access strategy was adopted. At the current stage, only anonymised transcription and annotation data are made available through the corpus platform. The original multimodal data (including handwritten images and audio recordings) are not publicly downloadable. This decision is partly due to the lack of suitable platforms that support view-only access without download functionality for sensitive multimodal data. Therefore, access to the original data is granted only upon request for research purposes, under controlled conditions and in compliance with the informed consent provided by participants.

3.3.3 Data encoding, storage, and retention period

To guarantee data security and anonymization, the study established a dual encoding mechanism:

First, at the identity protection level, each participant was assigned a personal code (01–39) upon entering the database. This code was assigned based on the order of participation and included an element of randomness, ensuring both uniqueness and preventing any direct link to the participant’s real identity for external users.

Second, at the corpus level, each piece of data was assigned an independent data code composed of five parts: the learner’s personal code, the number of times the learner participated, the session number, the data collection time, and the data modality. This design ensures both longitudinal and horizontal traceability of the corpus without including any personally identifiable information. In all public dissemination and results presentation, citations involving participants will use these codes instead of names, student numbers, or other personal identifiers.

Regarding data storage, all electronic data are currently temporarily held by me on an encrypted, password-protected hard drive, with regular local backups. Data is not uploaded to public cloud services during the research phase to prevent unauthorized access or leakage. All physical documents (including signed consent forms) are secured in a locked filing cabinet to prevent unauthorized direct association between data and identity. It should be emphasized that the statement in the ethical documents about data being “stored on the university server” refers to the final ownership and legal responsibility residing with the University of Bologna. Post-research, the data must be transferred to the university server or managed under the supervisor’s guidance to meet long-term archiving and institutional monitoring requirements. The current personal storage method is a transitional arrangement for the research phase and does not violate the ethical document’s requirements.

Concerning the retention period, the *Privacy Informativa Generica Ricerca* explicitly stipulates that participants’ language data will be retained for one year by default, extendable to two years with explicit participant consent, after which it must be permanently deleted or

subjected to irreversible anonymization. Documents related to ethical compliance (e.g., research protocols, approval decisions, consent form copies) must be retained by the department for five years to satisfy audit and oversight requirements (*Richiesta parere Comitato Bioetica*). It must be noted that these retention periods apply to original data that could still be identified (e.g., full handwritten samples, unprocessed audio). Once data undergoes irreversible anonymization—such as deleting or replacing names, removing potentially identifying features, and de-individualizing voiceprints or handwriting—it no longer constitutes “personal data” in the legal sense. Therefore, the data provided for external corpus release and teaching materials will be the strictly anonymized version, which is not subject to the 1–2 year retention limit and can be archived and publicly accessible online within a compliant framework.

In conclusion, this research has fully accounted for the privacy risks and legal responsibilities inherent in a multimodal corpus through its ethical application and implementation. From the protocol and privacy statement to the commitment declaration and final approval, all stages comply with regulatory and academic requirements.

3.4 Pilot study

3.4.1 Objectives of the pilot study

In the first half of 2023, I conducted a pilot study that included four stages: student recruitment, informed consent acquisition, metadata collection, and language data collection. The main purpose of the pilot study was to verify the feasibility of the data collection, test the rationality of the procedure settings, and provide operational and institutional groundwork for the formal data collection through a small-scale, complete rehearsal.

The main objectives of the pilot study included the following six aspects:

- i. Preliminary Observation of the Teaching Context. The pilot study also served as a classroom survey, aiming to understand the characteristics of the real teaching environment and the learners’ language use. After obtaining permission from the instructor, I audited classes and gave a brief presentation at the University of

Bologna's Chinese language program as part of the in-person participant recruitment. This process not only improved recruitment efficiency but also allowed me to directly observe the organizational models in classroom teaching, the balance maintained by the teacher between Pinyin and character instruction, the methods used to provide feedback on student writing errors, and the common difficulties students face in oral expression. These observations provided empirical reference for the subsequent corpus collection task design and corpus application research.

- ii. **Verification of Procedure Feasibility.** The construction of a multimodal corpus involves multiple steps (task assignment, recording, etc.). The pilot study rehearsed these steps completely to test whether the process was smooth, the time allocation was reasonable, the operational burden was acceptable, and to identify potential technical or organizational problems. This process provided the foundation for time planning, technical configuration, and procedure coordination for the formal collection phase.
- iii. **Testing the Task Design.** The pilot study was used to examine whether the data collection task design was practical. Although the initial plan was based on theoretical research, the task's difficulty level, clarity of instructions, time duration, and learner engagement still needed empirical verification. The pilot phase helped observe learners' understanding and execution of the tasks during actual operation, thereby judging the rationality of the task design at both the linguistic and operational levels.
- iv. **Verification of Data Collection Formats.** In addition to in-person collection, the pilot study tested the feasibility of remote collection via the Teams platform, assessing the audio quality, flow smoothness, and participant cooperation in an online environment. The purpose of this step was to evaluate whether online collection could serve as an effective supplement to in-person collection, ensuring the flexibility to handle cross-campus or remote participants in the formal study.
- v. **Verification of Ethics and Compliance procedures.** The entire process strictly adhered to research ethics guidelines. All participants signed informed consent forms, and data collection and storage followed the principles of anonymity. The purpose of this step

was to check the completeness of the ethical procedures, identify potential omissions in privacy protection and data management, and provide a basis for the formal ethics application (see Section 3.3 for details).

- vi. Collection of Learner Feedback. During the pilot phase, every participant underwent a brief interview after the data collection, primarily to understand their Chinese learning experience, difficulties, and subjective experience with the task. This was intended to evaluate the acceptability of the task design from the learner's perspective, provide reference for optimizing the content, instructions, and time control of the formal collection tasks, and offer practical basis for the corpus application.

Overall, the pilot study provided the practical foundation and methodological guarantee for the main data collection through observation of the teaching context, procedure testing, task verification, online format rehearsal, and ethical procedure checking. It not only helped confirm the technical and organizational feasibility of the data collection plan but also provided empirical reference for subsequent corpus construction.

3.4.3 Pilot study data collection procedure

In this subsection I will provide a detailed account of the design and procedures for the three processes: student recruitment, data collection arrangements, and the data collection process (including obtaining informed consent, metadata collection, language data collection, and participant feedback).

First, student recruitment. Due to ethical and privacy considerations, recruitment was confined internally to the University of Bologna. Recruitment adopted a purely voluntary registration and participation model, offering no monetary or academic credit compensation, but only providing language feedback to students who requested it after the collection was complete. In early 2023, I designed a project recruitment poster, posted it in public areas on campuses, and shared it online via department faculty and social media. In February–March 2023, after communicating with and receiving permission from several Chinese teachers on both campuses (Forlì and Bologna), I entered the classrooms to introduce the project and answer questions. Concurrently, I audited courses at different years to understand the teaching

approach to Pinyin and characters, and the difficulties students faced in writing and speaking. By the end of March 2023, a total of 55 students had registered to participate.

Second, data collection arrangement. In April 2023, I contacted each registered student individually via email to confirm their willingness to participate and schedule a data collection time. Participants could choose from three data collection methods: 1) an in-person one-on-one meeting; 2) an online one-on-one meeting via the university's Teams platform; or 3) completing the form themselves. I encouraged them to attend the in-person meeting in the email, as this was considered the best method (see Section 3.1); in fact, it was the most popular choice. Students who were unable to attend in person could choose an online meeting or complete the form independently.

Last, collection process. The data collection for the pilot study took place in April 2023. Each participant completed the full procedure in a one-on-one, face-to-face or Teams online meeting. The meeting included the following four steps:

Step1: Signing the Informed Consent Form (Approx. 5 minutes)

Before the meeting began, participants were required to read and sign the informed consent form. Although the pilot study was launched before the ethics approval was finalized, research ethics guidelines were strictly followed throughout. I explained to the participants the study's purpose, procedures, potential risks, data preservation methods, and their right to withdraw at any time. After confirming their understanding, participants signed the informed consent form, and I signed to confirm, with both parties keeping one copy.

Step 2: Metadata Collection (Approx. 5 minutes)

The consent form contained the basic information on the project, contact information etc., and needed to be signed by each participant and myself, in two copies, each keeping one. After signing up, participants needed to fill out a metadata collection form, including personal information, language background and proficiency self-assessment.

Step 3: Language Data Collection (Untimed)

The writing task was completed on A4 white paper and saved as a PDF by scanning with a mobile phone. The reading task was recorded simultaneously using a laptop and a mobile phone.

Step 4: Feedback (Approx. 5–10 minutes)

After completing the language data collection task, I engaged in an exchange with the participants to solicit their opinions. This included feedback on the data collection task: their feelings about the difficulty and format of the task, expectations for the corpus, and potential use scenarios. It also covered their daily Chinese learning situation: difficulties encountered in daily study, forms of daily classes and exams, and methods for after-class practice and self-study. This feedback was systematically recorded and serves as an important reference for corpus application research.

3.4.3 Pilot study data collection content

The data collection content includes metadata collection content and language data collection content.

For metadata collection in interlanguage corpora, standardized frameworks have long been limited and heterogeneous across studies. Recently, more systematic proposals have begun to emerge, most notably the learner corpus metadata schema proposed by Paquot et al. (2024), which aims to harmonize metadata categories across learner corpus projects. The design of MICICL's metadata collection predates the publication of this schema and was therefore developed independently, drawing on commonly adopted practices in learner corpus research and corpus linguistics more broadly (Granger, 2002; McEnery & Hardie, 2011). In MICICL, learner metadata is organised across three main dimensions: personal information, language-related background, and language proficiency, reflecting standard distinctions in learner corpus design. Personal information includes basic details such as name and gender, while language-related background covers aspects such as the learner's first language, home language(s), and other languages acquired. Despite being developed independently, MICICL's metadata structure shows substantial conceptual overlap with the dimensions proposed by Paquot et al. (2024), particularly with regard to learner background and proficiency-related variables, allowing for potential alignment and comparison in future corpus integration or reuse.

As for student Chinese proficiency information, while Italian has the Common European Framework of Reference for Languages (CEFR) for standardized proficiency assessment,

research on Chinese is less common (Bowles et al., 2021), and Chinese currently lacks a universally applicable, officially recognized standard for proficiency division. Although the widely known HSK exam is the most frequently taken Chinese proficiency test (Peng et al., 2021), it has been characterized as “a large-scale, high-stakes, and yet underrepresented language proficiency test” (Yan & Lin, 2023, p. 4). It is difficult to use the HSK as a proficiency reference for the students in this project because, firstly, it is not mandatory for Italian students, so not everyone has taken it and has a score. Secondly, the design of the HSK exam itself is questionable, as it only assesses listening, reading, and writing abilities for each level. Candidates wishing to be assessed on speaking ability must opt for the separate speaking test—the HSK Speaking Test (HSKK)—making the assessment incomplete. Therefore, during the pilot study phase, I attempted to understand the learners’ proficiency through self-assessment. The metadata collection form included a self-assessment of Chinese proficiency level, which was a multiple-choice question with three options.

Figure 3.1

Metadata collection form used in pilot study of MICICL

Metadata collection form

Name and surname		Gender	☐ Male ☐ Female ☐ Other
Country of birth		Mother tongue	
Degree course			
Grade	☐ BA ☐ MA ☐ PhD	☐ First year ☐ Second year ☐ Third year	
Home language (the most commonly used language in daily life)			
Whether you know other languages (if so, what language and at what level)			
How long you have been learning Chinese			
Chinese level	☐ Can read Pinyin, cannot read independently without Pinyin; can write Pinyin, but cannot write Chinese characters; can speak complete sentences orally. ☐ Able to read simple Chinese characters; able to write simple Chinese characters (100-200); able to complete daily conversations. ☐ Able to read aloud fluently; able to write more Chinese characters (300+); able to speak fluently and express one's own opinions.		
Whether you have been to China (if so, for how long)			

In the pilot study, I divided the participants' language proficiency into three levels: L (low), M (middle), and H (high), based on their literacy level. Each multiple-choice option was a short description of the level. Participants were to make their own choice and assess their own language level. The final metadata collection form used in MICICL pilot study is shown in Figure 3.1. All the participants filled the metadata collection form in English or Italian.

To ensure that the data collection tasks matched learners' abilities and yielded more targeted data, I implemented differentiated tasks in the pilot study. Based on learners' self-assessed proficiency in the metadata form, they were divided into three levels—Low (L), Medium (M), and High (H)—and were assigned different tasks accordingly.

Furthermore, during classroom observations and the pilot study, several teachers and students reported that learners at the initial stage of learning often lack the ability to write Chinese characters and tend to rely mostly on Pinyin when writing. In light of this, I did not require L-level learners to complete a handwritten task. Instead, I adopted an oral question-and-answer format, during which they were allowed to draft their responses in writing; however, these drafts were not included in the final corpus due to the predominance of Pinyin. For M-level and H-level learners, I adopted the “writing–reading” multimodal chain described in Section 3.1: learners first completed the writing task and then read aloud the content they had written.

The specific language data collection task³⁸ was as follows:

- L-level (Low Proficiency) Learners: this group was not assigned a writing task, only an oral task. The task format involved answering simple questions, requiring complete sentence responses. Questions included:
 1. *Please introduce your family.*
 2. *Please tell us something about Chinese culture/festivals/traditions that you know.*
 3. *Choose one of the following questions and answer it:*
 - *Please tell us about your favorite sport.*

³⁸ All the questions in the language data collection task were listed in English

- *Please tell us about your favorite place.*
- *Please tell us about your usual hobbies.*
- M-level (Medium Proficiency) Learners:
 Learners in this group still had limitations in Chinese character recognition and writing ability. The task required them to first write complete sentence answers and then read them aloud. The questions were the same as those for the L-level, but they had to be completed in written text form, and the content written had to be read aloud afterward:
 1. *Please introduce your family.*
 2. *Please tell us something about Chinese culture/festivals/traditions that you know.*
 3. *Choose one of the following questions and answer it: favorite sport / favorite place / usual hobbies.*
- H-level (High Proficiency) Learners:
 For learners who self-assessed as advanced, a more complex task was required to reflect their higher language level. This group was required to complete free writing of over 150 characters. The topics were:
 1. *Please talk about your impression of China.*
 2. *Please talk about your dream.*

After completing the writing, the learners were required to read the text they had written aloud. The recording was saved as the audio modality data.

3.4.4 Findings, problems, and adjustment strategies

The pilot study has shown that the overall process and data collection methods are feasible. These procedures were therefore retained for the main study. In addition, the mixed data collection approach used in the pilot study—combining face-to-face sessions, online sessions, and self-completed tasks—was also shown to be viable. The level of interaction achieved in online meetings was comparable to that of in-person sessions, and participants provided positive feedback, indicating that remote data collection is feasible both technically and in terms of participant experience. The data quality from self-completed tasks also met the

required standards. Based on these results, the main study formally adopted the 3-method data collection strategy. While maintaining face-to-face sessions as the primary mode, online meetings and self-completed options were offered to learners located in other regions or facing time constraints. This approach not only broadened participant coverage but also enhanced the continuity and flexibility of subsequent longitudinal tracking.

However, the pilot study also revealed two design issues in the data collection content: Firstly, language proficiency information collection. In metadata collection, the pilot phase used learner self-assessment to determine language proficiency. This choice was made due to practical limitations: the HSK exam is not mandatory in Italy, student coverage is low, and the exam only tests listening, reading, and writing, lacking a speaking component, thus failing to fully reflect actual language proficiency. However, in practice, the self-assessment showed a clear bias. However, in practice, the self-assessment results showed clear signs of bias. Learners' self-reported proficiency levels were not always consistent with several observable performance indicators available during the pilot phase. For instance, some participants who selected the Medium (M) level were able to engage in relatively fluent spontaneous conversations in Chinese and to produce written texts with coherent structure and limited lexical support. Despite these indicators of relatively advanced productive ability, these learners tended to avoid self-identifying as High proficiency. As a result, self-assessment responses led to a highly skewed distribution across proficiency groups: only 2 out of 30 participants classified themselves as High proficiency, while the majority clustered in the Medium level. This imbalance reduced both the discriminative power of proficiency-based grouping and the interpretability of proficiency as a metadata variable. To enhance the objectivity of proficiency classification, the main study therefore adopted a more structured questionnaire-based approach. Learners' proficiency levels were derived from composite numerical indices calculated from multiple variables, including length of Chinese study, frequency of language use, and learning context, rather than from self-reported categorical labels alone.

Secondly, in the initial stage of the study, a tiered task design was adopted for language

data collection. In the pilot study, low-level (L) learners completed only oral tasks, whereas intermediate-level (M) and high-level (H) learners completed reading and writing tasks of varying difficulty. The original intention of this design was to prevent situations in which learners were unable to complete the task due to insufficient writing ability, and to align task demands with learners' proficiency levels. However, practical outcomes revealed clear limitations in terms of cross-sectional comparability. Specifically, the L group lacked “reading–writing” data that could be compared with that of the M and H groups, making it impossible to evaluate their output on the same dimension. Meanwhile, the writing texts produced by the H group were significantly longer than those of the other groups, and such large discrepancies in text length undermined consistency. In addition, the fundamental structural differences among the three task types made it difficult to maintain uniformity in both the form and quantity of the corpus.

For these reasons, the level-based task design was removed in the main study, and all learners were assigned the same task. The task instructions explicitly stated that if learners encountered characters they could not write, they could use Pinyin, English, or Italian as substitutes. In fact, even when learners use Pinyin or foreign words, their writing samples still possess research value, as these substitution patterns reveal cognitive strategies and transfer phenomena within the Chinese writing system (see Sections 5.4 and 5.5 for detailed analyses). Furthermore, to avoid text-length variability affecting group comparison, explicit word-count requirements were established for writing tasks in the main study to ensure basic consistency. The results of the main study also demonstrated that learners' actual abilities often exceed researchers' expectations; overly restrictive assumptions embedded in task design may instead limit learners' natural output and obscure their true linguistic performance.

In conclusion, the pilot study revealed two key issues: collection format design and proficiency assessment. The corresponding adjustments ensure improvements in the formal corpus construction regarding operability, comparability, and sustainability. The pilot study not only verified the feasibility of the overall procedure but also provided a direct empirical basis for optimizing the formal data collection plan for the MICICL corpus.

3.5 Main data collection

The pilot study indicated that although certain adjustments were required in the design of metadata collection and language data elicitation tasks, the practical procedures—including participant recruitment, meeting organization, and meeting implementation—proved to be fully workable. Therefore, the main study continued to follow the procedures outlined in Section 3.4.2, and the specific operational steps will not be reiterated in this subsection.

3.5.1 Students recruitment and data collection timeline

After the pilot study helped refine and adjust the data collection procedures and content, I officially launched the MICICL data collection in September 2023. Between September and October 2023, the first round of learner recruitment was conducted. The recruitment was limited to the University of Bologna and publicized through in-class presentations, campus posters, and online dissemination via university-affiliated media channels. Participation was entirely voluntary. A total of 31 students registered, and 25 students eventually took part in the first round of data collection.

The first round of data collection took place from November 2023 to June 2024, comprising six sessions: November 2023, December 2023, February 2024, March 2024, April 2024, and May 2024. Since participation was voluntary and subject to students' personal schedules, the number of participants varied slightly across sessions.

In November 2023, an additional classroom-based data collection was also carried out with the assistance of the Chinese language instructor at the Department of Interpreting and Translation, University of Bologna. This supplementary classroom corpus included 13 students, 6 of whom also participated in the first round of MICICL data collection, while 7 students participated only in this classroom session.

After the first round, due to participant attrition and absences during the data collection process, a second round of recruitment was launched in September 2024 to increase the data volume and maintain sample balance. The recruitment methods and scope remained the same as in the first round. As a result, 7 new learners joined, and 5 students from the first round volunteered to participate again. The second round of data collection was conducted from

October to December 2024, comprising two sessions (October and December). Similarly, the actual number of participants in each session varied slightly due to the voluntary nature of participation and differences in academic schedules.

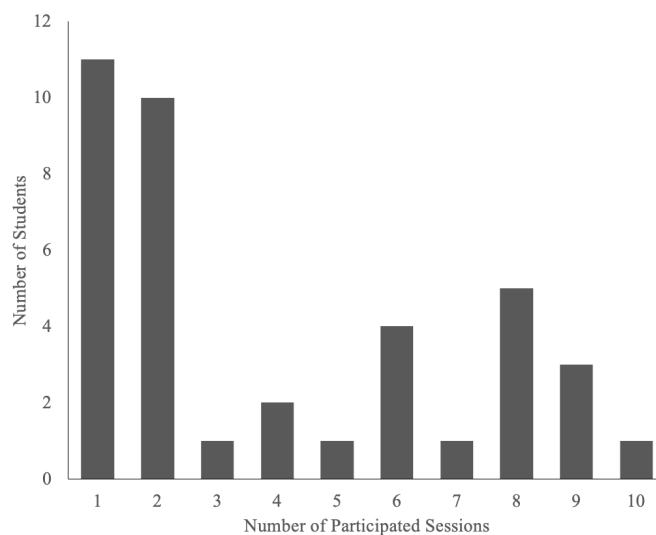
Overall, the main data collection phase of MICICL involved 39 students and spanned from November 2023 to December 2024, covering a total of ten formal data collection sessions.

It should be noted that among the 25 participants in the first round, 13 had already taken part in the pilot study. After being informed of the research objectives and procedures, all 13 students agreed to include their pilot study data in the final MICICL corpus and signed an additional consent form. Among them, 3 learners self-assessed as beginners during the pilot phase and provided only oral data. Considering MICICL's requirement to include both handwritten and reading-aloud data, their data were ultimately excluded. Consequently, pilot data from 10 learners were incorporated into the final corpus.

In summary, the MICICL corpus includes data from 39 learners, collected over a period extending from May 2023 to December 2024, across a total of 11 data collection sessions. Learners were assigned participant codes ranging from 01 to 39. Figure 3.2 illustrates the overall distribution: 11 students participated only once, 10 students participated twice, while 15 students took part in five or more sessions, providing a foundation for potential longitudinal research.

Figure 3.2

Distribution of students by number of participated data collection sessions



3.5.2 Student metadata collection results

In the main study, the collection of learners' metadata followed a similar structure to that of the pilot study (see Section 3.3), consisting of two parts: biographical information and language proficiency information. For biographical information, the items collected are shown in Figure 3.3.

Figure 3.3

Biographical information collection form

I. Biographical Information

Name and Surname		Gender	☐ male ☐ female ☐ other
Country of birth		Mother tongue(s)	
Current course level	☐ Bachelor ☐ Master ☐ PhD	Year of study	☐ First year ☐ Second year ☐ Third year
Degree course			
Home language (the most commonly used language in daily life)			
Which other languages do you know? (and at what level)			

Regarding language-related information, as discussed in Section 3.3, the self-assessment design adopted in the pilot study tended to lead learners to underestimate their own proficiency due to uncertainty. Therefore, in the main study, a more objective method was adopted to evaluate learners' proficiency levels. The evaluation referred to the multi-measure questionnaire based on the access testing materials provided by the Bilingual Language Profile³⁹ (BLP) website (Birdsong et al., 2012), which is an open-access instrument for assessing language dominance among bilinguals. It provides a systematic questionnaire and scoring framework that allow for a relatively objective assessment.

The BLP questionnaire consists of Biographical Information and four modules: Language History, Language Use, Language Proficiency, and Language Attitudes. These four modules

³⁹ <https://blp.coerll.utexas.edu/>

contain 19 questions in total, requiring respondents to give numerical responses on a Likert-type scale. Each response corresponds to a numerical value on the scale. Within each module, scores for each language are calculated according to specific criteria. To ensure equal weighting across modules in the overall language index, the module scores (calculated separately for each language) are multiplied by a correction coefficient, and the adjusted module totals are then summed to obtain the overall language score. A higher score indicates stronger language dominance.

Although the BLP was originally designed to evaluate bilinguals' dominant language, its framework can also be adapted to assess second or foreign language proficiency. For MICICL, I mainly referred to the Language History and Language Proficiency modules and modified several questions as needed, ultimately developing an eight-item MICICL Chinese Proficiency Questionnaire for learners. The specific questions are presented in Figure 3.4, and the complete metadata collection form is provided in Appendix A.

Figure 3.4

Language proficiency information collection form

II. Language History
In this section, we would like you to answer some factual questions about your language history by placing a check in the appropriate box.

1. How many years of classes (grammar, history, literature etc.) have you had in Chinese (primary school through university)?
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17 ☐ 18 ☐ 19 ☐ 20+

2. In an average week, how many hours do you spend on learning/exposure to Chinese? (Including taking classes, self-study, watching Chinese movies, etc.)
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17 ☐ 18 ☐ 19 ☐ 20+

3. How many months have you spent in a country/region where Chinese is spoken?
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17 ☐ 18 ☐ 19 ☐ 20+

III. Language Proficiency
In this section, we would like you to rate your language proficiency by giving marks from 0 to 6.

4. How well do you speak Chinese? *0=not well at all*
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 *6=very well*

5. How well do you understand Chinese? *0=not well at all*
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 *6=very well*

6. How well do you read Chinese? *0=not well at all*
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 *6=very well*

7. How well do you write Chinese? *0=not well at all*
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 *6=very well*

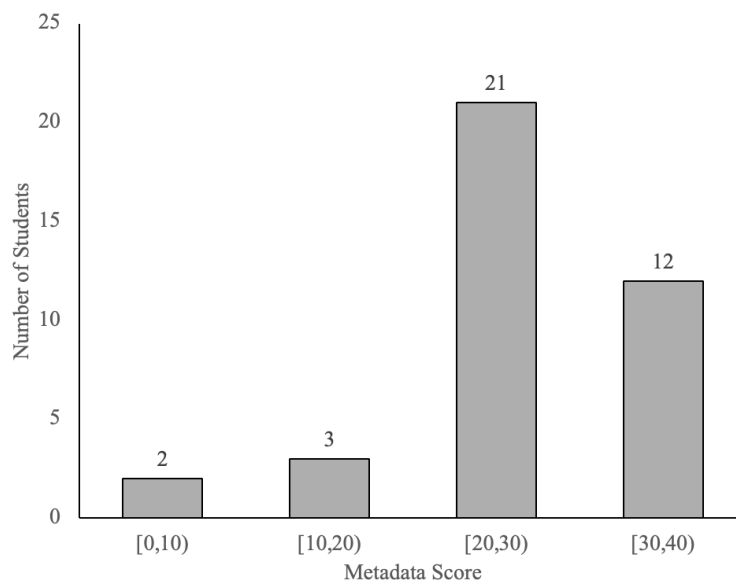
8. What's your HSK test level?
*(Each number represents a different HSK level.
 0 means that you have never taken the HSK test.)* ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

Among the eight items, Question 8 (“What’s your HSK test level?”) was used for reference only and not included in the final scoring, as not all students had taken the HSK exam. Since the scoring results of this questionnaire were intended solely for internal use within MICICL and not for external comparison with other corpora, no correlation coefficient adjustments were applied. The total sum of the numerical results from the first seven questions was used as the criterion for grouping students by proficiency.

The final scores of the 39 students ranged from 5.5 to 37 (see Figure 3.5 for specific score intervals distribution). Since the number of students falling within the 0–10 and 10–20 score ranges was relatively small, these two intervals were combined into the beginner group, while students scoring 20–30 were classified as intermediate, and those scoring 30–40 were classified as advanced.

Figure 3.5

Distribution of students’ metadata score



Among the 39 students, 6 were classified as beginner-level, 21 as intermediate-level, and 12 as advanced-level learners. Details of the proficiency grouping are provided in Appendix B.

It should be noted that:

- These three proficiency groups are used exclusively for internal comparative analysis within MICICL. The groupings represent relative proficiency levels rather than absolute

ones and should not be directly compared with external proficiency standards or other corpus classification systems.

- During the biographical information collection phase, participants' current academic degrees were recorded. Among the 39 participants, 30 were undergraduate students, 7 were master students, and 2 were doctoral students. However, academic degree did not directly correspond to language proficiency. Some undergraduate students, for instance, had lived in China for extended periods or received intensive language training, and their Chinese proficiency even surpassed that of some master students. Therefore, the degree information is included only as background reference rather than as a criterion for defining proficiency levels.
- Furthermore, as discussed in Section 3.2, MICICL integrates both quasi-longitudinal and longitudinal design features. Since no previous research has clearly determined the time span required for a qualitative restructuring of learners' interlanguage systems, the data collected over a period of 20 months in this study can still be validly analyzed through a quasi-longitudinal design. This approach allows for relative comparisons across proficiency groups and remains methodologically sound and comparable.

3.5.3 Language data collection results

Except for one classroom-based data collection conducted in November 2023 under the guidance of the course instructor, all other nine sessions of the main study were organized independently by myself outside regular class time. The data collection followed the same procedure as in the pilot study (see Section 3.4), primarily through in-person meetings, supplemented by online meetings and self-administered tasks.

A total of eleven rounds of data collection were conducted in this study, each assigned a specific code in file names for clarity and consistency: the pilot study conducted in May 2023 was coded as 00, the classroom-based task conducted in November 2023 was coded as ZY, and the nine subsequent main study sessions conducted between November 2023 and December 2024 were coded sequentially from 01 to 09.

The design of the pilot study task topics has been described in Section 3.4. The classroom-

based task (ZY) was designed by the Chinese language instructor, while the nine main data collection tasks were designed by myself. Each task featured a distinct topic, but all followed the principle of maintaining openness within a constrained theme, so as to prevent over-concentration of certain Chinese character types in the corpus. The tasks also covered different genres to examine learners' Chinese language abilities from multiple dimensions, see Table 3.1. Please also note that because the “Lower-level” oral response task used in the pilot study was not included in MICICL, only the task topics involving both handwritten and reading-aloud components (for Medium-level and High-level) are listed here for the pilot study session 00. All task instructions and questions were presented in English to ensure comprehension and consistency across participants. However, some task prompts contained isolated Chinese lexical items (e.g., proper names such as 小红) as part of the task content rather than as instructional language.

Table 3.1

Task topics of each session

Session	Topic
00	Medium-level proficiency: Write down the answers and read them aloud (please write complete sentences) (1) Please introduce your family. (2) Please tell us something about Chinese culture/festivals/traditions that you know. (3) (choose one): - Please tell us about your favorite sport - Please tell us about your favorite place - Please tell us about your usual hobbies High-level proficiency: Write a text about the topic and read it aloud (please write at least 150 characters) (1) (choose one): - Please talk about your impression of China

	- Please talk about your dream
01	Please select 5-7 words from the following 7 words and write a short passage: 起床 高兴 办法 可能 以前 鸡蛋 猫
02	Your friend 小明 plans to travel to Italy for a week. He wrote you an email asking for suggestions on visiting, accommodation, transportation, food, etc.
03	Your friend 小红 is coming to Forlì/Bologna to visit the campus. Please take him around and introduce him to the campus and its surroundings. How are you going to introduce it to him? Please write an article of no less than 200 characters.
04	Your teacher has asked you to do a sharing session in class and share one of your favorite movies/ books with the class. Please write a speech describing the plot of this movie/book and why you like it. Please write no less than 200 characters.
05	What is your favorite season, spring, summer, fall or winter? Why do you like this season? Do you have any fond memories from this season? Please write an article and tell me.
06	Your friend 小红 has recently had arguments with her parents due to different opinions. Please write a letter to comfort her. You can tell her whether you have had arguments with your parents, how you resolved them in the end, and give her some advice on getting along with parents. Please write no less than 200 characters.
07	Summer has passed. How did you spend this summer vacation? Did you meet any interesting people? Did any interesting things happen? Please write an essay to describe your summer vacation. Please write no less than 200 characters.
08	There must be someone who has had a major impact on your life. It could be a family member, a friend, or even a famous person like a singer or scientist. What happened between you two? Explain your relationship with them and how they've influenced your past, present, and future. In what ways this person has shaped your life and decisions?
09	Write an essay exploring how technology has influenced your life and the world around you. Discuss the positive or negative effects of technology and share your personal experiences or perspectives. How has technology shaped the way you live, you learn and so on?
ZY	Please introduce an Italian city — you may talk about its history, famous places, and other interesting aspects.

Before each writing task, I provided participants with some general instructions:

- i. If they could not recall or write a character, they were encouraged to substitute it with Pinyin, English, or Italian. This was to ensure the continuity and authenticity of the data.
- ii. They were asked to write at least 200 characters. Although not all participants met this requirement, it effectively served as a reminder to produce texts of substantial length.

After completing the writing task, students were given time to familiarize themselves with their own written texts in preparation for the reading task. Figure 3.6 presents an example of the task description provided to learners during the data collection process.

Figure 3.6

An example of a task description for students

Task Description

This task comprises two parts: writing and reading. Please note that this is not an examination, and we will not grade your language samples. It is simply a data collection process, and any form of error is acceptable. For our research, understanding and analyzing potential errors in your language sample are crucial. Therefore, we suggest the following:

- ***Avoid using any auxiliary tools*** such as dictionaries or translation software while completing the task.
- ***Present your true proficiency as accurately as possible.***

Thank you!

Writing:

Summer has passed. How did you spend this summer vacation? Did you meet any interesting people? Did any interesting things happen? Please write an essay to describe your summer vacation. Please write no less than 200 characters.

Please note:

1. Please complete by handwriting on white paper
2. If you forget how to write a character or word, you may use pinyin/English/Italian as a substitute.

Don't worry, any type of error is acceptable, so don't be afraid to make mistakes!

Reading:

Please read aloud the passage you've written in a quiet environment and record it.

The number of participants and the collected handwritten and reading-aloud samples from the eleven data collection sessions are shown in the Table 3.2. A total of 155 handwritten samples and 153 reading samples were collected. Two reading recordings were missing due to participants' non-submission or file corruption. The number of samples provided by each learner is presented in Appendix C. The number of language data samples across different proficiency groups is shown in Table 3.3.

Table 3.2

Number of students and collected samples of each session

Session Code	Number of Students	Number of Writing Samples	Number of Reading Samples
00	10	10	10
01	20	20	19
02	17	17	17
03	17	17	17
04	18	18	18
05	16	16	16
06	16	16	16
07	5	5	5
08	12	12	12
09	11	11	11
ZY	13	13	12

Table 3.3

Number of language samples across different proficiency levels

Proficiency level	No. of students	No. of writing samples	No. of reading samples
Beginner	6	25	24
Intermediate	21	77	77
Advanced	12	53	52

In addition, Table 3.4 presents the number of data collection sessions each student participated in and their continuity status (the shaded cells indicate the sessions in which the student participated).

Table 3.4*Distribution of learners' participation frequency and continuity in data collection*

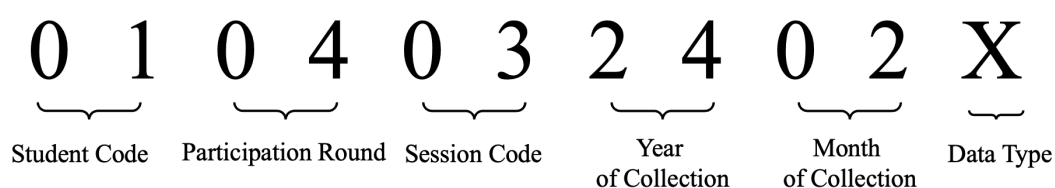
	Session code										
Student code	00	01	02	03	04	05	06	07	08	09	ZY
01											
02											
03											
04											
05											
06											
07											
08											
09											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											
33											
34											

35											
36											
37											
38											
39											

To facilitate data management and retrieval, and to provide anonymized metadata in a more accessible form, each handwritten sample and reading recording was assigned a unique 11-character code. This code encodes five types of information: the learner code, the participation round, the session code, the year and month of data collection, and the data type. As shown in Figure 3.7, the final character indicates the data type: X denotes handwritten data, and D denotes reading-aloud data, corresponding to the initial letters of the pinyin for “写 (xiě, writing)” and “读 (dú, reading).”

Figure 3.7

Structure of data codes used in MICICL



Chapter 4 Data Processing and Corpus Construction

4.1 Chapter overview

In the process of constructing an interlanguage corpus, data processing refers not merely to technical, automated operations, but rather a systematic procedure encompassing transcription, annotation, and correction. The core objective of this process is to transform learners' raw linguistic outputs from unstructured data (e.g., handwritten samples and voice recordings) into computer-readable, statistically analyzable textual resources. Upon this foundation, linguistic information is added to reveal the patterns and the developmental trajectory of the learner's interlanguage system. These operations can be executed manually or with the assistance of computational tools.

For corpora containing handwriting or reading-aloud data, transcription is the initial step, converting handwritten texts and oral utterances into digitized, retrievable text. This provides the necessary foundation for subsequent corpus statistics and multimodal comparisons. In the construction of the MICICL corpus, only through standardized transcription, meticulous annotation, and evidence-based correction can the corpus exhibit comparability and interpretability in subsequent analyses. Concurrently, this process determines the corpus's pedagogical applicability.

Therefore, the term "Data Processing" discussed in this chapter is not limited to narrow computational processing, but rather refers to a multi-layered linguistic refinement process guided by linguistic theories and aimed at achieving the analyzability of the corpus. This chapter will unfold in the following sections:

- Section 4.2 outlines the design of MICICL's data processing framework and the selection of tools used in this process.
- Sections 4.3 and 4.4 detail the MICICL handwritten and reading-aloud data's transcription, annotation and correction principles, along with the final results.

Through the content above, this chapter aims to address the core question: How can raw handwriting and reading samples be systematically processed into research resources that possess linguistic interpretability, computational operability, and pedagogical applicability within a learner-centered multimodal interlanguage corpus?

And please note that due to time and technical limitations, only the transcribed and annotated textual data are made publicly available⁴⁰ at present, while the original image and audio data can be accessed upon request.

4.2 MICICL data processing framework and tools

4.2.1 MICICL data processing framework

MICICL adopts a fully manual data processing workflow, applies different transcription strategies to its two data types, and adopts a three-level layering structure consisting of transcription, annotation, and correction, as detailed below.

Firstly, data processing was carried out through a manual workflow. With the rapid development of information processing technology, there is growing interest in the automation of corpus transcription and annotation tasks. Currently, various automated tools and models exist for Chinese handwriting and speech data, capable of achieving relatively high transcription and annotation accuracy. However, studies have shown that while these systems demonstrate strong performance when processing native speaker data, their accuracy declines when applied to learner language. Rosen (2017) and Abel et al. (2014) suggest that automatic annotation can maintain an acceptable level of accuracy for the written texts of intermediate and advanced learners. In contrast, other studies (Geertzen et al., 2013; Stymne et al., 2017) report that reliability decreases substantially when dealing with beginner learner corpora. A key reason is that most of these tools are developed and trained on native speaker corpora with relatively stable structures and normative grammar (Meurers, 2015), which limits their ability to handle non-normative or variable learner language.

⁴⁰ https://zenodo.org/records/19973878?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6Ijg3MGQ0OTcxLTdkOGQtNGY0YS1iNzczLWQ1ODlmYTc1NjQ0ZiIsImRhdGEiOnt9LCJyYW5kb20iOiJzMzQ5YzhiZTI5OWM5ZjQ0YmMyYTMzZjc4YWUiZmZlOSJ9.aKvU3YzCMaPkvDXPkFTN1rbc8X_UwQJ4LrOliDD3w_TvrFUEfBAjnUjA31cqGK7y3k9mmHuVLby3H6GVvA

The linguistic forms produced by Chinese interlanguage learners are complex and diverse, showing significant differences from native speakers in character structure, phonetic realization, and the sound-form correspondence. These non-standard features pose numerous challenges to the performance of automated tools in interlanguage corpus processing. Taking data transcription as an example: existing OCR systems primarily use printed and standard handwritten samples for training, while ASR systems are trained on native Mandarin or specific dialect speech. Learners' orthographic errors in writing, and deviations in initials, finals, and tones during reading-aloud, often lead to frequent recognition errors by the systems. This not only fails to reduce manual workload but also introduces significant noise data, undermining the reliability and research value of the corpus.

Given the complexity of learner language and the limitations of automated recognition, MICICL adopts manual transcription and annotation during the data processing phase. Although manual processing is time-consuming (and costly), it maximizes the authenticity of the corpus data and the accuracy of semantic restoration, ensuring that the learner's linguistic characteristics are accurately represented. Particularly in the initial phase of corpus construction, manual processing is not only necessary for quality assurance but also lays a reliable foundation for the subsequent establishment of annotation norms. Currently, no publicly available Chinese interlanguage corpus relies entirely on automated transcription.

Secondly, data type-specific transcription. Based on this principle, MICICL applies different transcription methods to its two data types: handwritten data is transcribed using Chinese characters, while reading-aloud data is transcribed using Pinyin with numerical tone marks (numbers 1–4 indicating the four lexical tones, with 0 used for the neutral tone). This differentiation ensures a more accurate representation of the unique linguistic features of each modality, maintaining the authenticity and comparability of the data form.

The decision to not use Chinese characters for transcribing reading audio is based on the following key points: 1) at the theoretical level, interlanguage in language acquisition research is viewed as an independent language system with its own rules and features. Recording the learner's output with Chinese characters often unconsciously pulls the learner's language closer

to the target language norms, thereby obscuring the unique characteristics of the interlanguage itself. For example, if a learner produces an inappropriate pronunciation, recording it with the corresponding Chinese character might align with the standard orthographic form, thereby concealing the error trace. Adopting Pinyin, however, allows for a neutral record of the pronunciation that does not directly map onto standard orthography, thus preventing the over-normalization of the interlanguage. Annotation with Chinese characters is unnecessary in spoken corpus transcription because the transcription primarily reflects the spoken output itself, not the written form. 2) at the practical level, Pinyin transcription facilitates the detailed recording of phonetic nuances. Learner phonological errors (such as tone errors or phoneme confusion) can be clearly captured through Pinyin and tone markings. If the transcriber encounters an uncertain or unintelligible segment and subjectively guesses the “correct” Chinese character based on the context, this violates the principle of authenticity and fails to remain faithful to the original data. A corpus constructed in this manner would consequently lose its value (Hu & Wang, 2011).

The choice of Pinyin over the International Phonetic Alphabet (IPA) is based on the target audience and practical utility of the corpus. MICICL primarily serves language teachers and learners; the Pinyin system is sufficient in teaching and research to represent syllabic and tonal features. While IPA is more detailed, it is complex to operate and less readable. If the corpus adopted the IPA system, general users would still often need professional phoneticians to re-listen and interpret the data to extract meaningful error information (Wang & Li, 2001; Zeng, 2008). This high analytical threshold does not align with MICICL’s pedagogical orientation and operability requirements.

Last, three-layer language information format. In addition to type-specific transcription, MICICL adopts a three-level layering structure consisting of transcription, annotation, and correction. First, transcription provides the raw data. Second, annotation is performed on the basis of the raw transcription, adding labels to record the learner’s linguistic features and error types. Finally, correction, which involves providing a hypothetical correct version based on grammatical or semantic plausibility, is carried out while retaining the annotation information.

This layered system is designed to balance readability and flexibility. Firstly, the layered structure allows users to select different levels of text for reading and retrieval according to their research needs, thereby improving the readability of the processed text (MacWhinney, 2000). Secondly, the design reserves space for future revisions and additions, ensuring the corpus construction remains open and extensible. Zhang & Cui (2013, 2022) proposed a “building block” dynamic construction strategy for corpora, and Granger (2003, p. 467) emphasized that the error annotation system should be flexible to allow for subsequent additions and revisions to the annotation layers.

It is crucial to note that the annotation and correction process for learner corpora is inherently subjective. When discussing the terms “error annotation” and “correction annotation” in interlanguage corpus research, Volodina et al. (2019) pointed out that the operation performed by researchers during correction is not an objective reproduction of the learner’s language behavior, but rather an interpretive, hypothetical act. They state: “what we are annotating is not the objective production of the learner, but the hypothesis made by the researcher based on their own judgment regarding the potential way it could be rewritten” (Volodina et al., 2019, p. 83). In other words, any correction result is only one of many possible interpretations; other researchers may reasonably propose different judgments. Leech (1993) also theoretically argued for the subjectivity of annotation, stating that annotation is essentially an understanding and interpretation of linguistic phenomena; thus, differences in annotators’ linguistic knowledge, experience, and analytical perspectives may lead to varied results. Precisely because of this, preserving the original transcribed data, untouched by any annotation or correction, is of paramount importance in the construction of MICICL. This not only ensures the traceability and methodological transparency of the corpus but also provides ample room for future researchers to propose new annotation and correction schemes under different theoretical frameworks.

Based on this, MICICL’s design ensures the original transcribed data is preserved independently, with no annotation or correction applied at the base layer. This approach guarantees data traceability and offers significant flexibility and extensibility for future

researchers to conduct re-annotation and re-correction under different theoretical frameworks.

4.2.2 MICICL data processing principles

Transcription: Principle of authenticity

In the data transcription layer, MICICL strictly adheres to the principle of authenticity, whereby learners' original linguistic features—including errors, pauses, repetitions, phonetic ambiguity, and non-standard character use—are preserved to the greatest extent possible. Rather than aiming at “correction” or “normalization”, transcription seeks to faithfully reproduce learners' actual linguistic output, avoiding information loss or distortion caused by formatting, filtering, or substitution. As the machine-readable textual representation of multimodal data, transcription constitutes the primary data layer and provides the foundation for subsequent analyses of written and spoken performance (Breiteneder et al., 2006; Wichmann, 2008). Consequently, transcription quality is crucial for ensuring the reliability and usability of the corpus (Gablasova et al., 2019).

Annotation: Principles of practicality and logical coherence

In learner corpus research, annotation has always been one of the core steps in data processing. Its purpose is not only to identify and classify learner errors but also to provide a systematic description of the learner's linguistic characteristics.

Regarding the question of what should be annotated in (Chinese) interlanguage corpora, two main academic positions can be identified. One view advocates comprehensive annotation of linguistic attributes beyond errors, aiming to maximize corpus functionality and research potential. This approach typically involves systematic annotation at multiple levels, including character, word, phrase, sentence, discourse, stylistic, semantic, pragmatic, and punctuation levels (Zhang, 2011; Díaz-Negrillo et al., 2010). The other view adopts a more cautious and practice-oriented stance, arguing that the theoretical frameworks supporting semantic, pragmatic, and discourse-level annotation in Chinese linguistics remain insufficiently mature (Leech, 1997; Wilson & Thomas, 1997). As a result, this approach prioritizes feasibility and operability in annotation practice by restricting annotation scope to linguistically well-defined and reliably identifiable phenomena—most commonly formal features and observable errors

at the character, word, and sentence levels—rather than attempting exhaustive annotation without solid theoretical support (Xiao & Zhou, 2014). While the former emphasizes the richness and extensibility of corpus functionality, the latter foregrounds annotation reliability, consistency, and reproducibility in actual corpus construction.

Taking these two viewpoints into consideration, the present study adopts a problem-oriented and demand-driven approach to corpus annotation (Zhang & Cui, 2016; Zhao & Lin, 2019). In the case of MICICL, annotation is designed primarily to support pedagogical applications while maintaining methodological transparency and analytical reliability. Accordingly, the annotation system focuses on phenomena that are both pedagogically relevant and linguistically observable. In addition to error annotation, MICICL includes the annotation of selected learner linguistic features manifested in the textual data, such as non-target-like forms, substitution patterns, and cross-linguistic influences that do not always constitute clear-cut errors but nonetheless reflect learners' interlanguage development. These features are annotated based on explicit and operational criteria, rather than on abstract or theory-dependent linguistic categories. From this perspective, MICICL's annotation system is oriented toward teaching and applied research needs, rather than toward the construction of a comprehensive linguistic ontology. At the same time, the annotation format is designed to remain extensible, allowing for future refinement or expansion as theoretical frameworks and research demands evolve.

In the design of annotation labels (tags), MICICL adheres to the principles of clarity and ease of comprehension. Details of the annotation labels are provided in Sections 4.3.2 and 4.4.3. Annotation codes are designed to be highly associative, allowing corpus constructors and users to readily infer their meanings. In the context of a Chinese interlanguage corpus, using Hanyu Pinyin as annotation codes is more transparent and memorable than relying on abstract alphabetic abbreviations (Zhao & Lin, 2019). For example, incorrectly written characters (错字, *cuòzì*) are annotated as /C, while incorrectly used characters (别字, *biézi*) are annotated as /B, with the tags derived from the initial letters of the corresponding Pinyin forms.

To enhance annotation consistency and reusability, MICICL draws on widely adopted

annotation conventions from existing Chinese learner corpora, particularly the HSK Dynamic Composition Corpus (HSK DEC). Adopting well-established and theory-neutral annotation models helps reduce ambiguity and misuse, facilitates tool development and cross-corpus collaboration, and avoids unnecessary duplication of resources. From a functional perspective, corpora can be broadly categorized as general-purpose or specialized: while general-purpose corpora typically require comprehensive multi-level annotation, specialized corpora focus on in-depth annotation of specific phenomena or research questions (Zhao & Lin, 2019). As a specialized corpus, MICICL selectively adopts and adapts elements of existing general-purpose annotation systems, including HSKDEC, while introducing targeted modifications to meet its specific pedagogical and research objectives.

Correction: Principle of minimal change

In the data correction layer, MICICL follows the principle of minimal change. Rather than modifying the original learner data, corrections are provided in a separate layer that represents a minimally adjusted target-like form. Corrections are limited to the smallest possible scope necessary to address identified errors, while preserving the original semantic intent and structural configuration of the learner's output. Napoles et al. (2017), in summarizing annotation experiences from multiple English learner corpora, noted that the "minimal change" strategy effectively balances readability and originality, resulting in a modified text that is both grammatically correct and faithful to the learner's actual expression. In Chinese learner corpus research, Zhou et al. (2007) also proposed the "simplification principle," advocating for maintaining the original sentence structure as much as possible during error correction, reducing additions, deletions, and substitutions to ensure the corpus retains its authentic characteristics. Similarly, Zhang (2013) emphasized that during correction and annotation, the "original flavor of the corpus should be preserved to the maximum extent," as the value of a corpus lies not only in providing a reference for correct language forms but also in presenting the dynamic development of the learner's language system. Consequently, the minimally modified text can be viewed as a grammatically correct reference variant rather than a re-creation of the learner's language.

Overall, the design of MICICL's data processing methods also adheres to economic efficiency. Corpus processing work is typically high-intensity and time-consuming. As Brezina et al. (2019, p. 119) noted in the the Trinity Lancaster Corpus⁴¹ project report, the transcription phase alone consumed over 3,500 hours of manual labor, with subsequent quality checks and data verification requiring substantial additional effort. Li (2021) also pointed out that corpus construction should balance time, cost, and skill input, ensuring the depth of correction and processing matches the project goals.

4.2.3 ELAN as the transcription tool

In the field of multimodal corpus construction, researchers have developed various tools for processing audio, video, and textual data. Current mainstream multimodal transcription and annotation software includes Praat⁴², Anvil⁴³, ELAN⁴⁴, Exmaralda⁴⁵, and MacVisSTA⁴⁶. Praat (Boersma & Weenink, 2007) is renowned for its acoustic analysis capabilities, suitable for detailed phonetic research on speech signals, but its functionality for handling multimodal synchronization (such as parallel presentation of images, text, and audio) is limited. Anvil is a lightweight multimodal annotation software focused on video annotation and visualization, supporting multi-layer timeline annotation and custom label systems in a single window, but its simple dependency structure design may not satisfy the needs of complex hierarchical annotation (Abuczki & Ghazaleh, 2013). Tools like Exmaralda and MacVisSTA perform well in spoken transcription and discourse-level analysis, but still face compatibility and stability limitations in supporting multimodal parallel alignment and multi-layer dependency annotation. In summary, existing tools possess specific advantages in their respective domains, but there are shortcomings in terms of multimodal synchronization, hierarchical dependency annotation, and extensibility. The construction of the MICICL corpus requires the processing of multimodal data, cross-modal comparison, and multi-layered structured retrieval within a

⁴¹ <https://cass.lancs.ac.uk/trinity-lancaster-corpus/>

⁴² <https://www.fon.hum.uva.nl/praat/>

⁴³ <https://www.anvil-software.de/>

⁴⁴ <https://archive.mpi.nl/tla/elan>

⁴⁵ <https://exmaralda.org/en/>

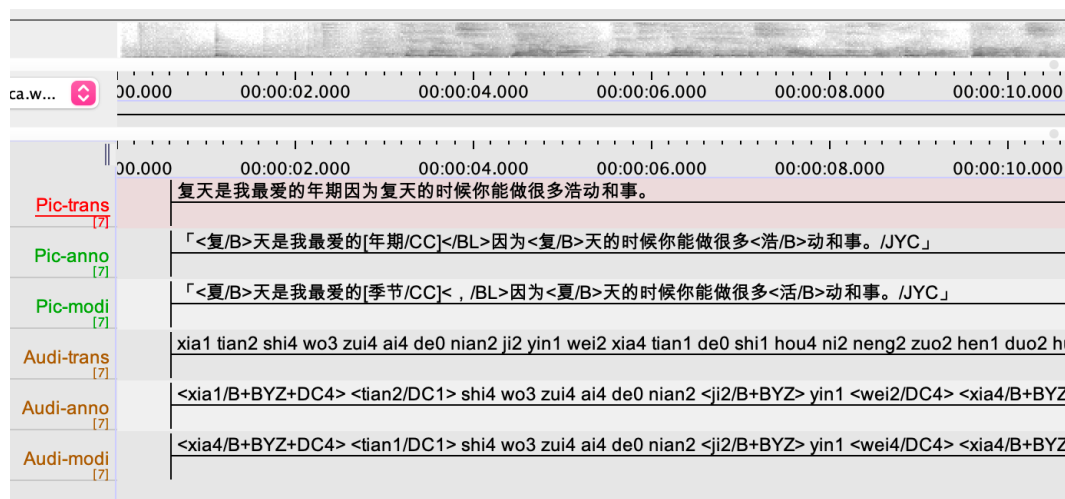
⁴⁶ <https://sourceforge.net/projects/macvissta/>

single environment. Based on this requirement, ELAN (Brugman & Russel, 2004) emerges as the most suitable choice for MICICL’s objectives.

ELAN is a multimodal corpus annotation platform developed by the Max Planck Institute for Psycholinguistics (MPI) in the Netherlands. It supports multi-layer annotation, dependency relationship settings, and multimedia synchronization. Its structured hierarchical system allows users to establish parent-child or sibling dependency relationships between multiple annotation layers, enabling clear visualization of complex corpus features (Brugman & Russel, 2004). Additionally, ELAN offers a highly customizable annotation scheme, allowing users to freely add or modify label types according to research goals, balancing generality and flexibility, making it a mainstream tool for international multimodal corpus construction (Liu, 2019, p. 2). Although ELAN does not currently support the import of image data, it remains overall the most suitable tool for MICICL’s needs.

Figure 4.1

Screenshot of the ELAN interface during the MICICL data processing



Specifically, MICICL uses ELAN to synchronously perform the transcription, annotation, and correction of both handwritten and reading-aloud data. Since almost all data in MICICL is a one-to-one correspondence between handwriting and reading-aloud, processing both modalities within the same ELAN file is possible. Figure 4.1 shows a screenshot of the ELAN interface during the MICICL data processing.

For each set of handwritten and reading-aloud data produced by a learner in a single data collection session, a new ELAN file is created to process this set of data (although only the reading audio can be imported; the handwritten data image must be viewed separately). Six tiers are constructed within the ELAN file: three for the handwritten data and three for the reading-aloud data. The three handwritten data tiers are named Pic-trans, Pic-anno, and Pic-modi, respectively used for transcription, annotation, and correction, with Pic-anno and Pic-modi having a child-parent relationship with Pic-trans. Similarly, the three reading-aloud data tiers, Audi-trans, Audi-anno, and Audi-modi, are used for transcription, annotation, and correction, respectively, with Audi-trans being the parent tier for the other two. Here, the tier label “modi” (modification) is retained for consistency with the actual ELAN file structure; conceptually, however, this tier corresponds to the correction layer described in this study and does not involve any alteration of the original transcription or annotation tiers. Transcription of handwritten and reading-aloud data is performed separately on the Pic-trans and Audi-trans tiers. Annotation is carried out on the Pic-anno and Audi-anno tiers, where raw data is manually entered with various annotation labels. Minimal correction of errors and non-standard features is performed on the Pic-modi and Audi-modi tiers while retaining the annotation labels.

4.2.4 Limitations

Considering time and human resource costs, while MICICL aims to comprehensively represent the learner’s interlanguage form, certain linguistic features have not been subjected to transcription, annotation, or other processing. These include cross-outs and corrections made by students in the handwritten data, and pause duration in the reading-aloud data. The preservation of the handwriting image and the reading audio provides the possibility for these features to be represented in machine-readable form in the future. Additionally, the entire process, from transcription to correction, was carried out by me. An internal secondary review was also conducted by me to check for consistency and obvious errors. While this procedure helped improve internal coherence, the absence of independent transcribers or annotators means that absolute accuracy cannot be assumed. Future work could benefit from the involvement of additional annotators, enabling inter-annotator comparison and further

enhancing corpus reliability. Users are welcome to report any errors they find during use. Finally, as Volodina et al. (2019) stated, any error annotation and correction performed are merely one hypothetical explanation proposed by the researcher. Different error and correction results may exist for the same linguistic sample, and MICICL merely offers one interpretation.

4.3 MICICL handwritten data processing rules and outcomes

4.3.1 Complexity of handwritten data

In alphabetical writing systems, the differences presented by handwriting and typing in learner studies are relatively minor, as both modes exhibit high consistency in form and expression. However, for the Chinese language, a significant divergence exists between handwriting and typing. Modern Chinese predominantly uses Pinyin computer input methods. The vast majority of these input systems are equipped with auto-completion and correction functionalities, which, to a certain extent, mask learners' actual command of character structure and writing rules. Consequently, collecting data solely through typed text may fail to accurately reflect the learners' true cognitive and operational proficiency with Chinese characters.

Based on these considerations, the collection of learners' handwritten data is widely regarded as having a high reference value in learner corpus research. Preserving original handwritten texts through optical scanning allows researchers to capture learners' authentic writing behavior with high fidelity, while also providing an essential visual basis for subsequent recognition, transcription, and analysis.

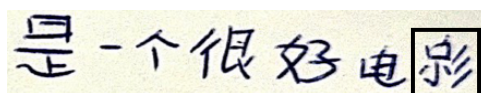
Learners' handwritten texts in Chinese interlanguage often exhibit various non-standard features. While the content—encompassing grammatical and pragmatic aspects—naturally contains numerous errors (which in corpus construction must be addressed in subsequent annotation, not transcription), the focus here is on the non-standard characteristics related to the form of the handwritten data. These formal complexities present the primary challenges during the transcription phase of handwritten data, including incorrectly written and incorrectly used characters.

Incorrectly written characters

Learners frequently produce incorrectly written characters during the writing process. These are characters that deviate from standard orthographic forms—often manifesting as self-invented characters due to anomalies in stroke, component, or structure. This category of error typically involves the addition or omission of radicals or strokes, the inversion of component positions, structural disproportions, or even the creation of entirely novel character forms by the learner (Zhang, 2016, p52). Since these altered forms lack corresponding encodings in the standard Chinese character system and cannot be input on a computer keyboard via conventional methods, they represent non-standard and non-digitizable characters in corpus transcription. As illustrated in Figure 4.2, an example taken from MICICL, the learner corpus constructed for the present study, shows that the intended target character is 影 (yǐng). Although the form remains contextually interpretable, several strokes are missing, causing it to deviate from the standard character form and resulting in an untypable handwritten character. The design and compilation of MICICL are described in detail in Chapter 3. Such errors not only reflect the learner’s deficiency in orthographic knowledge but also pose a significant challenge to the transcription process. The transcriber must strike a balance between faithfully representing the original output and ensuring the computational processing of the data. Therefore, establishing a methodology for the annotation and visual representation of these orthographic errors becomes a pivotal concern in the processing of Chinese handwritten data for interlanguage corpora.

Figure 4.2

An example of incorrectly written character⁴⁷



Source: MICICL 1902022312X

Incorrectly used characters

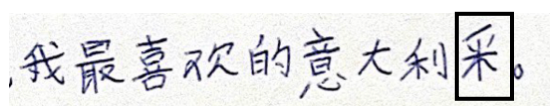
During the writing process, learners sometimes substitute the intended target character (A) with

⁴⁷ Boxes in the figure are overlaid by the author for the purpose of clarity only, and were not part of the learners’ original writing. The same holds for all boxes in the figures that follow.

a different character (B) that is similar in either orthography or phonology. The critical distinction from incorrectly written characters is that character B itself is a standard, typable character, but a contextually inappropriate one. For instance, in Figure 4.3, the intended character 菜(cài, “vegetable”) is replaced by the orthographically similar character 采(cǎi, “pick”). This character misuse reflects an unstable mastery in the learner’s form-sound-meaning integration mechanism, and it reveals a strong tendency to be influenced by either visual similarity or phonetic transfer during writing.

Figure 4.3

An example of incorrectly used character



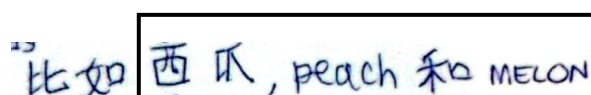
Source: MICICL 2102022312X

Code mixing

When learners are unable to recall or write a specific Chinese character, they often employ Pinyin or foreign letters/words as a substitution—a classic phenomenon of strategic avoidance. Furthermore, some learners may append Pinyin, English, or other language glosses alongside the characters, resulting in a mixed script where Chinese characters and Latin alphabet letters coexist. As seen in Figure 4.4, the student substitutes the intended characters with an English word; in Figure 4.5, Pinyin is used in place of an unfamiliar character.

Figure 4.4

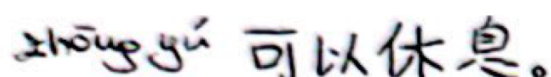
Example 1 of code mixing



Source: MICICL 2405052404X

Figure 4.5

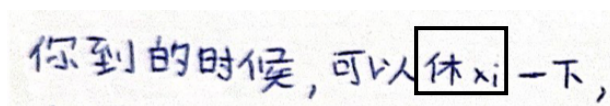
Example 2 of code mixing



More complex scenarios include Pinyin lacking tone marks or having erroneous tone annotations (Figure 4.6), or the simultaneous occurrence of a miswritten character and Pinyin/foreign word mixing (Figure 4.7). This multi-system, potentially erroneous, code-mixing phenomenon significantly increases the complexity of text transcription and subsequent annotation.

Figure 4.6

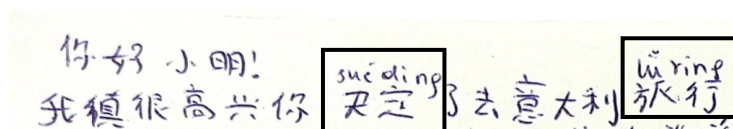
Example 3 of code mixing



Source: MICICL 0903022312X

Figure 4.7

Example 4 of code mixing



Source: MICICL 0703022312X

Manuscript markings and corrections

During the physical act of handwriting, learners frequently make immediate corrections to the text. These manifest as a variety of features: cross-outs, smudges, insertion marks, or paragraph reordering symbols. Examples include striking out the original character and rewriting above it, using an insertion symbol “^” to add a word, or employing arrows to adjust sentence sequence (Figure 4.8 and 4.9). These non-textual symbols and revision traces constitute a crucial record of the learner’s authentic writing process. However, if improperly handled during transcription, they can easily become conflated with the main text. Therefore, maintaining the information value of these corrections while ensuring the readability and computational tractability of the text is a key challenge in establishing transcription standards.

Example 1 of correction

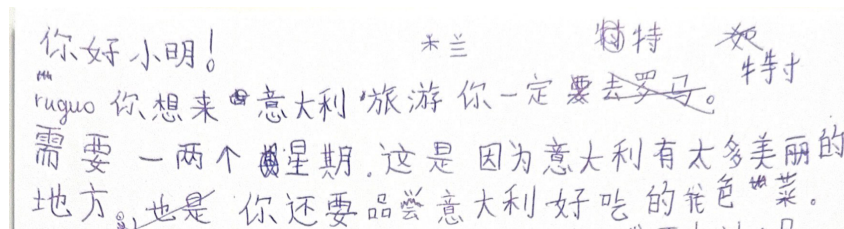
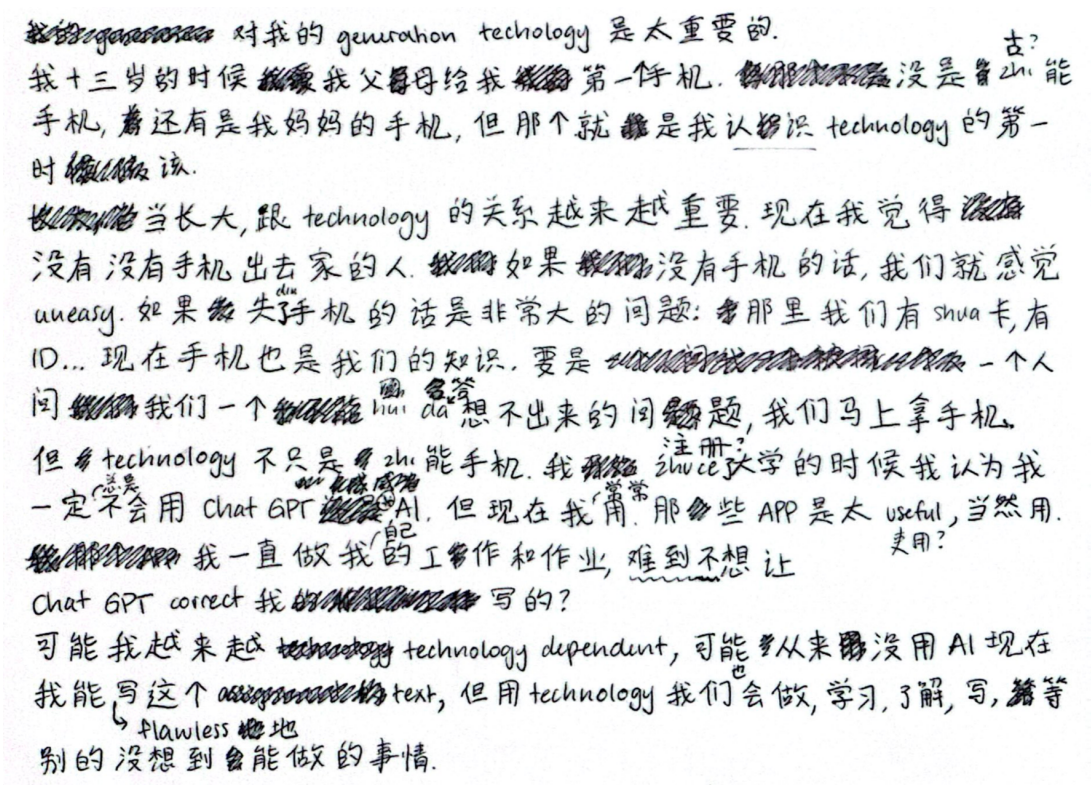


Figure 4.9

Example 2 of correction



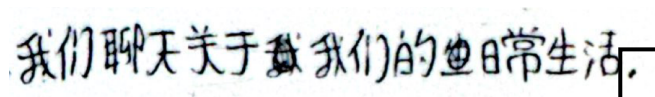
Source: MICICL 3102092412X

Punctuation is an integral part of the language system, and learners' writing often reveals diverse punctuation errors. Chinese uses full-width punctuation (全角, *quán jiǎo*), which differs significantly in form from the half-width punctuation (半角, *bàn jiǎo*) found in

languages like English (e.g., the period is “。” instead of “.”). Chinese also employs unique symbols such as the book title mark (“《》”⁴⁸). Learners are often influenced by their L1 writing conventions, leading to the mixing or misuse of punctuation. For instance, the student in Figure 2.9 uses a half-width period instead of the required full-width Chinese period.

Figure 4.10

An example of incorrect punctuation



Source: MICICL 0204052404X

In summary, the formal complexity inherent in learners’ handwritten data—ranging from miswritten characters and character misuse to Pinyin intermixture, correction marks, and punctuation errors—places high demands on the text transcription process. To ensure the consistency and computational processability of corpus data, it is imperative to establish clear, operative transcription guidelines that balance the faithful representation of the learners’ original output with the necessity of achieving standardization and structuralization. This leads to a central methodological question: When faced with such diverse and non-standard writing forms, how should the transcription system achieve an optimal balance between authenticity and processability? In the following, I examine several existing Chinese interlanguage writing corpora and analyze the approaches they use to transcribe handwritten data, focusing on how they address the aforementioned special issues.

At the annotation level, complexity is first manifested in the diversity of error types. Learners’ writing may feature incorrectly written characters, incorrectly used character, and intermixture of Pinyin or foreign letters. These categories reflect the learners’ mastery status at different levels, such as fragile form memory, sound-form decoupling, or the use of cross-code strategies. To accurately reveal these distinctions in analysis, the corpus’s annotation scheme

⁴⁸ The book title mark 《 》 is a pair of angle brackets used in written Chinese to enclose the titles of books, films, newspapers, journals, and other standalone works. It functions as a title-marking convention in Chinese, where italics or quotation marks are not traditionally used.

must separately encode each error type to enable clear typological differentiation and statistical comparison. However, a multi-category annotation system inevitably increases the number of labels and operational complexity, simultaneously demanding higher inter-annotator consistency.

Second, the common phenomenon of code-mixing in learner writing introduces an additional challenge. These instances take various forms, ranging from the direct substitution of characters with Pinyin or English to the parallel use of Pinyin and Hanzi. Annotation must be designed to faithfully represent the formal characteristics of this code-mixing while accurately conveying its language-strategic function (e.g., compensatory use or temporary substitution). This requires the annotation scheme to maintain a delicate balance between formal representation and explanatory power. Furthermore, the scope of corpus annotation often extends beyond single characters to include information at the word, sentence, and discourse levels. A key design challenge in a multi-layered annotation system is how to maintain the text's readability and structural clarity while providing sufficient linguistic information. Excessive embedding of annotation information renders the text difficult to parse, while insufficient layering may lead to the loss of crucial research-relevant linguistic cues. Therefore, establishing an annotation strategy that is hierarchically distinct, logically coherent, and operationally feasible is paramount to ensuring the corpus's usability.

At the correction level, if the annotation involves error information, corresponding correction data must be provided concurrently. The mode of correction presentation (e.g., inline embedding or a separate correction layer) and the granularity control of the correction are critical considerations. Over-correction risks obscuring the authentic features of the learner's interlanguage, while under-correction compromises the corpus's comparability and analytical utility. The dilemma of providing standardized, analyzable correction information while simultaneously maintaining the authenticity of the data is another key problem that corpus designers must resolve.

4.3.2 Transcription of handwritten data

For learners' handwritten data, MICICL adopts a transcription approach based on standard

Chinese characters, supplemented by a limited set of explicit annotation symbols to represent non-standard forms and special writing phenomena (e.g., non-typeable characters). This approach avoids embedding images or screenshots in the transcribed text, thereby ensuring a simple and uniform text format. The specific rules are as follows.

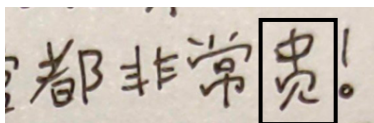
Non-typeable Chinese characters

If a learner wrote an incorrectly written character that does not exist in the standard Chinese system, it was not skipped or corrected during transcription, but marked with a special symbol string. In MICICL, a simple and intuitive rule is adopted for marking incorrect characters: the symbol “*” is used to indicate an incorrectly written character that cannot be typed or a character that is illegible. This method allows readers to know that an incorrect character exists without needing an image. Compared to embedding images, this purely textual marking is more convenient for retrieval and statistical analysis.

Given that handwritten data is highly variable and prone to arbitrary forms, a clear definition of an incorrectly written character is required. Referencing existing literature (Zhang, 2014, 2016), MICICL defines an incorrectly written character as one that involves: missing strokes, missing components, mistakenly added strokes, mistakenly added components, misplaced strokes, misplaced components, or obvious misrepresentation of strokes, misrepresentation of components, blending of strokes, or blending of components—any instance that clearly alters the number, shape, relationship, or structure of the character’s strokes or components, making it non-compliant with standard Chinese orthography, is counted as a incorrectly written character. For instance, in Figure 4.11, the last character should be 贵 (guì) but is missing a horizontal stroke. This is a stroke quantity error and is judged as an incorrectly written character.

Figure 4.11

An example of incorrectly written character 贵(guì)

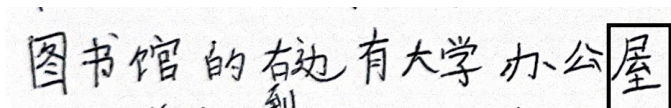


Source: MICICL 1502022312X

In Figure 4.12, the student intended to write 屋 (wū) but added an extra dot. This is a stroke quantity error and is judged as an incorrectly written character. In Figure 4.13, the last character should be 楼 (lóu), but the top part is missing two strokes.

Figure 4.12

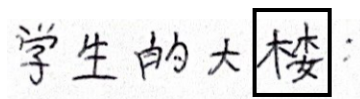
An example of incorrectly written character 屋(wū)



Source: MICICL 2103032402X

Figure 4.13

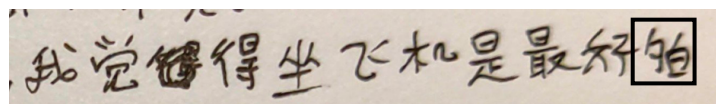
An example of incorrectly written character 楼(lóu)



Source: MICICL 2103032402X

Figure 4.14

An example of incorrectly written character 的(de)

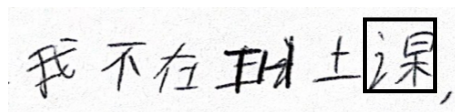


Source: MICICL 1502022312X

In Figure 4.14, the last character writes the left and right components of “的 (de)” in reverse order. This is judged as an incorrectly written character.

Figure 4.15

An example of incorrectly written character 课(kè)

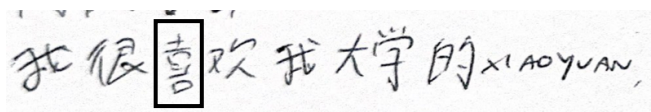


Source: MICICL 0604032402X

In Figure 4.15, the vertical stroke in the character “课 (kè)” (from “上课 (shàngkè)”) should pass through the upper part. It is judged as an incorrectly written character. In Figure 4.16, the upper part of “喜 (xǐ)” should be “士 (shì),” where the first horizontal stroke is longer than the second. The student wrote it as “土 (tǔ),” where the second horizontal stroke is longer. This is judged as an incorrectly written character.

Figure 4.16

An example of incorrectly written character 喜(xǐ)

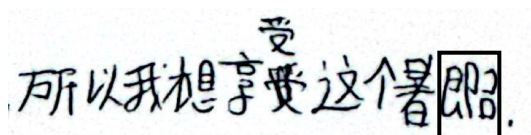


Source: MICICL 0604032402X

In Figure 4.17, the last character is a student-created form that does not exist in the Chinese character system. It is judged as an incorrectly written character and replaced with “*” in the transcribed text.

Figure 4.17

An example of incorrectly written character 假(jià)



Source: MICICL 0204052404X

If the written strokes or components are somewhat non-standard but do not cause confusion with other strokes or components, they are not counted as incorrect characters. Any lack of standardization in strokes or structure caused by writing habits (e.g., a vertical stroke habitually ending with a hook) or writing environment (e.g., uneven stroke thickness) is treated leniently if the character can be clearly distinguished from others, and is not counted as an incorrectly written character. For example, in Figure 4.18, the structure of “容易 (róngyì)” is not standard, appearing top-heavy; and in Figure 4.19, the components of the two characters in “饭馆 (fàn guǎn)” are not written tightly enough. However, since these only involve issues of font standardization and aesthetics, and the characters themselves are correct, they are not counted as incorrectly written characters and are transcribed as written. Chinese characters about which

native Chinese speakers themselves have writing controversies are treated leniently and are not counted as incorrect characters.

Figure 4.18

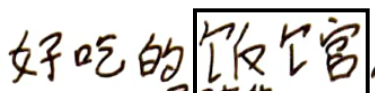
An example of correctly written character 容(róng) and 易(yì)



Source: MICICL 2602092412X

Figure 4.19

An example of correctly written character 饭(fàn) and 馆(guǎn)



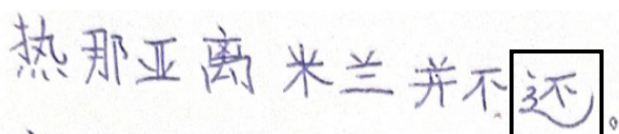
Source: MICICL 1104032402X

Inputtable Chinese characters

Any Chinese characters that can be typed (well-formed, existing characters in Chinese) appearing in the learner's composition are transcribed verbatim into the corresponding text, regardless of whether they are used correctly or not. Incorrectly used characters are also transcribed according to the actual text.

Figure 4.20

An example of incorrectly used character



Source: MICICL 2102022312X

For example, in Figure 4.20, the learner wrote “热那亚离米兰并不还 (Rènàyà lí Mílán bìng bù hái).” The character 还 (hái) is an incorrectly used character; the target character should be 远 (yuǎn). During transcription, it is written as “热那亚离米兰并不还”.

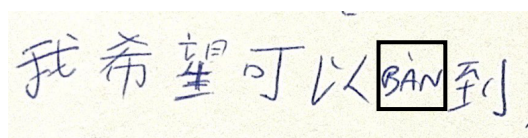
Latin Letters

When encountering content written by the learner in a foreign language, such as Pinyin or English, the same string of Latin letters is directly retained in the transcription, ensuring the transcribed text covers all information points in the original text.

Regarding Latin letters, some Italian native speakers habitually use all capital letters in their writing, including for Pinyin which is typically lowercase (as seen in Figure 4.21). The case of the letters does not affect the interlanguage analysis. For ease of reading, statistics, and transcription efficiency, all letters are transcribed in lowercase, except for the first letter of proper nouns, which remains capitalized.

Figure 4.21

An example of Pinyin written in capital letters

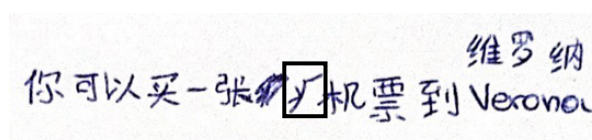


Source: MICICL 0602012311X

Sometimes a student writes a Chinese character/Pinyin/foreign word and then supplements it in the blank space with another form. This may reflect the student's hesitation or uncertainty about a word. In such cases, the content within the text line is transcribed first, followed by the supplementary content from the blank space. As shown in Figure 4.22, the transcribed text is: “你可以买一张*机票到 Verona 维罗纳”.

Figure 4.22

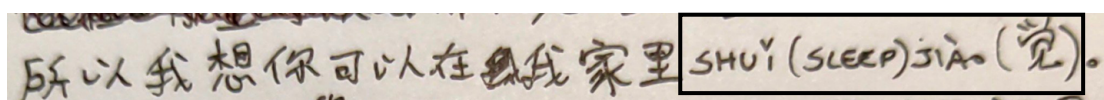
An example of coexistence of characters and latin letters



Source: MICICL 0903022312X

Figure 4.23

An example of multi-format writing in parentheses



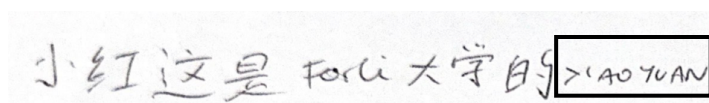
Source: MICICL 1502022312X

If a student uses parentheses to present multiple forms within a single line of text. As shown in

Figure 4.23, it is transcribed according to the actual text: “shuǐ(sleep)jiào(觉) (shuǐ (sleep) jiào (jiào)”. If a student produces incorrect Pinyin or Pinyin with missing components (e.g., missing tone marks), the Pinyin form is transcribed verbatim without completion. As shown in Figure 4.24, the student’s written Pinyin is missing a tone mark. It is therefore transcribed exactly as produced: “小红这是 Forli 大学的 xiaoyuan (xiǎohóng zhè shì Forli dàxué de xiàoyuán)”.

Figure 4.24

An example of incomplete Pinyin

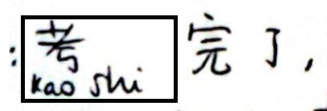


Source: MICICL 0604032402X

Sometimes a student only writes part of the Chinese characters for a word in the text line, leaving a blank space for the missing character, while providing Pinyin or a foreign word in the blank space to explain what they intended to write. In this case, the corresponding position is left blank in the transcription, followed by the Latin letter string. As shown in Figure 4.25 and 4.26, the texts are transcribed as: “考 kaoshi 完了 (kǎo kaoshi wán le),” and “冬季的时候还有 假 hánjià (dōng jì de shíhou háiyou jiǎ hánjià).”.

Figure 4.25

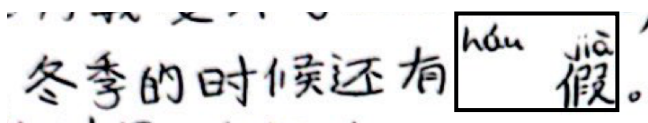
Example 1 of Pinyin/foreign word with character blank



Source: MICICL 0404052404X

Figure 4.26

Example 2 of Pinyin/foreign word with character blank



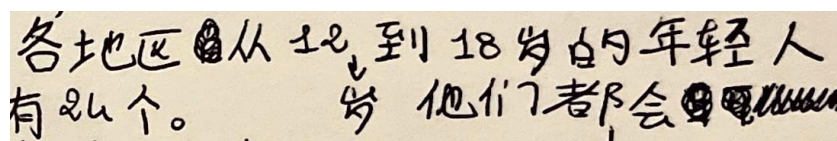
Source: MICICL 1006052404X

Arabic numerals

If the student uses Arabic numerals, they are transcribed in their original form as the corresponding Arabic numerals. As shown in Figure 4.27, it is transcribed as: “各地区从 12 岁到 18 岁到年轻人 (gè dìqū cóng 12 suì dào 18 suì dào niánqīngrén)”.

Figure 4.27

An example of using Arabic numerals



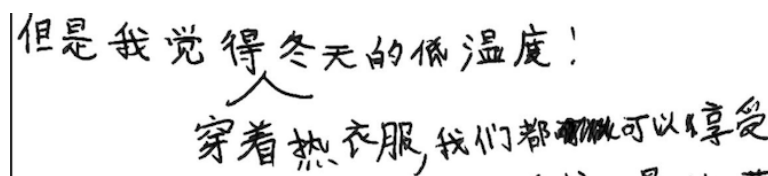
Source: MICICL 1005042403X

Handwriting corrections

For instances where the manuscript includes arrows for insertion, added content, or cross-outs for deletion, MICICL directly transcribes the content after the student's handwriting correction. The deletion/correction symbols themselves are not transcribed or annotated. While these corrections are interesting, reflecting the student's thought process and possibly language proficiency, MICICL provides the scanned image of the writing data, giving users the opportunity to browse this information visually. As shown in Figure 4.28 and 4.29, the student uses various symbols for text addition, position swapping, and extensive scribbling.

Figure 4.28

An example of handwriting corrections (text addition)



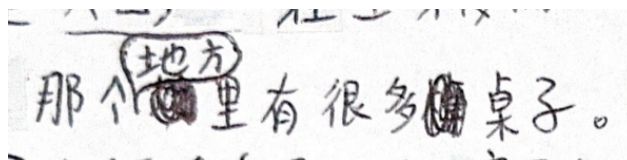
Source: MICICL 2105052404X

These are not transcribed. Instead, the text is transcribed directly after the student's own additions, deletions, and rearrangements—which is also the text read by the student during the reading task. As shown, the texts are transcribed as: “但是我觉得穿着热衣服, 我们都可以

享受冬天的低温度 (dànshì wǒ juéde chuānzhe rè yīfu, wǒmen dōu kěyǐ xiǎngshòu dōngtiān de dī wēndù)!", and “那个地方里有很多桌子 (nàge dìfāng lǐ yǒu hěn duō zhuōzi).”.

Figure 4.29

An example of handwriting corrections (text addition and scribbling)



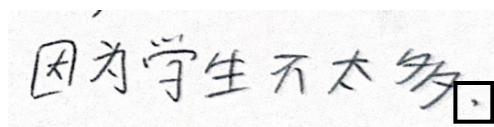
Source: MICICL 0504032402X

Punctuation and formatting

Punctuation, capitalization, line breaks, and other formatting features in the learner's composition were generally transcribed as they occurred, unless they are clearly unintentional slips of the pen. For example, extra periods or missing quotation marks will be preserved. Chinese uses full-width punctuation, which differs from the half-width punctuation used in Italian. If a student used a half-width period, as shown in Figure 4.30, this was transcribed using the half-width period: “因为学生不太多. (yīnwèi xuéshēng bù tài duō).”.

Figure 4.30

An example of incorrect punctuation



Source: MICICL 0604032402X

Through these strategies, MICICL aims to make the transcribed handwritten text fully retain the learner's writing behavior while minimizing the intrusion of images or hard-to-decode symbols. The transcribed composition text can be searched, segmented, and analyzed like ordinary digital text, while all deviations from the norm are marked with unified symbols, facilitating filtering or separate statistical analysis. Compared to previous solutions, this approach avoids embedded screenshots and achieves unity between information retention and text clarity: researchers can directly understand the content of the composition from the transcribed text and identify the error points through the markings. Furthermore, machine

processing can selectively ignore or statistically analyze these error details based on the annotation tags applied to the transcribed text. The original handwritten images are permanently stored as reference material for researchers who need to analyze writing habits.

4.3.2 Annotation of writing data

Annotation covers four levels: Character (including punctuation), Word, Phrase, and Sentence. The inclusion of the phrase level, which is not commonly adopted in existing Chinese interlanguage corpus annotations, is motivated by the following considerations.

During the examination of the MICICL corpus, it was observed that certain errors occur above the lexical level but below the sentential level. For instance, in some cases, words appear in an incorrect order within a sentence. For example, a learner may produce a sentence such as: “我常常去超市周末(*wǒ chángcháng qù chāoshì zhōumò*)” intended as “我周末常常去超市”(*wǒ zhōumò chángcháng qù chāoshì*, “I often go to the supermarket on weekends”). In this case, the error lies in the ordering of the phrase “周末”(*zhōumò*, “on weekends”) relative to the rest of the sentence, rather than in any individual word or in the overall sentence structure. Annotating this error at the lexical level would fail to capture the relational nature of the problem, while annotating it at the sentential level would treat the entire sentence as erroneous without specifying the locus of the word-order issue.

Therefore, a more fine-grained level of annotation is required to accurately represent such intermediate-level errors. Another common situation involves collocational errors, such as inappropriate combinations of measure words and nouns. If such errors were marked only at the lexical level—by labeling the measure word as an “incorrect word”—the learner’s actual interlanguage performance would not be fully represented. In these cases, the problem does not lie in the use of an incorrect word form but in the combination of words.

Therefore, an intermediate phrase level was introduced between the lexical and sentential levels to annotate these types of errors more precisely. According to the widely accepted definition, a phrase is a structural unit composed of words that functions as a sentence constituent (Wang, 1984, p. 12). The internal structure of a phrase is relatively stable and usually functions as a whole with other constituents in the sentence (Zhou & Yu, 1996, p. 1).

From a pedagogical perspective, phrase-level annotation holds significant value. Systematic annotation of word-order errors and collocational errors enables teachers to identify learners' difficulties in syntactic and semantic organization, as well as the patterns of word combination that characterize their interlanguage system. For example, some learners, influenced by their first language, may tend to transfer their native word order or lexical co-occurrence patterns into Chinese. Statistical analysis of such errors can provide valuable insights for targeted instruction, the design of vocabulary and syntax teaching activities, and even the development of teaching materials and automatic error-correction systems.

In terms of annotation logic, MICICL follows the “bottom-up” principle, meaning annotation proceeds in order from the smallest to the largest unit. It starts at the Character level, where learners' deviations in character form and use are annotated. This then extends to the Word and Phrase levels, and finally, a comprehensive analysis is conducted at the Sentence level. This hierarchical annotation system helps establish connections between different linguistic units, thereby revealing the developmental patterns of the learner's Chinese interlanguage system. Note that words and characters are distinguished semantically, not by the number of characters. Monosyllabic words are annotated as words, not characters.

In terms of annotation labels, the four levels of annotation—Character, Word, Phrase, and Sentence—are represented by: `</>`, `[/]`, `{/}`, and `[/]` respectively. The content within the brackets represents the erroneous segment, and the annotation label indicating the error type is placed at the end of the brackets. The reason for choosing not to use the parentheses () is to avoid confusion, as the original text written by the students may contain parentheses. In designing the specific annotation labels at each layer, all tags were constructed based on the initial letters of the corresponding error-category names in Chinese Pinyin (see the next sections). This design principle is consistent with that of the HSKEDC. On the one hand, using Pinyin initials helps maintain simplicity and systematicity in the tag system; on the other hand, given that users of the corpus are expected to possess at least basic knowledge of Chinese, this approach makes the tags easier to remember and apply. If more than one error occurs at the

same level, a plus sign (+) is used to mark two or more errors (e.g., if a character is both incorrectly written and redundant, it is marked with </C+D>).

4.3.4 Character level annotation of writing data

Table 4.1

Character level (including punctuation marks) annotation labels and definitions⁴⁹

Code	Name	Annotation Content	Example
</C>	错字 (Cuòzì)	Incorrectly written character	Pic-trans: 罗马的历史很丰* Pic-anno: 罗马的历史很丰<*/C>
	别字 (Biézi)	Incorrectly used character	Pic-trans: 在罗马可以吃意试面条 Pic-anno: 在罗马可以吃意<试/B>面条
</K>	空缺 (Kòngquē)	A position where a character should appear but was intentionally left blank by the learner	Pic-trans: 这种 器是这么重要 Pic-anno: 这种</K>器是这么重要
</L>	漏字 (Lòuzì)	A character that should appear in the sentence but is missing	Pic-trans: 最后他们决定了结婚,他还有三个孩子 Pic-anno: 最后他们决定了结婚,他<们/L>还有三个孩子
</D>	多字 (Duōzì)	A character that appears in the sentence but should not be there	Pic-trans: 所以你在建筑中走走的时候 Pic-anno: 所以你在建筑中走<走/D>的时候
</F>	繁体字 (Fántǐzì)	Traditional Chinese character	Pic-trans: 我希望你最近過得很好 Pic-anno: 我希望你最近<過/F>得很好
</S>	数字 (Shùzì)	Arabic numerals	Pic-trans: 一共有 24 个 Pic-anno: 一共有<24/S>个
</P>	拼音	The use of Pinyin in	Pic-trans: 在这儿有 xǐ 手间

⁴⁹ Capital letters in the Pinyin forms of the label names (e.g. Cuòzì, Biézi, Pīnyīn Cuòwù) reflect the actual letter(s) used in the corresponding annotation tags (e.g. /C, /B, /PC), rather than sentence-initial capitalization.

	(<i>Pīnyīn</i>)	place of a Chinese character	Pic-anno: 在这儿有<xǐ/P>手间
</PC>	拼音错误 (<i>Pīnyīn Cuòwù</i>)	Incorrect Pinyin	Pic-trans: 我希望你喜 huàn 住在这儿 Pic-anno: 我希望你喜<huàn/PC>住在这儿
</PQ>	拼音声调缺失 (<i>Pīnyīn shēngdiào Quēshī</i>)	Pinyin written without tone marks	Pic-trans: 我想去华人 jie 看看 Pic-anno: 我想去华人<jie/PQ>看看
</PZ>	拼音汉字同用 (<i>Pīnyīn hànZì tóngyòng</i>)	<i>Mixed Pinyin and corresponding Chinese character</i>	Pic-trans: 这儿博洛尼亚大学的校 xiào 园不太大 Pic-anno: 这儿博洛尼亚大学的<校 xiào/PZ>园不太大
</BC>	标点错误 (<i>Biāodiǎn Cuòwù</i>)	Incorrect punctuation mark	Pic-trans: 因为学生不太多. Pic-anno: 因为学生不太多<./BC>
</BL>	标点漏掉 (<i>Biāodiǎn Lòudiào</i>)	A punctuation mark that should appear in the sentence but is missing	Pic-trans: 寄生虫是我最喜欢的电影之一 Pic-anno: </BL>寄生虫</BL>是我最喜欢的电影之一
</BD>	标点多余 (<i>Biāodiǎn Duōyú</i>)	A punctuation mark that appears in the sentence but should not be there	Pic-trans: 故事开始的时候她们五-六岁 Pic-anno: 故事开始的时候她们五<- /BD>六岁

Note: The examples in the table are all drawn from the MICICL corpus. Since the original data often involves more than one error, to ensure clearer representation in the table, all errors but one have been corrected.

Some further by-category explanations:

- `</B>`: Used to mark cases where character A was written as character B. In such cases, B refers to an existing Chinese character that can be entered via a standard input method.
- `</K>` and `</L>`: `</K>` is used to mark a blank space intentionally left by the learner for a character that should have been written. In this situation, a visible empty position appears in the handwritten text and is represented as a space in the transcription. In contrast, `</L>` marks omitted characters—those that should appear in the text but were not written and for which no space was left in the original data. In short, `</K>` indicates that the learner was aware that a character should appear in this position but, due to insufficient mastery, intentionally left it blank; whereas `</L>` indicates an unconscious omission made out of carelessness or inattention.
- `</S>`: Arabic numerals are marked with `</S>` because their use is not merely a matter of writing habit but represents a cross-script writing strategy. This reflects a form of functional substitution, in which learners use symbols from another orthographic system to compensate for incomplete mastery of the target script. From a semiotic perspective, Arabic numerals belong to a non-logographic symbol system and do not share the orthographic properties of Chinese characters. Thus, the appearance of numerals in a Chinese orthographic context can be considered an act of cross-script transliteration or heterographic borrowing. Such numerals are highlighted in the transcription, and in the subsequent correction process they are converted into their corresponding Chinese numerals (e.g., “2” → “二 (*èr*)”, “10” → “十 (*shí*)”) so that the corpus returns to a consistent Chinese orthographic system.
- `</P>`: Used to indicate cases where the learner wrote Hanyu Pinyin instead of Chinese characters, and the Pinyin used is accurate and complete.
- `</PQ>`: Used to mark Pinyin written without tone marks. This label was added because such cases were frequently observed in the MICICL data. The omission of tones may

result from learners' limited tonal competence or intentional avoidance of tonal marking during writing. These cases can later be cross-referenced with the reading-aloud data to examine whether the learners were able to pronounce the correct tones.

- **</PZ>**: Used to mark cases where learners wrote both the Pinyin and the corresponding Chinese character. If both forms are correct, the instance is not considered an error, but it indicates that the learner was uncertain or not yet fully automatized in character writing. In cases where the learner wrote both but the Chinese character was incorrectly written or incorrectly used, the combined labels **</PZ+B>** or **</PZ+C>** are used.
- **</BC>**: Used to mark incorrect punctuation use, such as using half-width punctuation where full-width punctuation should be used.
- During the transcription process, no illegible characters were encountered; therefore, no label was defined for unrecognizable handwriting.

4.3.5 Other levels of annotation of writing data

For word level annotation, the specific details can be seen in Table 4.2:

Table 4.2

*Word level annotation labels and definitions*⁵⁰

Code	Name	Annotation Content	Example
[/CC]	词 错 误 (<i>CiCuòwù</i>)	Incorrect word	Pic-trans: 我觉得坐火车是最好的方法 Pic-anno: 我觉得坐火车是最好的[方法/CC]
[/CD]	词 多 余 (<i>CiDuōyú</i>)	A word that appears in the text but should not be there	Pic-trans: 你不用担心吧 Pic-anno: 你不用担心[吧/CD]
[/CL]	词漏(<i>CiLòu</i>)	A word that should appear	Pic-trans: 从米兰机场差不多

⁵⁰ Capital letters in the Pinyin forms of the label names reflect the actual letter(s) used in the corresponding annotation tags, rather than sentence-initial capitalization.

		in the sentence but is missing	两个小时的火车 Pic-anno: 从米兰机场[/CL]差不多两个小时的火车
[/CS]	生造词 (<i>Shēngzàocí</i>)	A self-created word	Pic-trans: 还是中国的省都 Pic-anno: 还是中国的[省都/CS]
[/CLH]	离合词错误 (<i>LíHécí Cuòwù</i>)	A separable verb error	Pic-trans: 如果人类都多睡觉一点儿 Pic-anno: 如果人类都多[睡觉/CLH]一点儿
[/W]	外文词 (<i>Wàiwéncí</i>)	A foreign word	Pic-trans: 晚上我举办一个小 party Pic-anno: 晚上我举办一个小 [party/W]
[/WZ]	外文词和汉字同用 (<i>Wàiwéncí he hànZì tóngyòng</i>)	Mixed foreign word and corresponding Chinese character	Pic-trans: 成为一位英雄 hero Pic-anno: 成 为 一 位 [英 雄 hero/WZ]

Some further explanations of the tags:

- [/CC]: Used to mark incorrect words. In such cases, the characters within the word have already been annotated at the character level, so this label focuses only on the word level. Typical cases include word-internal character order errors, e.g., 首先 [/CC 先首] (“firstly”) and 众所周知 [/CC 众所知周] (“as everyone knows”).
- [/CS]: Indicates invented words created by the learner, or words that exist in other languages but not in Chinese.

- [/CLH]: Used to mark errors related to separable verbs (离合词, *lihéci*). The label is placed after the erroneous separable verb, indicating that the preceding expression contains a misuse or mis-segmentation of a separable verb structure.
- In cases involving foreign words, if the foreign word is a common noun, it should be uniformly annotated. If the foreign word is a proper noun, it is treated case by case:
 - i. When there is an official Chinese translation (e.g., Bologna → 博洛尼亚), the foreign word should be annotated as an error.
 - ii. When the proper noun (e.g., a brand name, pet name, or shop name) lacks an official translation but is necessary for meaning and cannot be replaced, it should be kept in its original form. Example: “学校里有一家咖啡馆，它的名字是 Volume。” (There is a café at school called “Volume.”)
- Errors caused by omission or redundancy of prepositions, localizers, or particles, as well as misuse of parts of speech, are all treated as word-level errors.
- Confusion among the structural particles “的,” “地,” and “得”⁵¹ is also treated as a word error.

Phrase level annotation labels are outlined in Table 4.3.

Table 4.3

*Phrase level annotation labels and definitions*⁵²

Code	Name	Annotation content	Example
{/DXC}	短 语 顺 序 错 误 (<i>Duányǔ shùnXù Cuòwù</i>)	Phrase order error	Pic-trans: 常常最贵的宾馆在城市中心 Pic-anno: {常常最贵的宾馆/DXC} 在城市中心
{/DPC}	短 语 搭 配 错 误	Phrase	Pic-trans: 还有一个咖啡馆

⁵¹ Function words that mark attributive, adverbial, and complement relations in Chinese.

⁵² Capital letters in the Pinyin forms of the label names reflect the actual letter(s) used in the corresponding annotation tags, rather than sentence-initial capitalization.

	(<i>Duányǔ</i> <i>dāPèi</i> <i>Cuòwù</i>)	collocation error	Pic-anno: 还有{一个咖啡馆/DPC}
--	--	----------------------	-------------------------

It should be noted that if two or more consecutive words are missing in the corpus, the case should be annotated using the word-level label /CL, rather than the phrase-level label. This follows the bottom-up logic of the annotation hierarchy.

As discussed in Section 4.2, existing learner corpora typically adopt theory-driven, grammar-oriented sentence-level annotation systems, with a primary focus on syntactic and grammatical analysis. However, such annotation schemes have limited direct value for Chinese language teaching, as teachers often cannot apply them directly in classroom settings without further interpretation or adaptation.

Since the MICICL corpus is designed with pedagogical applications in mind, its annotation framework emphasizes practicality rather than theoretical orientation. At the sentence level, MICICL abandons traditional grammar-based annotation and instead classifies errors according to the types of ill-formed sentences commonly encountered in Chinese language teaching. This allows teachers to conveniently retrieve relevant examples of erroneous sentences, gain an intuitive understanding of learners' actual performance, and directly incorporate these examples into teaching activities. Drawing on the HSK test framework and previous research (Chen, 2003; Tan & Zhao, 2005; Jia & Zhu, 2012; Duan, 2015), this study categorizes erroneous sentences into nine types: collocation error, word order error, incomplete component, redundant component, word choice error, logical error, sentence structure confusion, unclear expression, and ambiguous sentence. Sentence level annotation labels are shown in Table 4.4:

Table 4.4

*Sentence level annotation labels and definitions*⁵³

Code	Name	Annotation content	Example
[/JPC]	句内搭配	Collocation	Pic-trans: 校园里有食堂，还有一个咖啡馆，

⁵³ Capital letters in the Pinyin forms of the label names reflect the actual letter(s) used in the corresponding annotation tags, rather than sentence-initial capitalization.

	错误 (<i>Jùnnèi dāPèi Cuòwù</i>)	error	我很喜欢这家咖啡馆。 Pic-anno: 「校园里有食堂, 还有{一个咖啡馆/DPC}, 我很喜欢这家咖啡馆。/JPC」
「/JXC」	句内顺序 错误 (<i>Jùnnèi shùnXù Cuòwù</i>)	Word order error	Pic-trans: 常常最贵的宾馆在城市中心。 Pic-anno: 「{常常最贵的宾馆/DXC}在城市中心。/JXC」
「/JQ」	句子成分 残缺 (<i>Jùzi chéngfēn cànQuē</i>)	Incomplete sentence component	Pic-trans: 从米兰机场差不多两个小时的火车。 Pic-anno: 「从米兰机场[/CL]差不多两个小时的火车。/JQ」
「/JD」	句子成分 多余 (<i>Jùzi chéngfēn Duōyú</i>)	Redundant component	Pic-trans 冬季的时候还可以做冬季运动, 比如说滑冰, 可能还有滑雪。 Pic-anno: 「冬季的时候还可以做冬季运动, 比如说滑冰, [可能/CD]还有滑雪。/JD」
「/JYC」	句子用词 错误 (<i>Jùzi Yòngcí Cuòwù</i>)	Word choice error	Pic-trans: 如果你想看很多城市, 我觉得坐火车是最好的方法。 Pic-anno: 「如果你想看很多城市, 我觉得坐火车是最好的[方法/CC]。/JYC」
「/JLC」	句子逻辑 错误 (<i>Jùzi Luójí Cuòwù</i>)	Logical error	Pic-trans: 你的身体应该比较好因为天气很热。 Pic-anno: 「你的身体应该比较好因为天气很热。/JLC」
「/JSZ」	句式杂糅 (<i>JùShì Záróu</i>)	Sentence structure confusion	Pic-trans: 另外, 为了学习中文, 手机特别有用。 Pic-anno: 「另外, 为了学习中文, 手机特别有用。/JSZ」
「/JBY」	句子表意	Unclear	Pic-trans: 我真想你, 但是我知道你一直努力!

	不 明 (<i>Jùzi BiǎoYì bùmíng</i>)	expression	Pic-tanno:「我真想你,但是我知道你一直努力! /JBY」
「/JQY」	句 子 有 歧 义 (<i>Jùzi yǒu QíYì</i>)	Ambiguous sentence	Pic-trans: 在那儿,我跟别的意大利同学上了中文课,也参加了很多活动。 Pic-anno:「在那儿,我跟别的意大利同学上了中文课,也参加了很多活动。/JQY」
「/JC? 」	存 疑 句 错 误 (<i>cúnyí Jù Cuòwù</i>)	Uncertain sentence error	Pic-trans: 虽然他们相爱,但是一个人做一凶的事。 Pic-anno:「虽然他们相爱,但是一个人做一凶的事。/JC? 」

Sentence segmentation in MICICL is based on sentence-final punctuation marks in Chinese, including periods, exclamation marks, question marks, and ellipses. If learners used incorrect punctuation marks (in most cases, a mixture of half-width and full-width forms), segmentation is performed according to the target-correct punctuation.

Some sentence-level labels may partially overlap in form or function with labels at other annotation levels. For instance, “/JPC” and “/JXC” correspond functionally to the phrase-level labels “{/DPC}” and “{/DXC}”, while “/JYC” may overlap with the word-level label “[/CC]”. This overlap is intentional, as sentence-level annotation is designed to help teachers quickly and accurately identify completely erroneous sentences, making it easier to extract representative examples for teaching purposes.

In fact, the determination of sentence-level labels is derived from the annotation results at lower levels. Since MICICL follows a bottom-up annotation principle, the word- and phrase-level annotations are completed before the sentence-level annotation begins. Thus, the sentence-level labels serve as a comprehensive generalization of lower-level errors. For example, when a word-level error label “[/CC]” (incorrect word) appears in a sentence, the corresponding sentence-level label is “/JYC” (word choice error). When extra or missing words are marked, the sentence-level labels “/JD” (redundant component) or “/JQ” (incomplete

component) are added. Similarly, when phrase-level labels “{/DXC}” or “{/DPC}” (phrase order or collocation errors) appear, the sentence-level labels “/JXC” or “/JPC” are correspondingly applied. This progressive, hierarchical annotation logic maintains consistency across levels and enhances the interpretability and pedagogical usefulness of sentence-level annotations.

For sentences of uncertain error types or ambiguous meaning, the label “/JC?” is used to indicate unclear or indeterminate sentence errors.

4.3.6 Correction of writing data

In the correction layer, revisions are made based on the annotations in the transcription layer, with each annotation corresponding to one correction.

Character-level correction:

- Incorrectly written characters are corrected to their intended target characters, replacing the “*” symbol in the text, based on context and the original handwriting.
- Incorrectly used characters are replaced with the correct ones in the same manner.
- Traditional characters are converted to simplified forms.
- Arabic numerals are replaced with the corresponding Chinese numerals.
- All forms of Pinyin—including correct and complete Pinyin, Pinyin without tone marks, and incorrect Pinyin—are converted to the correct Chinese characters.
- When both Pinyin and Chinese characters appear together, the Pinyin is deleted while the character is retained (if the character itself contains a writing or use error, it is also corrected).
- For omissions or redundant characters (labels </K>, </L>, etc.), the corresponding characters are added or deleted.
- Incorrect punctuation marks are replaced with correct ones, and punctuation omissions or redundancies are similarly corrected.

Word-level correction:

- Incorrect or invented words are replaced with the most contextually appropriate words.
- Errors in separable verbs are corrected according to the minimal revision principle.

- Redundant or missing words are deleted or added to fit the context.
- Foreign words are translated into suitable Chinese equivalents.
- If both foreign and Chinese words co-occur, the foreign word is deleted while ensuring the correctness of the Chinese equivalent.
- For proper nouns, contextual judgment is applied: when an official Chinese translation exists (e.g., *Bologna* → “博洛尼亚”), it is used; for proper names such as café names or pet names without an official translation, the original Latin alphabet is preserved.
- Uncertain word errors are minimally corrected according to contextual appropriateness.

Phrase-level correction:

- Adjustments are made to word order or collocational relationships based on the identified type of phrase-level error.

Sentence-level correction:

- For some labels (e.g., sentence collocation errors or sentence word choice errors), corrections are already resolved during lower-level (word or phrase) processing. Other labels are corrected according to the minimal revision principle, providing a revised version that maintains the learner’s original intention while ensuring grammatical completeness.

4.3.6 Writing data transcription and annotation overview

A quantitative overview of MICICL writing data transcription can be seen in Table 4.5.

Table 4.5

Writing data transcription overview

Element type	No. of occurrences
Chinese characters	30524
* (Incorrectly written characters)	607
Pinyin	1900
Foreign words	1184
Punctuation marks	4505

Numbers	145
Total	38865

Note that the number of Chinese characters does not include incorrectly written characters, Pinyin, foreign words, or punctuation marks. Pinyin refers to all cases including Pinyin without tone marks, incorrect Pinyin, and Pinyin that appears together with Chinese characters. The term “number” refers to individual digits; for example, “311” is counted as three digits. The statistical results of the annotated labels are as follows:

Table 4.6

Writing data annotation label numbers

Annotation Level	Label	Frequency of occurrence
Character Level	C	607
	B	437
	K	32
	L	119
	D	68
	F	270
	S	63
	P	304
	PC	113
	PQ	1322
	PZ	161
	BC	344
	BL	156
	BD	28
Word Level	CC	832
	CD	294
	CL	812
	CS	9

	CLH	6
	W	1130
	WZ	54
Phrase Level	DXC	258
	DPC	75
Sentence Level	JPC	70
	JXC	230
	JQ	502
	JD	208
	JYC	557
	JLC	18
	JSZ	11
	JBV	60
	JQY	9
	JC	6

A total of 9,165 annotation labels were applied. Among them, 4,024 annotations were at the character level, 3,137 at the word level, 333 at the phrase level, and 1,671 at the sentence level. All the errors represented by these annotations were subsequently corrected in the correction layer.

4.4 MICICL reading-aloud data processing rules and outcomes

4.4.1 Complexity of reading-aloud data

Regardless of the type, audio data from Chinese interlanguage learners is often replete with non-standard features, posing numerous challenges for the data processing. Common phenomena include:

Phonological errors

Learners frequently make or confuse initial and final sounds (consonants and vowels), as well as make tone errors. These errors are abundant in interlanguage corpora and must be accurately captured and recorded during transcription.

Repetitions and self-corrections

During speaking or reading aloud, learners may repeat certain characters or words due to uncertainty, or they may immediately correct a mistake after uttering it. Consequently, the audio contains repeated segments or sentences that are revised midway. Even when the corrected content is accurate, these phenomena indicate the learner's lack of fluency with the pronunciation of a specific character or a particular grammatical point.

Fillers and pauses

Learners often produce filler words, such as “um,” “ah,” or “er,” during moments of thought or hesitation. These vocalizations typically carry no semantic meaning but reflect the learner's lack of fluency in that segment. Additionally, due to insufficient proficiency or the need for more processing time, learners may generate unnatural pauses in places where natural pauses between sentences do not occur, reflecting non-fluency characteristics.

The above features can appear in both reading-aloud and spoken data. In reading-aloud data, the following situations can further complicate transcription:

Non-textual vocalizations

These include instances where a learner asks a question during reading-aloud due to ambiguity in the text, or offers an explanation about their reading due to special circumstances. Such vocal content is irrelevant to the text being read and is often directed at the researchers present, sometimes even using the language of research communication (such as English) rather than Chinese.

Inconsistency between reading and text content

Learners may omit or misread parts of the text. In this context, “misreading” is not an error in initials, finals, or tones, but rather the substitution of a correct character/word in the text with a different character/word. In such cases, the content the learner reads is phonologically correct

(when isolated from the text) and may even be grammatically sound within the flow of speech; the error lies only in the disjunction with the source text. This also indicates the learner's uncertain or weak memory and mastery of certain characters/words.

The management of this complex information significantly impacts the final textual representation of the data. Below, we will examine the transcription schemes of several existing Chinese interlanguage corpora, compare their data processing methods, and thereby summarize current best practices and shortcomings in reading-aloud data processing.

4.4.2 Accuracy and reliability of reading-aloud data transcription

In the transcription of the reading-aloud data, the analysis focuses on the phonetic realization of Chinese syllables, including both segmental structure and tonal information. Transcription was carried out by the author, relying on auditory perception and native-speaker linguistic intuition to identify syllable segments and tones. Chinese syllables were transcribed using Hanyu Pinyin with numerical tone codes, while English or Italian words produced by the learners were transcribed by reproducing their original spoken forms without phonological adaptation.

I am a native of Henan, China, which belongs to the Central Plains Mandarin dialect area. My first language is Standard Mandarin, not a local dialect, and I grew up in a standard Mandarin phonetic environment. I participated in the Putonghua Proficiency Test (普通话水平测试, *pǔtōnghuà shuǐpíng cèshì*)⁵⁴ organized by the National Language and Writing Commission of China (国家语言文字工作委员会, *guójiā yǔyán wénzì gōngzuò wěiyuánhui*)⁵⁵, achieving a score in the Level One Grade B (一级乙等, *yījí yíděng*) range (92–96.9 out of 100). This is the second-highest level in the official standard, indicating normative pronunciation, correct use of initials, finals, and tones, minimal regional accent, and meeting professional requirements for language teaching and broadcasting.

Regarding the reliability of auditory judgment, extensive empirical research has demonstrated the accuracy and consistency of native speakers in perceiving Mandarin

⁵⁴ <https://www.cltt.org/#/webHome>

⁵⁵ http://www.moe.gov.cn/jyb_sy/China_Language/

segments and tones. These findings are directly relevant to learner data, as native speakers' sensitivity to segmental and tonal deviations provides a reliable basis for identifying non-target-like realizations in second language speech, particularly in controlled reading-aloud tasks where the intended lexical forms are known.

Accuracy of segmental perception

Zhang, Peng, and Wang (2016) found that native Chinese speakers exhibit typical categorical perception features in both tone and consonant perception tasks, reacting most sensitively at the phonetic continuum boundaries, while non-native speaker listeners lack clear categorical boundaries. This suggests that long-term linguistic experience modulates cue weighting, enhancing the ability to distinguish segmental differences. Research by Lin & Francis (2014, p. 2832) further indicates that the accuracy of native Mandarin speakers in segmental identification tasks consistently remains between 78% and 90%, showing strong perceptual stability. An experiment by Lai (2012, p. 246) showed that native speakers' accuracy in nasal contrast tasks (e.g., [m–n] and [n–ŋ]) exceeds 90%, demonstrating extremely high discrimination accuracy. The cross-language study by Tsao, Liu, and Kuhl (2006) also confirmed that native language experience forms stable phonetic category representations at the neural processing level, leading to greater consistency and automatization in consonant discrimination among native Mandarin speakers. These findings collectively establish a solid basis for the reliability of the phonetic judgment in this study's speech transcription.

Accuracy of tone perception

Numerous experimental studies indicate that native listeners' tone perception typically reaches a "ceiling effect" in quiet environments, meaning recognition accuracy is near perfect. Even under noisy or incomplete speech signal conditions, native speakers maintain high recognition levels (Gottfried & Suiter, 1997; Li et al., 2016; Lee et al., 2009). Native listeners possess categorical perception ability for tones (Hallé et al., 2004; Yang, 2021), allowing them to automatically ignore acoustic variations within a category and focus on semantically distinctive tone differences, thus maintaining highly stable recognition performance across different speakers. Experimental data from Liu (2013) showed that the accuracy of native listeners in

identifying single-syllable tones was about 88% under single-speaker conditions and 91% in context. Under multi-speaker conditions, these accuracies were 82% and 90%, respectively, demonstrating good cross-speaker consistency. Neurophonetic research by Krishnan et al. (2005) further revealed that the human brainstem's Frequency-Following Response (FFR) can track the pitch contours of Mandarin tones with high precision, and native experience significantly enhances the clarity and sensitivity of this encoding. This suggests that native speakers possess greater neural sensitivity to tone changes, providing physiological and cognitive support for tone annotation guided by auditory intuition. The latest experiment by Zhang et al. (2025, p. 8-9) also indicated that the perceptual accuracy of native listeners for the four tones in a natural speech environment is between 93% and 97%.

Coexistence of reliability and limitations

Undeniably, manual auditory transcription inherently involves a degree of subjectivity at the operational level, and the lack of objective verification for every syllable via automatic acoustic measurement constitutes a potential limitation of MICICL's reading-aloud data transcription. However, based on multiple lines of evidence from linguistic theory, experimental phonetics, and neurophonetics, the significant advantages of native Mandarin speakers in segmental and tonal perception renders auditory-based annotation, guided by native speaker intuition, highly reliable. In the specific operation, I made every effort to re-listen to each audio track multiple times to ensure consistent judgment, thereby minimizing subjective bias.

Therefore, although MICICL's Pinyin and tone transcription is primarily based on one-person manual auditory perception rather than an automated acoustic recognition system. This process is sufficiently supported by empirical evidence in theory, and its scientific validity and reliability are ensured through repeated verification and cross-checking in the methodology.

4.4.3 Transcription of MICICL reading-aloud data

For learners' reading-aloud data, MICICL adopts a method of Pinyin transcription with numerical tones plus necessary symbolic markers, deliberately avoiding the use of Chinese characters. Specific rules are as outlined below.

Transcription format

- All reading-aloud data are transcribed into Pinyin, following the Hanyu Pinyin scheme for the learner's pronunciation of each character.
- Pinyin is written entirely in lowercase, and the four tones (Yinping, Yangping, Shangsheng, Qusheng) and the neutral tone are marked using the numbers 0, 1, 2, 3, and 4 respectively.
- Syllables are delimited by spaces to enhance readability.
- Foreign words are transcribed in their corresponding original form.

Simplified labeling for filler words

In the MICICL corpus, all filler words (such as *ēn*, *ā*, *è*, etc.), regardless of their specific phonetic form, are uniformly transcribed using the symbol /en. This design is based on three main considerations: First, clarity and recognizability. The form /en is visually distinct from other Pinyin forms in the MICICL reading-aloud data, allowing readers to instantly recognize that the segment is not normal speech but a marker of learner hesitation or pause. This symbolic treatment simplifies the visual structure of the text, making the corpus presentation cleaner and more intuitive. Second, pedagogical and research focus. MICICL's emphasis is on examining the learner's oral fluency, rather than scrutinizing the subtle phonetic differences between different types of filler words. Adopting a unified, undifferentiated marker accurately records the frequency of learner hesitation while avoiding the subjective judgment errors and annotation burden associated with distinguishing pronunciations like *ēn* and *ā*. While this simplification has limitations (e.g., from a discourse function perspective, different fillers may carry different pragmatic functions), this information loss is negligible for MICICL's core goal of teaching and assessment. Last, operability. This scheme significantly reduces the transcription workload. The transcriber only needs to mark /en once for each instance of a filler pause, regardless of how many times the learner repeated the filler, thereby avoiding transcription inconsistencies and improving efficiency. For machine processing, /en serves as a unified token for automated analysis, facilitating the statistical counting of hesitation frequency or the filtering of spoken noise.

Other disfluency phenomena

In addition to filler words, MICICL draws on existing corpus experience to establish transcription rules for common oral disfluency phenomena such as repetitions and self-corrections. For repetitions, MICICL adheres to the principle of recording every instance. All repetitions are strictly transcribed according to the audio, without omission or merging. Self-corrections are also recorded verbatim.

Pauses

In the transcription of MICICL's reading-aloud data, the pause is recorded as a crucial feature affecting speech fluency. To balance corpus readability and analytical utility within a pedagogical framework, MICICL adopts a classification method based on semantic and written correspondence, dividing pauses into two types: Natural Pause (/ZT) and Non-Natural Pause (/FZT).

- Natural Pause (/ZT): Refers to a pause that occurs naturally according to syntactic or semantic structure (e.g., between clauses, at sentence ends, or between paragraphs). These pauses usually correspond to punctuation in the written text and are normal prosodic features of speech flow. In the transcription, a pause at a sentence end or a semantic boundary is marked as /ZT.
- Non-Natural Pause (/FZT): Refers to an unexpected pause caused by factors such as lack of proficiency, hesitation, word recall, or difficulty in speech planning. These pauses typically appear mid-clause or interrupt semantic flow, affecting continuity and fluency. Such pauses are marked as /FZT in MICICL.

In terms of annotation methodology, MICICL does not use precise time measurements to differentiate pause duration, but relies on auditory judgment to identify and record the existence of a pause. I only classified and annotated pauses that were clearly perceptible in the audio, without quantifying their duration. This approach simplifies the transcription process while ensuring the corpus's operability and consistency.

While pause duration has been discussed in existing research, no unified standard has been reached. Hu & Wang (2011) summarized that most scholars lean toward a dichotomy of

“short pause” and “long pause,” but consensus is lacking on the specific time threshold for “short” versus “long”. For Chinese learners, the frequency and position of pauses directly impact their fluency and discourse organization skills. Since systematic experiments are lacking to define which duration falls into the “non-natural” category, MICICL avoids time thresholds, using semantic interruption and perceived prominence as the differentiating criteria. This approach aligns better with the functional positioning of a pedagogical corpus: helping researchers and teachers clearly identify fluency issues during the learner’s reading process by distinguishing between natural and non-natural pauses.

Exclusion of punctuation marks

In MICICL’s reading-aloud data transcription system, no written punctuation marks are used. This design principle is based on the fundamental difference in symbol system function between spoken and written language. Written punctuation is designed to aid reading and semantic structure and does not perfectly correspond to spoken features like pauses, intonation, or turn boundaries. However, in some transcription systems, researchers often borrow written punctuation to denote prosodic features or discourse boundaries, which can lead to ambiguity. In transcription practice, the meaning of punctuation often differs from its written function (Du Bois et al., 1993; Hepburn & Bolden, 2017). Mixing punctuation in interlanguage spoken corpus transcription could lead to misinterpretation as grammatical pauses or syntactic segmentation, thereby obscuring the learner’s genuine phonetic features and fluency performance.

Therefore, MICICL avoids using written punctuation in its transcription symbol system, opting instead for dedicated transcription labels (such as /ZT, /FZT, /en, etc.) to represent spoken phenomena like pauses and hesitation. This approach ensures a one-to-one correspondence between the transcription symbols and the phonetic phenomena, avoiding interference from the written language system, while making the data formally unified and computable.

Handling of non-reading content

During the reading task, learners occasionally interrupt to explain or ask questions about a segment of the text. Such non-reading speech is also considered part of the corpus and is transcribed in the normal Pinyin format to fully present the learner's authentic performance during the task. If the learner uses a foreign language for explanation or comment, the original foreign language text is retained and subsequently labelled (by using the label [/W]) in the annotation layer to reflect code-switching behavior. This method ensures the authenticity of the data and provides a basis for analyzing the learner's language mobilization and cross-code behavior during task execution.

Recording of intonation

A common occurrence in the corpus is the rising intonation at the end of a clause. This rising intonation may be due to the reading of an interrogative sentence, or it may arise from the learner's lack of fluency or self-doubt. The latter can indicate the learner's uncertainty about the content being read or reflect hesitation and lack of proficiency in the pronunciation of specific words. To preserve this phonetic feature, MICICL marks intonation changes in the transcription. Any noticeable rising intonation at a sentence end or semantic boundary is annotated as /SY (上扬 *shàngyáng*, rising), allowing for subsequent quantitative analysis of speech fluency and intonation control. Explicitly marking the rising intonation helps teachers or researchers identify points of uncertainty in the learner's comprehension or pronunciation, thereby providing data support for phonetic instruction and individualized feedback.

4.4.4 Annotation of reading-aloud data

Annotation format and content

Similar to the processing order for handwritten data, annotation of MICICL's reading-aloud data was performed after the transcription was complete. The annotation system for the reading section is divided into three structural levels: Syllable, Segment, and Phrase, corresponding to the symbols: </>, [/], and {/}, respectively.

This study did not set a "Sentence" level for the reading-aloud data because the only phonetic feature likely to affect the syntactic level in reading material is the intonation pattern,

especially rising intonation. Since this feature has already been recorded in the transcription layer using a dedicated symbol (e.g., /SY), there is no need to redundantly establish a Sentence tier at the annotation level.

The design of the three annotation tiers is based on the linguistic features of the reading-aloud data and the practical research needs:

Syllable level (</>)

Chinese uses a monosyllabic writing system, where each character corresponds to a single syllable unit. Therefore, syllable annotation in the reading-aloud data directly applies to the character level and is used to record the learner's accuracy in syllable pronunciation (e.g., errors or deviations in initials, finals, and tones).

Segment level ([/])

The Segment tier primarily targets the annotation of foreign language lexical items (such as English or Italian) that appear during the reading. Since foreign words often contain multiple syllables and differ structurally from the Chinese monosyllabic system, the Segment tier annotation is used to identify the overall phonetic production of non-Chinese words.

Phrase level ({})

The Phrase tier is utilized to mark phenomena such as repetitions and self-corrections during reading-aloud. MICICL's actual reading-aloud data shows that when learners hesitate or make an error, they often tend to re-read a phrase unit. The length of this phrase unit is not fixed; it could be a multi-character phrase or a single character. Regardless of the phrase length, a unified bracket marking method is used to ensure consistency in the annotation format.

This layered design enables the MICICL reading corpus to simultaneously capture both micro-level phonetic errors (syllable and segment) and macro-level fluency features (repetition and correction). This achieves a balance between annotating phonetic accuracy and fluency, while also maintaining structural consistency with the hierarchical system used for the written corpus.

Annotation labels and definitions

Table 4.7 presents the syllable-level annotation labels in MICICL, which are used to annotate segmental and tonal deviations in learners' reading-aloud production.

Table 4.7*Syllable level annotation labels and definitions⁵⁶*

Code	Name	Annotation content	Example
</DC1/2 /3/4/0>	声 调 错 误 (<i>ShēngDiào Cuòwù</i>)	tone error	Pic-trans: 我是意大利人 Audi-trans: wo1 shi4 yi1 da1 li1 ren2 / Audi-anno: <wo1/DC3> shi4 yi1 <da1/DC4> <li1/DC4> ren2
</SC>	声 母 错 误 (<i>Shēngmǔ Cuòwù</i>)	errors in the pronunciation of the initial (consonant) of a syllable	Pic-trans: 我爸爸 Audi-trans: wo3 pa4 pa0 Audi-trans: wo3 <pa4/SC> <pa0/SC>
</YC>	韵 母 错 误 (<i>YùnmǔCuòwù</i>)	errors in the pronunciation of the final (vowel or vowel combination) of a syllable	Pic-trans: 很多大的城市 Audi-trans: hen3 dou1 da4 de0 cheng2 shi4 Audi-anno: hen3 <dou1/YC> da4 de0 cheng2 shi4
</QC>	全 部 错 误 (<i>QuánbùCuòwù</i>)	A case where both the tone and the syllable pronunciation are incorrect (e.g., tone + final, tone + initial)	Pic-trans: 我跟我的同学常常去 看电影 Audi-trans: wo3 gen1 wo3 de0 tong2 xue2 chang2 chang2 chu1 kan4 dian4 ying3 Audi-anno: wo3 gen1 wo3 de0 tong2 xue2 chang2 chang2 <chu1/QC> kan4 dian4 ying3

⁵⁶ Capital letters in the Pinyin forms of the label names reflect the actual letter(s) used in the corresponding annotation tags, rather than sentence-initial capitalization.

</BDC>	变调错误 (<i>BiànDiào Cuòwù</i>)	A case where tone sandhi should occur but was not applied	Pic-trans: 我们一起去 Audi-trans: wo3 men2 yi1 qi3 qu4 Audi-anno: wo3 men2 <yi1/BDC> qi3 qu4
</BYZ>	不一致(<i>BùYìZhì</i>)	Inconsistent with handwriting content	Pic-trans: 我姐姐二十四岁 Audi-trans: wo2 jie3 jie1 shi2 er4 si4 sui4 Audi-anno: wo2 jie3 jie1 <shi2/BYZ> <er4/BYZ> si4 sui4

Note. The examples provided in the table are authentic data from the MICICL corpus. However, in order to more clearly illustrate the corresponding labels, information outside of the demonstrated labels has been displayed after correction.

Some further explanations:

- </DC1/2/3/4/0> (Tone Error): The number following *DC* indicates the correct tone that should have been produced for that character.
- </BDC> (Unchanged Tone): This tag is used when the learner pronounces the character's base tone correctly, but fails to apply a mandatory tone sandhi rule in sequence with surrounding words.
- {/BYZ} (Inconsistency with Written Text):
 - i. Misreading: For example, the written text is “老古 (*lǎo gǔ*)”, but the learner read “gu2 lao3”.
 - ii. Correct Utterance of Incorrectly written Character: This also covers cases where the learner wrote incorrectly used characters, but read the text as if they had written the target character. For instance, if they wrote “垂觉 (*chuí jiào*)” instead of 睡觉 (*shuì jiào*), but still read it as “shui4 jiao4”.

Table 4.8 presents the segment-level annotation labels in MICICL, focusing on the identification of foreign words and other segment-level features in learners' oral production.

Table 4.8*Segment level annotation labels and definitions⁵⁷*

Code	Name	Annotation content	Example
[/W]	外 文 (<i>Wàiwén</i>)	A foreign word	Pic-tran: 她在 post office 工作 Audi-trans: ta1 zai4 post office gong1 zuo4 Audi-anno: ta1 zai4 [post office/W] gong1 zuo4

Table 4.9 lists the phrase-level annotation labels in MICICL, which are used to annotate discourse-related phenomena such as repetition, self-correction, and explicit explanation in learners' reading-aloud data.

Table 4.9*Phrase level annotation labels and definitions*

Code	Name	Annotation content	Example
{/CF} {/ CFH}	重复/重复 后 (<i>ChóngFù/ ChóngFù Hòu</i>)	Repeated content	Pic-trans: 我学了一点儿中国文化 Audi-trans: wo3 xue2 le0 yi4 dian3 er0 zhong1 guo2 zhong1 guo2 wen2 hua4 Audi-anno: wo3 xue2 le0 yi4 dian3 er0 {zhong1 guo2/CF} {zhong1 guo2/CFH} wen2 hua4
{/XZ} {/ XZH}	修正/修正 后 (<i>XiūZhèng / XiūZhèng Hòu</i>)	Self-correction	Pic-trans: 我跟我爸爸去过中国 Audi-trans: wo3 gen1 wo3 ba4 ba0 qu4 guo0 gong2 zuo2 qu4 guo0 zhong1 guo2 Audi-anno: wo3 gen1wo3 ba4 ba0 {qu4 guo0 <gong2/QC> <zuo2/SC>/XZ} {qu4 guo0 zhong1 guo2/XZH}

⁵⁷ Capital letters in the Pinyin forms of the label names reflect the actual letter(s) used in the corresponding annotation tags, rather than sentence-initial capitalization.

{/JS}	解 释 (<i>JiěShì</i>)	Explanation	Pic-trans: 因为我们有不同的观点 Audi-trans: yin1 wei4 wo3 men2 im not sure about this one you3 bu4 tong2 de0 guan1 dian3 Audi-anno: yin1 wei4 wo3 men2 {im not sure about this one{/JS} you3 bu4 tong2 de0 guan1 dian3
-------	--------------------------	-------------	---

Some further explanations:

- **{/CF}** and **{/CFH}**: Used to mark repetitions made by the learner during reading-aloud aloud. The repeated segment is enclosed by [/], followed by /CF at the end to indicate a repetition. If the segment is repeated multiple times, /CF is added accordingly for each repetition. The label /CFH is used to mark the final repetition in a series of repeated utterances.
- **{/JS}**: Used to annotate verbal explanations made by the learner to the researcher that are not part of the original text. These utterances sometimes occur during the reading process, and learners may provide explanations or comments in either Chinese or English.
- **{/XZ}** and **{/XZH}**: Used to mark self-corrections made by the learner while reading aloud. /XZ marks the incorrect segment that was immediately followed by a self-correction, while /XZH marks the corrected segment. The corrected segment may still contain pronunciation errors. The distinction between repetition and self-correction lies in the nature of the repeated content: repetition involves the same pronunciation being repeated, whereas self-correction involves a change in pronunciation, showing that the learner is attempting to produce a more accurate sound.

4.4.5 Correction of reading-aloud data

The correction of the reading-aloud data primarily utilizes the Pinyin + numerical tone format, referencing the image of the written text to transcribe the correct pronunciation.

The following operations are performed:

- **Filler Words and Explanations:** Filler words and explanatory comments are retained in both the transcription and annotation tiers.
- **Repetitions:**
 - i. The content within the [/CF] labels is deleted, but the label itself is retained.
 - ii. Only the content of [/CFH] (the final repetition) is retained.
- **Self-Corrections:**
 - i. The content of [/XZ] (the error segment) is deleted, but the label itself is retained.
 - ii. Only the content of [/XZH] (the corrected segment) is retained, and any existing syllable errors within this segment are modified.
- **Pauses:** Based on the writing data in the image text, Natural Pauses (“/ZT”) are annotated in places where a pause is necessary (such as between clauses or short sentences). The Non-Natural Pause label “/FZT” is deleted from the transcription and annotation tiers.
- **Foreign Language Data:** Foreign language data is annotated only and not translated.

4.4.6 Reading-aloud data processing overview

A reading-aloud data transcription overview can be seen from Table 4.10:

Table 4.10

Reading-aloud data transcription overview

Content type	No. of occurrences
Tone 0	2651
Tone 1	9324
Tone 2	8731
Tone3	5698
Tone4	7941
/en	1054
foreign words	1251
/SY	13
/FZT	101

The result of summing up all the tones is the total number of syllables in the Chinese characters read by the learners, which is 34345. Reading-aloud data annotation overview can be seen from Table 4.11.

Table 4.11

Reading-aloud data annotation overview

Annotation label	No. of occurrences
DC0	196
DC1	1410
DC2	1482
DC3	2308
DC4	3742
W	1251
BYZ	226
CF	369
CFH	343
XZ	187
XZH	169
JS	22

In total, the corpus contains 9,364 syllable-level annotations, 1,251 segment-level annotations, and 1,090 phrase-level annotations. All annotations were revised accordingly during the revision process.

Chapter 5. Character Level Handwritten Data Analysis

5.1 Introduction

As established in Section 1.3, this study takes the Chinese character as the fundamental unit of analysis. Building on this foundation, this chapter focuses on examining the handwritten data to uncover how learners construct, represent, and negotiate orthographic knowledge during the process of Chinese character acquisition.

The analysis in this chapter is organized around three major types of character-related deviations that emerged in the MICICL handwritten data: incorrectly written characters (the /C tag), incorrectly used characters (the /B tag), and the use of Pinyin (/P, /PC, /PQ, /PZ tags). Detailed definitions of these annotation tags are provided in Section 4.3. These categories, while distinct in form, are closely interrelated in reflecting how learners' orthographic systems develop and how the learners cope with difficulties in Chinese character production. Each section investigates these phenomena from two complementary perspectives: between learners (how many and which learners produced each type of error) and between character types (which Chinese characters were most frequently involved), in order to reveal both inter-learner regularities and inter-character patterns.

Section 5.2 focuses on incorrectly written characters, which are represented by the symbol “*” in MICICL (see Section 4.5). Far from being random mistakes, these instances generally display a certain internal logic within learners' interlanguage orthography. By examining the frequency and the distribution of these cases, Section 5.2 explores how structural and visual features of Chinese characters influence learners' writing performance and reveal their developing orthographic competence.

Section 5.3 turns to incorrectly used characters. Their occurrence stems from the intricate relationship between sound and form in the Chinese writing system, where homophones, near-homophones, and graphically similar characters frequently overlap. For learners, insufficient consolidation of these correspondences can lead to substitution errors.

Section 5.4 addresses another phenomenon closely related to orthographic uncertainty—the use of Pinyin. Many learners in the MICICL corpus occasionally replaced target Chinese characters with Pinyin forms, sometimes mixing Pinyin and characters within the same word or phrase. These instances are annotated with /P, /PC, /PQ, and /PZ, representing various forms of Pinyin use, such as missing tones, partially correct spellings, or co-occurrence with characters. The analysis of Pinyin substitution reveals how learners employ phonologically driven compensatory strategies to maintain textual coherence and communicative flow when orthographic retrieval fails. It also sheds light on the evolving relationship between sound and form in learners’ interlanguage orthography.

In the MICICL corpus, learners occasionally incorporated foreign words into their handwritten texts. These instances are annotated as /W and are briefly discussed in Section 5.5. As this phenomenon primarily concerns word-level rather than character-level analysis, it is only addressed in a concise overview without extensive quantitative examination.

Finally, Section 5.6 provides a comparative analysis and synthesis of the first three major categories—incorrectly written characters, incorrectly used characters, and Pinyin use—summarizing their distributional features and implications for Chinese character acquisition among L1 Italian learners.

Through this multi-level examination, Chapter 5 aims to construct a comprehensive picture of learners’ handwriting performance at the character level, highlighting the systematicity underlying different types of orthographic deviations. Rather than viewing these phenomena as mere errors, the chapter interprets them as developmental markers that reveal how learners conceptualize, reconstruct, and internalize the relationship between sound, form, and meaning in the process of learning to write Chinese.

5.2 Character writing: incorrectly written characters

5.2.1 Research questions

Building on previous research on Chinese character writing (see Section 1.3) and addressing the identified research gap, this section examines the overall distribution of incorrectly written

characters in the MICICL writing data, drawing on the machine-readable transcription, annotation, and correction layers.

Subsequently, I conducted a multimodal classification analysis of these errors by incorporating the corresponding image data from MICICL. While text-based data provide information on frequency and distribution, the transcription process necessarily replaced illegible handwriting forms with symbolic markers, thus losing the specific visual and structural features of the learners' written output. By introducing the original image data, I was able to reconstruct the learners' actual writing attempts and classify incorrectly written characters more precisely according to their form and structural characteristics. This allowed for an in-depth exploration of *how* L1 Italian learners made writing errors, rather than simply *where* or *how often* these occurred.

The multimodal analysis examines patterns of character writing in L1 Italian learners' Chinese output by integrating handwriting and reading-aloud data, with the aim of providing empirically grounded evidence for Chinese orthography teaching to learners from alphabetic-language backgrounds. Writing errors are treated not as random "slips of the pen" but as systematic deviations observable in learner output. By analyzing the distribution and types of these errors across different levels of graphical organization, the study identifies recurrent difficulty areas and formal characteristics of learners' written production, thereby offering a detailed description of how Chinese characters are produced and patterned in learner writing.

Therefore, this section addresses the following two research questions:

- i. What is the overall distribution of incorrectly written characters produced by L1 Italian learners in the MICICL writing data? This question is explored by analyzing the frequency and distributional patterns of such errors across learners and across specific Chinese characters.
- ii. What are the internal differences among various error types? The ratios of distinct incorrectly written character types are compared to identify differences in learners' mastery across character hierarchy levels. The internal concentration within each

category is further analyzed to determine whether the errors reflect group-wide systematic patterns or idiosyncratic individual habits.

The following points provide additional clarifications:

- i. The focus of this section is on the phenomenon of “unformed characters”—instances where the writing failure is so severe (e.g., excessive or missing strokes, misplaced components, disrupted structure) that the resulting form is illegible. This corresponds to instances labeled /C in the MICICL corpus. Minor inaccuracies that still yield a legible character are not discussed here, while fully formed but contextually incorrectly used characters (labeled /B) are examined in Section 5.3.
- ii. In this study, the target character refers to the standard Chinese character that the learner intended to write, while the incorrectly written character refers to the illegible form actually produced due to errors in stroke, component, or structure.

5.2.2 Distribution of incorrectly written characters among learners

Incorrectly written characters appeared in 117 out of the 155 writing samples in the MICICL corpus, a prevalence rate of 75.48%. It should be noted that a writing sample does not correspond to a single learner: the 155 samples were produced by 39 learners, with each learner contributing multiple samples (see Section 3.5). Within these samples, I tallied a total of 607 incorrectly written character tags (the /C tag), which accounted for 1.84% of the corpus’s total character count of 33031 (excluding punctuation, number and foreign words). Among all these 607 tags, 559 were for characters that appeared alone (only marked as “/C”), while the remaining 48 appeared in combination with other error tags (such as “/C+PZ” or “/C+PQ”). The combination tags indicated that learners not only produced a graphic error but may have simultaneously resorted to strategies like Pinyin substitution, reflecting their uncertainty regarding the character and their reliance on multiple recovery strategies. At the character level annotation, the MICICL corpus recorded 3,496 error tags in total (excluding punctuation and lexical errors); the incorrectly written character tag accounted for 17.36% of this total.

A clearer understanding of the incorrectly written character situation in MICICL can be obtained by calculating the proportion of incorrectly written characters among all the characters

written or attempted by each learner. This ratio is derived by dividing the number of incorrectly written characters by the total number of characters, where the total includes correct Chinese characters, incorrect characters, and Pinyin substitutions, but excludes numbers, foreign words and punctuation marks.

$$\text{Student's Incorrectly Written Character Ratio} = \frac{\text{Total Number of Incorrectly Written Characters Produced by the Student}}{\text{Total Number of Characters Produced by the student}} \times 100$$

All calculation results are rounded to two decimal places. The same formula and rounding convention are applied to calculate the error ratios for other groupings (e.g., students of different proficiency levels) throughout this section.

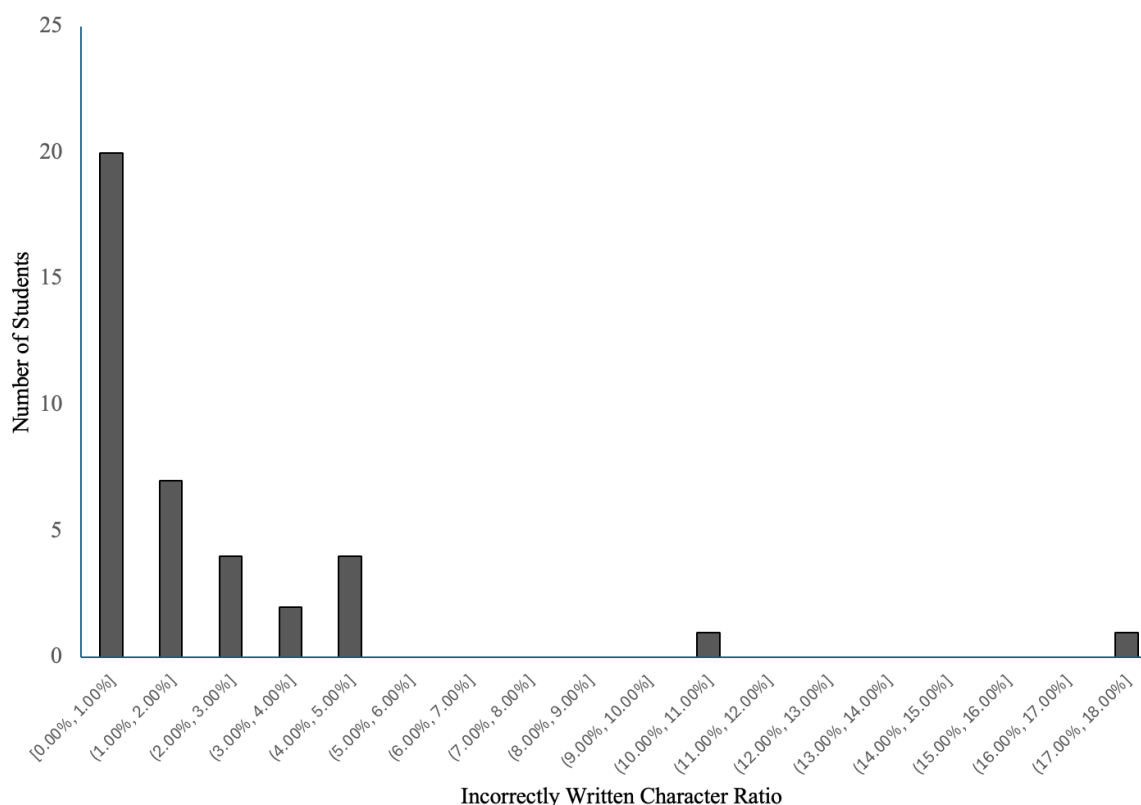
The incorrectly written character ratio of 39 students is shown in Appendix D. 32 out of the 39 participants produced incorrectly written characters in their data. The 7 learners who did not produce this error type showed varied circumstances: Student No.12 (Advanced) produced 989 characters and 573 character types across 6 data collections without a single incorrectly written character, demonstrating a high level of Chinese character mastery. Student No.17 (Beginner) participated in only one collection and, due to weak Chinese character writing skills, wrote almost the entire content in Pinyin (64/65), resulting in zero errors of this type. Student No. 18 (Advanced), Student No. 33 (Intermediate), Student No. 34 (Advanced), Student No. 35 (Intermediate), and Student No. 36 (Intermediate) each participated in only one session. Although their data were limited in quantity, all of them showed clear, well-structured writing with stable control over character forms, indicating good overall orthographic competence.

I found that the highest proportion of incorrectly written characters was produced by Student No. 16 (Beginner), reaching 17.86%. However, this learner produced only 10 incorrectly written characters in total, and this high proportion was skewed by the short data length (participating in only one collection with a total of 56 characters). The learners who produced a high absolute number of incorrectly written characters were Student No. 7 and Student No. 4, who produced 65 and 62 errors, respectively. Student No. 7 participated in eight

data collections with a total of 1889 characters, while Student No. 4 participated in six collections with a total of 1450 characters.

Figure 5.1

Number of students by incorrectly written character ratio



As shown in Figure 5.1, the majority of students fell within the 1%–3% range, meaning that about 1 to 3 characters out of every 100 were ill-formed. Incorrectly written characters were observed in a large proportion of learners (32 out of 39), indicating that such errors are a widespread phenomenon rather than isolated cases. However, the instructional and acquisitional value of these errors lies not merely in their frequency, but in the extent to which they reveal systematic patterns in learners’ character production. As will be shown in the following sections, many of these errors are not random, but reflect recurring difficulties at the level of strokes, components, and character structure, which can provide useful insights for both teaching and learning.

5.2.3 Distribution of incorrectly written characters by character type

Based on the correction information provided by MICICL, the specific target character corresponding to every incorrectly written character could be determined. Using the target character as the reference point for the following descriptive statistics, a total of 245 incorrectly written character types were identified, meaning 245 different Chinese characters had been incorrectly written into a unformed shape in the learners' production. The specific character forms and their frequencies are detailed in Appendix E.

Table 5.1

Top 10 target characters most frequently written incorrectly by students

Target character type	No. of students who wrote the character incorrectly	No. of students who wrote the character correctly
喜(xǐ)	10	22
或(huò)	9	9
帮(bāng)	7	11
住(zhù)	6	18
春(chūn)	5	5
知(zhī)	5	18
候(hòu)	4	25
夏(xià)	4	9
爱(ài)	4	22
影(yǐng)	4	16
望(wàng)	4	8
旅(lǚ)	4	14
漂(piāo)	4	15

Table 5.1 displays the top ten Chinese target characters whose incorrectly written forms were produced by the highest number of unique students in the MICICL corpus. It should be specially emphasized here that: the number of students who wrote a given character incorrectly is not equivalent to the number of students who wrote it correctly, nor to the total number of students who used that character.

This is because the same student may have written the character correctly in one task but

incorrectly in another. The two numbers are neither additive nor mutually exclusive; they simply reflect how many students have ever produced an error for that character and how many students have successfully produced it. Therefore, this statistic reflects not an “error rate” but the scope of correct or erroneous activation of the character within the learner population—that is, its cross-learner distribution.

The data shows that these characters exhibit clear convergence and commonality among the student population. For example: 或 (*huò*): 9 students had incorrectly written the character, and 9 students had correctly written it. These two groups of students may be identical or partially overlapping; it cannot be simply interpreted as “9 students wrote it incorrectly”.

From the overall trend, incorrectly written characters are not individualized or accidental errors but a systematic phenomenon that repeatedly appears across different learners. Some characters (like 或(*huò*) and 春(*chūn*)) showed errors almost universally among those who wrote them, indicating a general writing difficulty at the structural or component-combination level. Characters such as 喜(*xǐ*), 帮(*bāng*), and 住(*zhù*), while not incorrectly written by all students, show recurrent errors across a large number of learners, revealing a common weak point in their visual recognition and orthographic processing. In contrast, the error distribution for characters like 候(*hòu*), 爱(*ài*), 影(*yǐng*), and 旅(*lǚ*) is more dispersed, yet they still present a stable cross-learner characteristic, suggesting that while these errors are not highly concentrated, they possess a degree of commonality.

To gain a clearer understanding of how the phenomenon of incorrectly written characters is distributed across different target Chinese character types, I analyzed the target character types with the highest frequency of incorrectly written phenomena in the MICICL corpus.

In calculating the total frequency of each target character, I included not only the Chinese characters correctly written by the learners but also the target characters corresponding to the incorrectly written characters (i.e., characters the learner attempted but failed to complete correctly) and the target characters corresponding to the Pinyin substitution phenomenon. This calculation method comprehensively reflects the learner’s mastery of each target character type, regardless of whether that character type appeared in the output as correct, erroneous, or in

Pinyin form.

It must be pointed out that the target characters corresponding to incorrectly used characters were not included in the total frequency count of target characters. Unlike incorrectly written characters, the target character of an incorrectly used character did not actually appear in the corpus, because the learner wrote a different, legitimate Chinese character instead. Including these non-occurring “hypothetical target characters” would not only cause frequency inflation but also violate the principles of frequency calculation based on the actual corpus. Therefore, the total frequency of each character in this statistics includes only three categories:

- i. The correctly written Chinese characters actually produced by the learner;
- ii. The target characters corresponding to the incorrectly written characters;
- iii. The target characters corresponding to the Pinyin substitution phenomenon.

In calculating the ratio of incorrectly written instances for each target character, I set a minimum frequency threshold (≥ 10 times); i.e., quantitative analysis was only performed on characters that appeared at least 10 times in the corpus. This setting is designed to prevent bias caused by low-frequency characters, avoid the interference of extreme values on the overall trend judgment.⁵⁸

At the same time, to ensure that results reflect a common phenomenon across learners, rather than just one student’s personal writing style, memory lapse, or a one-off operational error, I also set a minimum student number threshold: incorrectly written characters that appeared in the writing of only 1 student are excluded from the statistics in Table 5.2. That is to say, even if a certain character was repeatedly incorrectly written by one student, I still excluded it if it only appeared in that single student’s output. For example, the character 京 (jīng) appeared 17 times in the corpus and was incorrectly written 14 times by 1 student, giving it the highest error rate of 82.35%. I excluded it because it lacks the characteristic of “cross-learner distribution” and therefore cannot be said to represent a group-level difficulty for Italian native speakers.

⁵⁸ A considerable number of Chinese characters in the MICICL corpus only appeared once or twice; if an incorrectly written instance happened to occur once, its error rate would be calculated as 100%. Such a ratio does not genuinely reflect the learner’s mastery of that character but is merely an isolated individual event.

Table 5.2

The 10 character types with the highest frequency of incorrectly written characters

Target character type	Total frequency	Incorrectly written times	Incorrect ratio	No. of students who wrote the character incorrectly
春(<i>chūn</i>)	29	16	55.17%	5
争(<i>zhēng</i>)	10	5	50.00%	2
庭(<i>tíng</i>)	11	5	45.45%	3
或(<i>huò</i>)	40	17	42.50%	9
夏(<i>xià</i>)	35	11	31.43%	4
言(<i>yán</i>)	16	5	31.25%	2
假(<i>jià</i>)	22	6	27.27%	2
喜(<i>xǐ</i>)	180	47	26.11%	10
趣(<i>qù</i>)	12	3	25.00%	3
秋(<i>qiū</i>)	13	3	23.08%	3

From the data in Table 5.2, the incorrectly written phenomenon shows clear distribution stratification across different character types.

The three characters 喜 (*xǐ*), 或 (*huò*), and 春 (*chūn*) have high total occurrences and high error counts, forming the most representative high-frequency error types. 喜(*xǐ*) appeared 180 times in total and was incorrectly written 47 times (26.11%); 或(*huò*) appeared 40 times and was incorrectly written 17 times (42.50%); 春(*chūn*) appeared 29 times and was incorrectly written 16 times (55.17%). These characters were also ranked highly in Table 1, indicating that they are the characters incorrectly written by most learners, and also the characters most likely to be incorrectly written among all written instances. They possess the dual features of “distribution breadth” and “error concentration”.

In contrast, the total occurrences of characters like 庭 (*tíng*), 争 (*zhēng*), 假 (*jià*), 言 (*yán*), 趣 (*qù*), and 秋 (*qiū*) are relatively low (between 10–22 times). However, their incorrectly written count still ranges from 3–6 times, with an error ratio between approximately 25%–50%. The characteristic of this type of character is that although the frequency of appearance is not high, the proportion of errors within the limited output is large, exhibiting strong intra-character concentration. In other words, these characters have a higher probability

of error once they are written.

A combined view of Table 2 and Table 1 provides a more comprehensive insight into the two levels of the learner's incorrectly written phenomenon:

- Table 5.1 focuses on the “number of students,” reflecting the distribution breadth (cross-learner similarities) of the incorrectly written phenomenon, i.e., which characters were incorrectly written by more students.
- Table 5.2 focuses on the “total occurrences” and “incorrectly written count,” reflecting the concentration (intra-character concentration) of the incorrectly written phenomenon at the character type level, i.e., which characters are more likely to be written erroneously when repeatedly written.

It is worth noting that some characters (e.g., 或, 春, 喜) rank high in both tables, these characters possess both high group-level distribution breadth and strong intra-character error concentration. Conversely, some characters that only appear in Table 5.2 but not in Table 5.1 (e.g., 庭, 言, 趣) reflect another type of error characteristic: while their error ratio is high, the errors are concentrated among a small number of students, indicating a type of locally high-incidence error with high concentration but narrow distribution. Characters that only appear in Table 1 (e.g., 爱, 旅, 望) indicate that their error distribution range is wide, but the overall error ratio is low, reflecting an “easily confused but controllable” error type.

Overall, the complementary analysis of Table 5.1 and Table 5.2 reveals two layers of the incorrectly written phenomenon in the Chinese character writing acquisition of Italian native speakers:

- i. At the commonality level, certain characters repeatedly show incorrectly written instances across multiple learners, presenting stable cross-individual consistency. This suggests that learners universally face difficulties in mastering certain character types, exhibiting a clear group distribution characteristic.
- ii. At the individual level, some characters, although appearing in only a few learners' output, repeatedly show errors when written multiple times, demonstrating a persistent error tendency within the individual.

This dual structure indicates that learners' incorrectly written instances involve both common rules shared across learners and stable patterns driven by individual differences. From a pedagogical perspective, this result suggests that teachers designing writing instruction should consider both levels: for characters with high frequency and a high incorrectly written ratio, the focus should be on reinforcing component awareness, structural rules, and structural contrast to help learners establish clear orthographic concepts. For characters with high distribution but a low ratio, the focus should be on contextual discrimination and orthographic recognition practice to improve the learner's ability to distinguish between similar characters and reduce the occurrence of repetitive errors.

5.2.4 Incorrectly written characters by category

Before classifying incorrectly written characters, it is useful to recap the basic constitutive elements of the Chinese character.

From the analysis in Section 1.3, it can be seen that if stroke order errors (which this study did not address) and character misuse (which was discussed in Section 5.3) are excluded, existing research on the classification of incorrectly written characters largely employs a dichotomous approach, dividing errors into the stroke level and the component level.

However, if the classification criterion is explicitly defined as the “hierarchical level where the error occurs,” then the structure level must be included. This is because for compound characters, even if all components themselves are correctly written, a deviation in their positional relationship during the combination process cannot be simply categorized as a component error; it must be localized as an error at the structural level. Although existing studies have involved this type of error to varying degrees, most still subsume it under categories like “component mirror errors,” placing it at the component level. Only Wei (2017) and Feng (2019) established the structure level separately: However, Wei's definition of structural errors leaned more toward the quality of the character's presentation (e.g., aesthetics or standardization) rather than strict correctness; and Feng similarly followed this pattern, but additionally classified some incorrectly used characters resulting from graphic similarity errors into the structure error category, such as writing 呆 (*dāi*) as 杏 (*xìng*).

Furthermore, the lack of uniformity in classification criteria and boundary definition is one of the major issues in current research. First, blurred lines often exist in the categorization across different studies, meaning similar error types may be subsumed under different categories in different systems. For example, the examples provided by Wei (2017) for “stroke errors” and “component errors”—classifying writing 目 (*mù*) as 𠂇 (*rì*) as a stroke error, and writing 古 (*gǔ*) as 吉 (*jí*) as a component error—demonstrate a certain logical ambiguity. Both types of errors resulted in a legitimate Chinese character form, and no explicit criteria were provided for why these were defined as incorrectly written characters at the writing level (stroke or component error) rather than as incorrectly used characters at the semantic-use level.

Moreover, some classification criteria rely on the subjective judgment of the researcher, lacking objective and operational standards. For instance, within Wei’s definition of “structure errors,” if miswriting 项 (*xiàng*) as “工页” is considered a structural error, to what extent must the spatial distance between the two components deviate to constitute a structural error? When the arrangement of components shows only minor changes in density, should it be classified as a component error or a structure error? These questions lack standardized quantitative or descriptive criteria in existing research, where researchers often rely on visual inspection for demarcation. This type of subjective judgment directly impacts the comparability and reliability of error statistics and, to a certain extent, limits systematic comparison across corpora and studies.

Finally, the simple dichotomous approach has clear limitations when dealing with real-world data. Learners’ incorrectly written characters are often not single-dimension but compound errors that cross hierarchical levels. For example, a character might simultaneously feature a stroke omission, an incorrect component sequence, and overall structural misalignment. The dichotomous method is unable to fully describe this complexity. Although scholars (e.g., Fan, 1993; Kang, 2010) have proposed categories like “comprehensive errors” or “multiple errors” to cover these cross-level phenomena, most research remains within the dichotomous framework, lacking a systematic description and analysis of cross-hierarchical errors.

Based on the preceding discussion, this study performed a moderate revision and expansion of previous classification systems. First, the structure incorrect characters category was added, strictly limiting errors to the three concrete hierarchical levels of stroke, component, and structure to avoid ambiguity and overlap between categories. Second, a new category, multi-errors characters, was introduced to cover instances where two or more error types appeared simultaneously, or where atypical errors could not be systematically categorized.

In summary, in this section I divided the incorrectly written character into four main categories: stroke incorrect characters (SCs), component incorrect characters (CCs), structure incorrect characters (STCs), and multi-error characters (MCs). Due to constraints of scope and research focus, internal sub-classification of each category is not detailed further in this thesis.

SCs

This category refers to errors that occur solely at the individual stroke level. Typical scenarios include errors in the starting/ending points, direction, turning, length proportion, or inter-stroke relationships (separation, connection, intersection, top/bottom alignment, inside/outside relations). For instance, in Figure 5.2, the “丿” (piě, left-falling) in the character 看 (*kàn*) was written too long, extending above the top boundary; in Figure 5.3, the vertical hook in 除 (*chú*) was too short; and in Figure 5.4, the vertical-turn-hook was placed too far from the “丿” (piě, left-falling) in 现 (*xiàn*), disrupting the necessary continuous relationship. It is important to note that this study classifies the addition or omission of strokes under the component incorrect characters category. This is because, whether a single-component or a compound character, a stroke incorrect character error is only defined as an instance where the number of strokes is correct, but the form or position is inappropriate. In other words, a stroke error implies that the learner has mastered the character’s overall structure and component composition, correctly estimating the required stroke count, but still exhibits deviations in the specific writing action. This type of error reveals that the learner’s mastery of stroke norms, movement control, and “inter-stroke constraint” rules has not yet been fully automated.

Figure 5.2

Example 1 of SCs



Target character: 春(*chūn*)

Source: MICICL 0203042403X

Figure 5.3

Example 2 of SCs

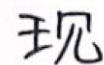


Target character: 除(*chú*)

Source: MICICL 1402022312X

Figure 5.4

Example 3 of SCs



Target character: 现(*xiàn*)

Source: MICICL 0401002305X

CCs

A component can function as a phonetic radical or a semantic radical, or it can be a subcomponent that only serves a structural purpose without phonetic or semantic function. Consequently, errors at the component level can occur in both single-component and compound Chinese characters. If one component of a Chinese character is written incorrectly while all other components are correct, it is judged as a component error. For example, in Figure 5.5, the upper part of the character 高 (*gāo*) was incorrectly written as the 宝盖头 (*bǎogàitóu*, roof radical). In Figure 5.6, the right-side component of 淮 (*zhǔn*) was missing one stroke, resulting in 主 (*zhǔ*). Errors involving the addition or omission of strokes in single-component characters are also included in this category, such as the omission of the dot stroke in 书 (*shū*)

in Figure 5.7.

Figure 5.5

Example 1 of CCs



Target character: 高 (gāo)

Source: MICICL 0401002305X

Figure 5.6

Example 2 of CCs



Target character: 准 (zhǔn)

Source: MICICL 0701002305X

Figure 5.7

Example 3 of CCs



Target character: 书 (shū)

Source: MICICL 1005042403X

This type of error reflects a weakness in the learners' component lexicon and its functional representation; that is, their form recognition and functional differentiation (semantic/phonetic/structural subcomponent) of components are still unstable. It also indicates a deficiency in their awareness of component substitution rules (which strokes or structures can and cannot be substituted) and a low sensitivity to component placement, showing they have yet to master the legitimate distribution patterns of components within the spatial structure.

STCs

This refers to instances where the learner correctly writes the form of all individual components (i.e., the strokes, proportions, and identity of the components themselves are correct), but

makes an error in their two-dimensional combination relationship. These errors manifest as mistakes in the orientation, arrangement, nesting relationship, or relative positioning of components, such as miswriting a left-right structure as an up-down structure, writing an enclosed structure as a side-by-side one, or reversing the component order or relative position. Typical examples include Figure 5.8, 品 (*pǐn*), and Figure 5.9, 爸 (*bà*), which were written upside down. Another example is in 或 (*huò*), Figure 5.10, where the positions of the 横 (*héng*, horizontal stroke) and 口 (*kǒu*, mouth radical) were reversed, leading to an incorrect overall structural template.

Figure 5.8

Example 1 of STCs



Target character: 品 (*pǐn*)

Source: MICICL 0202022312X

Figure 5.9

Example 2 of STCs

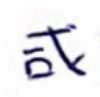


Target character: 爸 (*bà*)

Source: MICICL 0403042403X

Figure 5.10

Example 3 of STCs



Target character: 或 (*huò*)

Source: MICICL 0701002305X

The principle for determining a structure incorrect character error is as follows: if the component form itself is incorrect, it is still classified as a component incorrect character error. Only when the internal strokes of the components are correct but the spatial relationship or

structural template between components is flawed is it classified as a structure incorrect character error. This type of error suggests that the learner has insufficient mastery of the Chinese character spatial configuration schema, failing to accurately grasp the geometric constraints and overall balance of the character framework.

MCs

This refers to the simultaneous occurrence of two or more error types in the same character, or instances where the character form is highly disordered and cannot be categorized into a single hierarchical level. Examples include concurrent stroke omission and component misalignment, or when components are written correctly but the structural configuration is wrong, and simultaneously, one stroke is poorly executed. For instance, in Figure 5.11, the target character 假 (*jià*) was written by the learner with multiple errors and overall imbalance, making single-level classification impossible.

Figure 5.11

Example of MCs



Target character: 假 (*jià*)

Source: MICICL 0204052404X

For this type of error, this study does not force a “single-best-fit” judgment but annotates it accurately as a multi-error character. This approach more faithfully reflects the learner’s actual processing path during writing and prevents the dilution of hierarchical information caused by forced categorization during statistical analysis. Multi-error characters often signal that the learner has weaknesses across multiple levels simultaneously (stroke norms, component representation, structural layout), indicating the overall instability of their writing processing system.

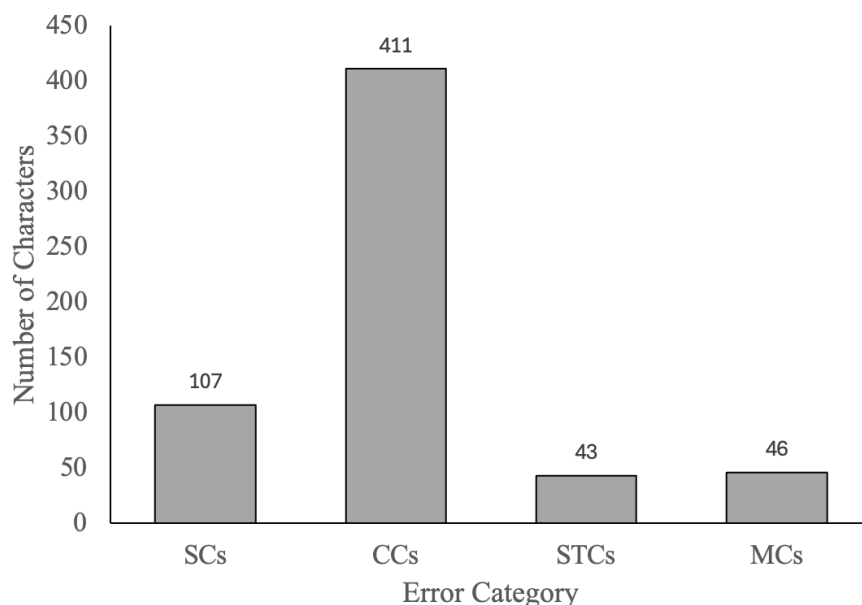
5.2.5 Categorical analysis of incorrectly written characters

Following the classification rules outlined above, and by combining the original handwriting forms from the image data, the classification results for the 607 incorrectly written characters

identified in the MICICL corpus are shown in Figure 5.12.

Figure 5.12

Classification results of incorrectly written characters in MICICL



From the overall distribution, the proportions of the four types of errors were highly imbalanced: CCs held the dominant position, SCs were substantially fewer, while STCs and MCs occurred at similarly low frequencies, each accounting for only a small proportion of the total errors:

First, CCs accounted for the absolute majority, this indicates that the primary bottleneck for learners in Chinese character acquisition is concentrated at the component level of recognition and processing, a finding consistent with the statistical results of Deng & Hu (2022) and Wei (2017).

Next, the quantity of SCs was relatively low, their ratio (17.63%) of the total incorrectly written characters being markedly less than that of CCs. This suggests that learners have formed a preliminary visual template for the overall character shape and stroke count, but still experience difficulty with fine-grained control at the stroke level. This primarily manifests as an imbalance in stroke length proportion, errors in inter-stroke relationships (improper separation, intersection, or connection), and deviations in stroke starting and ending points. This type of error indicates that the learner's writing movements have not yet achieved complete procedural control. That is, stroke generation still relies on visual imitation rather

than a motor generation mechanism dependent on internalized rules. For L1 Italian speakers, this phenomenon partly stems from the linear, alphabetic nature of their native writing system: letter strokes have limited variation and simple form rules, making the writing process closer to static reproduction than dynamic motor planning. When this linear writing model is transferred to Chinese characters, learners tend to neglect the dynamic dependency relationship between strokes, leading to a disconnect between writing fluency and stroke normativity. Therefore, stroke incorrect characters reflect a problem not of a holistic misunderstanding of the character structure, but rather a lack of fine motor control and awareness of stroke proceduralization in the writing motor process.

Third, although the ratio of STCs was low (7.08%), this error type is still noteworthy. These errors predominantly occur when the learner can correctly write the form of all individual components but is unable to organize the component relationships rationally in two-dimensional space. This type of error reflects the learner's persistent difficulty with spatial processing and structural template mastery (spatial compositional schema). Unlike the linear arrangement of alphabetic scripts like Italian, Chinese characters constitute a visual-spatial symbolic system based on a two-dimensional structure, requiring the learner to integrate the relative positions of strokes and components on a plane (Li, 2014). For L1 Italian speakers, this non-linear spatial layout lacks a direct native language reference, causing them to rely on perceptual position estimation rather than abstract morphographic rules during writing. In other words, their abstract cognition of Chinese character spatial organization remains unstable; even when component forms are mastered, maintaining structural consistency and symmetry in writing is challenging.

Finally, MCs were the least frequent (error ratio 7.58%), but their research value should not be overlooked. This type of error is often accompanied by cumulative errors across the stroke, component, and structure levels, revealing bottlenecks in multiple stages, including character form recall, visual representation, and stroke execution. For L1 Italian speakers, this breakdown phenomenon can also be explained as a result of cross-orthography transfer: their native writing process relies on stable sound-form mapping, and the lack of such clues in

Chinese characters makes it difficult for them to initiate the form generation mechanism when phonological support is unavailable (Baroni, 2011).

It is also necessary to further analyze the high-frequency character types and their degree of concentration within each category. Merely knowing the total number of errors in each category was insufficient to reveal the learners’ true difficulties in writing—some categories might be concentrated on a few “high-risk” characters, while others might have a more dispersed error distribution. By examining whether errors within each category are concentrated on a limited number of characters or dispersed across many items, this analysis clarifies whether the observed patterns are driven by a few particularly difficult characters or represent more general and systematic difficulty patterns at a given hierarchical level. This analysis aids in more accurately judging the learners’ mastery across different levels (stroke, component, structure) and provides more specific direction for subsequent pedagogical intervention.

SCs

The top-5 character types in SCs are showed in Table 5.3. Combined with the image data, it can be seen that learners’ stroke-level errors exhibit a highly consistent trend of concentration. That is, the same character often shows stroke errors at the identical position and in similar forms across the writing of different learners. This indicates that stroke errors are not randomly distributed but rather present systematic “vulnerability points” within the character structure.

Table 5.3

Top 5 SCs target character types

Target character Type	Frequency	No. of students
喜 (xǐ)	34	7
京 (jīng)	14	1
看 (kàn)	6	2
书 (shū)	5	2
早 (zǎo)	5	3

Character 喜 (xǐ) had the most frequent occurrence of stroke errors (34 instances, involving 7 students). All errors were concentrated in the upper component 士 (shì). Learners commonly miswrote 士 as 土 (tǔ)—specifically, the top horizontal stroke was written too short, and the bottom horizontal stroke was written too long, as shown in Figure 5.13, Figure 5.14 and Figure 5.15. This error repeatedly occurred across multiple learners, suggesting a consistent misjudgment regarding the number of strokes in this component, either in visual memory or structural recognition.

Figure 5.13

Example 1 of character 喜 (xǐ) in SCs



Source: MICICL 1902022312X

Figure 5.14

Example 2 of character 喜 (xǐ) in SCs



Source: MICICL 2801082410X

Figure 5.15

Example 3 of character 喜 (xǐ) in SCs



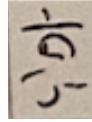
Source: MICICL 3101082410X

For character 京 (jīng), there were 14 stroke errors, produced by only one student. All errors were concentrated in the vertical-bending hook at the bottom of the character. The learner wrote this stroke too short, resulting in a break in the lower structure, as seen in Figure 5.16. This

type of stroke deficiency error remained stable across the same learner's multiple writings, showing clear individualized consistency.

Figure 5.16

Example of character 京 (jīng) in SCs

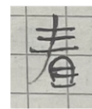


Source: MICICL 0708ZY2311X

There were 6 stroke errors of the character 看(kàn), produced by 2 students, all concentrated at the position of the upper left-falling stroke (撇, piě), as shown in Figure 5.17 and Figure 5.18. Both students exhibited the “protruding left-falling stroke” phenomenon, where the stroke was too long, extending beyond its required structural boundary.

Figure 5.17

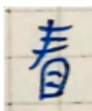
Example 1 of character 看 (kàn) in SCs



Source: MICICL 0203042403X

Figure 5.18

Example 2 of character 看 (kàn) in SCs



Source: MICICL 1108082410X

There were 5 stroke errors of character 书(shū), manifested in the dot at the beginning. Learners consistently misplaced the dot, causing the top of the character to be unbalanced, as seen in Figure 5.19 and Figure 5.20.

Figure 5.19

Example 1 of character 书 (shū) in SCs



Source: MICICL 0701002305X

Figure 5.20

Example 2 of character 书 (shū) in SCs



Source: MICICL 2802092412X

For character 早(zǎo), there were also 5 stroke errors, involving 3 students. The errors were concentrated in the central vertical stroke. Some wrote it too long, while others wrote it noticeably too short, as seen in Figure 5.21, Figure 5.22 and Figure 5.23.

Figure 5.21

Example 1 of character 早 (zǎo) in SCs



Source: MICICL 0702012311X

Figure 5.22

Example 2 of character 早 (zǎo) in SCs



Source: MICICL 2201012311X

Figure 5.23

Example 3 of character 早 in SCs



Source: MICICL 2702092412X

Overall, stroke-level errors form a structural commonality across multiple learners, manifested as an unstable mastery of specific stroke features. This consistency provides important clues for subsequent investigations into the mechanism of learner character-form construction at the component and structural levels.

CCs

The top 5 character types in CCs are shown in Table 5.4. Combining the image data, it can be seen that the error forms and positions of the first four characters are highly consistent, exhibiting obvious common characteristics.

Table 5.4*Top 5 CCs target character types*

Target character Type	Frequency	No. of students
候 (hòu)	23	4
春 (chūn)	12	4
喜 (xǐ)	13	5
夏 (xià)	10	3
帮 (bāng)	9	6

The character 候 (hòu) is the most prominent incorrectly written character in terms of component-level errors, appearing 23 times and concentrated among 4 learners. Specifically, Student No. 2 (Advanced) and Student No. 4 (Advanced) contributed over 90% of the error frequency: the former accumulated 5 errors across three tasks, and the latter accumulated 16 errors across four tasks. The error form for both students across different times and tasks was completely consistent—they both added an extra horizontal stroke above the 矢 (shǐ) component (see Figure 5.24 and Figure 5.25).

Figure 5.24*Example 1 of character 候 (hòu) in CCs*

Source: MICICL 0204052404X

Figure 5.25*Example 2 of character 候 (hòu) in CCs*

Source: MICICL 0404052404X

This repetition across tasks indicates that the error is not accidental but has persisted and remained uncorrected across multiple writing instances, showing a clear fossilization tendency.

This suggests that the learners' error in a specific orthographic pattern has solidified into a stable writing habit, reflecting a persistent processing deviation in the representation and generation of the character at the component level.

Character 春 had a total of 12 component errors, appearing in 4 learners. Three of the learners exhibited the same type of error, miswriting the bottom component 日 (*rì*) as 目 (*mù*), which is a typical component substitution error (see Figure 5.26). Student No. 9 (Advanced), however, exhibited a mixed error of one stroke omission and one stroke addition (see Figure 5.27). This demonstrates that most learners have a consistent error on the same component, reflecting a common pattern in component processing.

Figure 5.26

Example 1 of character 春 (chūn) in CCs



Source: MICICL 0404052404X

Figure 5.27

Example 2 of character 春 (chūn) in CCs



Source: MICICL 0906052404X

Character 喜 (*xǐ*) appeared with 13 component errors, involving 5 learners. The errors mainly manifested as miswriting the upper component 吉 (*jí*) as 古 (*gǔ*) (see Figure 5.28), or omitting the top 口 (*kǒu*) component (see Figure 5.29). Notably, 喜 (*xǐ*) was also a high-frequency error character at the stroke level (SCs), and some students showed both stroke errors and component errors on the same character (e.g., Student No. 4). This indicates instability in both stroke execution and component construction for the character 喜 (*xǐ*) across different levels of writing processing, showing a multi-level overlap characteristic. In other words, the writing error may stem simultaneously from both stroke operation and component formation,

reflecting the learner's failure to form a complete and stable structured cognition in the orthographic representation of this character.

Figure 5.28

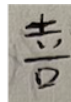
Example 1 of character 喜 (xǐ) in CCs



Source: MICICL 0203042403X

Figure 5.29

Example 2 of character 喜 (xǐ) in CCs



Source: MICICL 2502042402X

夏 (xià) had a total of 10 component errors, all showing the same tendency across 3 learners: simplifying the upper double-horizontal structure into a single horizontal stroke. This is a component reduction error achieved through stroke simplification (see Figure 5.30 and Figure 5.31). The consistent error form across different learners shows a high degree of cross-learner consistency.

Figure 5.30

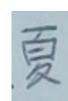
Example 1 of character 夏 (xià) in CCs



Source: MICICL 2405052404X

Figure 5.31

Example 2 of character 夏 (xià) in CCs



Source: MICICL 2109092412X

By contrast, character 幫 (*bāng*) had a total of 9 component errors, distributed among 4 learners, but the error types varied significantly. Student No. 28 (Intermediate) incorrectly wrote the bottom component 巾 (*jīn*) as 王 (*wáng*), which is a component substitution error (Figure 5.32). Student No. 19 (Beginner) incorrectly wrote the left component 𠂇 as 力 (*lì*), which is a component deformation error (Figure 5.33). Different learners erred on different components of the character 幫 (*bāng*), showing that its errors are multi-sourced and dispersed.

Figure 5.32

*Example 1 of character 幫 (*bāng*) in CCs*



Source: MICICL 2801082410X

Figure 5.33

*Example 2 of character 幫 (*bāng*) in CCs*



Source: MICICL 1909092412X

Overall, similar to the stroke level, component-level errors also exhibit the coexistence of clear cross-learner commonality and individual differences. On the one hand, many errors are concentrated on the same location and similar orthographic patterns, indicating common vulnerability points in learners' component processing, which are highly systematic and regular. On the other hand, different learners still show some variation and diversification in the specific error forms; even when facing the same character, they may make errors on different components or at different levels.

STCs and MCs

As can be seen from Table 5.5, the overall distribution of STCs and MCs is quite dispersed, with only a few character types showing relatively high frequency.

Table 5.5*Top 5 STCs and MCs target character types*

STCs			MCs		
Target character type	Frequency	No. of students	Target character type	Frequency	No. of students
或 (<i>huò</i>)	11	4	假 (<i>jià</i>)	4	1
庭 (<i>tíng</i>)	3	1	咱 (<i>zán</i>)	3	1
最 (<i>zuì</i>)	3	1	室 (<i>shì</i>)	3	1
知 (<i>zhī</i>)	3	2	教 (<i>jiào</i>)	3	1
秋 (<i>qiū</i>)	3	1	帮 (<i>bāng</i>)	2	1

Specifically, in the Structural Error category, only 或 (*huò*) appears to be concentrated, with 11 occurrences, making it the most significant type of incorrectly written character in this category. The remaining character types (such as 庭 (*tíng*), 最 (*zuì*), 知 (*zhī*), 秋 (*qiū*), etc.) appeared only sporadically and were distributed among a small number of students, thus not forming a noticeable convergence.

The errors in the writing of 或 (*huò*) were mainly produced by Student No. 7 (Intermediate) and Student No. 4 (Advanced). Both learners repeatedly produced the same type of error across different tasks: they reversed the relative position of the internal 口 (*kǒu*) component and the upper diagonal stroke, resulting in a character form with an inverted vertical structure (see Figure 5.34 and Figure 5.35). This indicates that the error does not stem from a failure to recognize the component itself but rather occurs at the level of processing the spatial relationship between components. That is, the learners showed deviations in the spatial organization of the orthographic structure, disrupting the overall balance and normative structure of the character. This phenomenon can be seen as a typical manifestation of structural bias, suggesting that the learners' internal orthographic system is still unstable when mastering component positioning and structural combination rules.

Figure 5.34

Example 1 of character 或 (huò) in STCs



Source: MICICL 0402032402X

Figure 5.35

Example 2 of character 或 (huò) in STCs



Source: MICICL 0701002305X

As for the MC category, the character type distribution is even more uniform, with no high-frequency clustering like that seen with 或(huò). Most of these incorrectly written instances appeared in the individual writing of a single student, with low frequency and scattered types, failing to form a systematic error pattern. This suggests that, at this level, errors largely exhibit individualized characteristics rather than group-level regularity.

Classification Analysis Summary

As shown in Table 5.6, the distribution of the four types of incorrectly written characters in the corpus shows clear contrasts in their distributional profiles.

Table 5.6

Statistics of four incorrectly written characters type

Category	No. of characters types	Top-5 ratio
SCs	33	59.8%
CCs	192	16.3%
STCs	20	63.9%
MCs	33	32.6%

SCs involved fewer character types but had a high degree of concentration (Top-5 ratio = 59.8%), showing a distinct tendency toward aggregation. CCs, on the other hand, were both

the most frequent (411 instances) and covered the largest number of character types (192), indicating a wide range but a relatively low concentration (Top-5 ratio = 16.3%). However, as discussed in the previous analysis, despite their lower concentration, learners displayed strong cross-learner regularity in component processing, similar to the consistency observed at the stroke level. These two categories therefore represent the most systematic and collectively representative patterns of writing bias among learners.

By contrast, STCs involved the fewest character types, yet the top five alone accounted for nearly two-thirds of this category (Top-5 ratio = 63.9%), indicating that structural deviations also show cross-learner commonality, though they are confined to a few fixed-problem characters. MCs had limited coverage and a moderate concentration (Top-5 ratio = 32.6%), representing more individualized and sporadic errors that emerged from the overlap of multiple processing levels. These two categories therefore lack strong group-wide regularity and instead reflect learners' individual strategies and occasional deviations during character production.

Overall, this distribution pattern reveals the coexistence of systematic regularity and individual variation in learners' Chinese character writing acquisition, offering empirical evidence for the hierarchical design and targeted intervention of pedagogical practices in Chinese orthography instruction.

5.3 Character use: Incorrectly used characters

5.3.1 Research questions

On the surface, incorrectly used characters involve the substitution of a formally correct character in an incorrect context. This phenomenon of a “legitimate but erroneous” character compels us to investigate a critical question: Why do learners substitute the target character A with the incorrect one B? Does this substitution suggest that they primarily rely on phonological similarity during character mastery, or is it a result of orthographic confusion of visually similar forms? Alternatively, does it stem from the instability of semantic connections? Specifically, how do learners mobilize their existing character lexicon within a given writing

context? The occurrence of incorrectly used characters serves as a crucial observational window, enabling us to identify the inherent instability in the learners' integrated connection among form, sound, and meaning, thereby exposing specific difficulties in the acquisition process. Furthermore, by examining the distribution and type of these errors across different learning stages, we can observe whether this phenomenon evolves as the interlanguage system develops.

At the pedagogical level, incorrectly used characters offer instructors direct, intuitive evidence: they highlight which characters are most susceptible to misapplication, often involving obstacles learners encounter when differentiating or retrieving existing knowledge. By systematically categorizing and analyzing these errors, teachers can pinpoint specific problematic areas. Consequently, they can provide more targeted instruction and focused practice on these common confusion points, ultimately enhancing learners' accuracy in the practical writing and application of Chinese characters.

This section, therefore, aims to provide a targeted, supplementary perspective by utilizing the MICICL Corpus to specifically examine the use of incorrectly used character by L1 Italian learners. The research will revolve around the following two questions:

- i. What characteristics and developmental trajectory emerge in the acquisition of the sound-form connection of Chinese characters among L1 Italian learners? I addressed this by examining the overall pattern of incorrectly used characters in learners' writing and comparing the differences in use across three proficiency levels.
- ii. Are learners more susceptible to interference from a character's sound or its graphic form during the acquisition process? Specifically, do incorrectly used characters more frequently reflect phonological similarity or graphic similarity to the intended target? This question is explored through an analysis of the distributional patterns of incorrectly used characters across different categories.

Three methodological clarifications are in order.

- i. In a situation where the learner should have used character A but instead used character B, I designated B as the incorrectly used character and A as the target character.

- ii. When writing, learners' intended target and actual written outcome do not always coincide. In some cases, learners appear to aim at producing the correct target character (A) but, due to insufficient mastery of form or structure, unintentionally produce another existing and legitimate character (B). In such cases, although the written form is a real Chinese character, it does not correspond to the presumed target. Some scholars have treated this mismatch between intended target and written outcome as a distinct category, referring to it as near-incorrectly used characters (cf. Zhang, 2025). However, in a corpus-based study, the learner's actual intention cannot be directly accessed. The identification of a target character therefore necessarily involves an analytical assumption made by the researcher, based on linguistic competence, contextual interpretation, and task design, rather than on explicit evidence of learner intention. As the available data only allow observation of what learners actually produced, rather than what they intended to produce, all legitimate Chinese characters that appear in learner output, can be electronically inputted, but do not match the inferred target (and are consequently tagged as "/B" in the corpus), are uniformly classified as incorrectly used characters and included in the analysis.
- iii. Throughout this chapter, Top-N analyses are used as a descriptive tool to highlight high-frequency patterns. The value of N is adjusted according to the size of the dataset and the analytical granularity of each subsection, and frequency ties at the cutoff point are systematically retained.

5.3.2 Distribution of incorrectly used characters among learners

Initially, by tallying all instances tagged with "/B" (incorrectly used characters) across the 155 handwriting type data samples within MICICL corpus, I identified a total of 437 such characters. This count represents 1.32% of the corpus's total character count of 33031, indicating that in the handwritten data of MICICL corpus, there is likely to be around one incorrectly used character for every 100 characters written.

As I did in Section 5.2, to have a clearer understanding of the incorrectly used character situation among MICICL learners I calculated the incorrectly used character ratio for each

student, using the following formula:

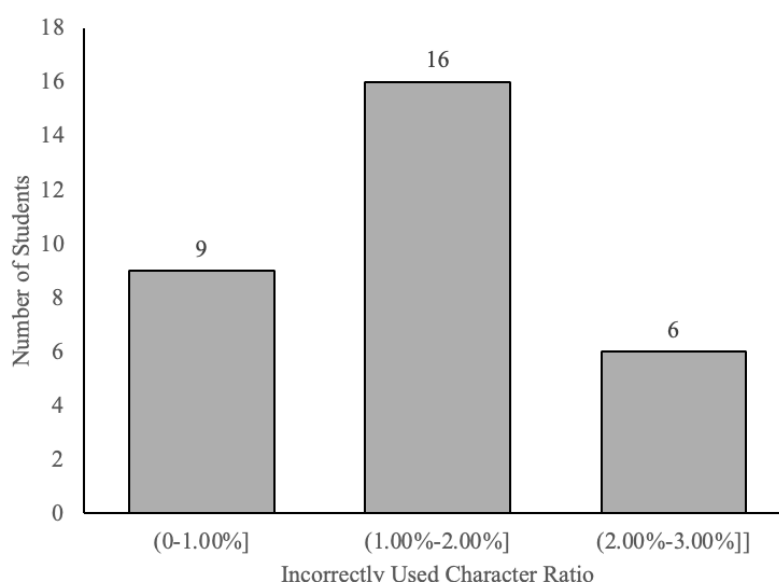
$$\text{Student's Incorrectly used Character Ratio} = \frac{\text{Total Number of Incorrectly - used Characters Produced by the Student}}{\text{Total Number of Character Produced by the student (excluding punctuation)}} \times 100$$

Same as in Section 5.2, all calculation results are rounded to two decimal places. The same formula and rounding conventions are applied to calculate the substitute character ratios for other groupings (e.g., students of different proficiency levels) throughout this section. The incorrectly used character ratio of 39 students is shown in Appendix F.

The counts show that 118/155 writing samples contain incorrectly used characters, with 31 of the 39 participating students producing incorrectly used characters. Among the 8 students who showed an incorrectly used character count of 0: Student No.13 (Beginner) used Chinese characters in only one of their 9 handwritten data submissions, with the other 8 exclusively using Pinyin. Student No.17 (Beginner) submitted only one handwritten sample, which was entirely written in Pinyin. Students No.33, No.35, No.37, and No.38 (all Intermediate) each participated once and demonstrated generally good writing performance.

Figure 5.36

Distribution of students by incorrectly used character ratio



The student with the highest relative ratio of substitute characters was Student No.32 (Intermediate), despite participating in only 2 data collection sessions, totaling a mere 452 handwritten characters; 11 of these were incorrectly used characters, resulting in a relatively

high incorrectly used character ratio of 2.43%. In contrast, Student No.7 (Intermediate) exhibited the highest absolute count of 40 incorrectly used characters, found across 8 handwritten samples that totaled 1,889 characters, giving an incorrectly used characters ratio of 2.12%. Based on the ratio interval, group statistics were calculated as shown in Figure 5.36. Among the 31 students who produced incorrectly used characters, 9 had a ratio below 1.00%, 13 fell within 1.00%–2.00%, and 9 exceeded 2.00%. Although this is only an overview grouping and further analysis should consider learners' proficiency levels, the data indicate that most learners made fewer than two incorrectly used characters per 100 Chinese characters written. The overall rate is not particularly high, yet its consistent presence across learners suggests that such substitution errors, while not dominant, form a stable feature in their writing.

5.3.3 Distribution of incorrectly used characters by character type

Similar to Section 5.2.2, the correction information provided by MICICL was used to identify the number of unique character types that were incorrectly used, with the target character serving as the reference point. The results showed that the 437 instances of incorrectly used characters were distributed across 225 distinct character types, indicating that learners had unstable mastery of these 225 characters and often substituted them with other characters. The detailed list of character types and frequencies is provided in Appendix G.

Subsequently, I examined what specific characters these 225 target characters were replaced with. The analysis revealed that the 437 incorrectly used characters involved a total of 231 distinct substituted forms, meaning that some target characters were replaced by more than one incorrect form. Table 5.7 lists the nine most frequently incorrectly used characters. Up to the ninth rank, characters are clearly differentiated by frequency. From the tenth position onward, multiple characters share the same frequency (three occurrences), forming a frequency tie. Including the tenth-ranked character would therefore require including all characters with the same frequency, substantially expanding the list without adding proportional analytical value. For this reason, the table is restricted to the top nine items.

Table 5.7*Top 9 Target characters most frequently incorrectly used by students*

Character type	No. of students who used the character incorrectly	No. of students who used the character correctly
他(<i>tā</i>)	8	28
采(<i>cǎi</i>)	5	0
描(<i>miáo</i>)	5	1
欢(<i>huān</i>)	5	27
酉(<i>yǒu</i>)	4	0
交(<i>jiāo</i>)	4	9
共(<i>gòng</i>)	4	7
比(<i>bǐ</i>)	4	26
那(<i>nà</i>)	4	24

Note. Due to space constraints, only the most common pronunciation and English gloss are provided for each character.

This table not only identifies which characters were most widely incorrectly used but also compares the distribution between incorrect and correct use of the same character. The numbers of “students who used the character incorrectly” and “students who used the character correctly” are not mutually exclusive, as some learners may have used the same character correctly in certain contexts while misusing it in others. For instance, eight students incorrectly used the character 他 (*tā*), yet 28 students also used it correctly, suggesting that the general understanding of this pronoun remains unstable. In contrast, characters such as 采(*cǎi*) and 酉(*yǒu*), which were used incorrectly but never correctly, indicate that their semantic or collocational functions have not been fully acquired by the learner group.

Overall, the distribution shows that incorrectly used characters are relatively widespread and exhibit a certain degree of cross-learner commonality, indicating that such errors are not isolated incidents but reflect shared difficulties in character use among learners.

Subsequently, I calculated the error rate of incorrectly used characters (Table 5.8). The function of Table 5.8 is to supplement Table 5.7 by shifting the focus from which characters were incorrectly used by more students to how likely each character was to be used incorrectly once it appeared. While Table 5.7 identifies the characters most frequently incorrectly used across learners, Table 5.8 examines the proportion of incorrect occurrences among all uses of each character—essentially, the “density” of errors. As in Section 5.2, only characters with a total occurrence frequency of ten times or more were included in the analysis.

Table 5.8

The 10 character types with the highest frequency of incorrectly used characters

Character type	Total frequency	Incorrectly used times	Ratio	No. of students
采(cǎi)	11	11	100.00%	5
复(fù)	14	7	50.00%	2
佛(fó)	16	8	50.00%	3
汽(qì)	11	5	45.45%	2
共(gòng)	18	5	27.78%	4
担(dān)	22	6	27.27%	3
哪(nǎ)	15	4	26.67%	3
通(tōng)	12	3	25.00%	3
陪(péi)	13	3	23.08%	1
交(jiāo)	26	6	23.08%	4

Note. Due to space constraints, only the most common pronunciation is provided for each character.

As shown in Table 5.8, the character 采(cǎi) reached a 100% error rate, meaning that every instance of its use was incorrect, making it a high-risk character. The characters 复(fù) and 佛(fó) each showed a 50% error rate, indicating highly unstable mastery. Meanwhile, 共(gòng), 担(dān), 哪(nǎ), 通(tōng), and 交(jiāo) exhibited error rates around 25%, which cannot be considered low—one in four occurrences was incorrect. For learners of Chinese, such

proportions suggest incomplete mastery of semantic boundaries, collocational patterns, and distinctions between semantically or graphically similar characters.

Table 5.8 thus highlights the characters occupying an “unstable zone” in learners’ interlanguage systems—those with a high probability of being incorrectly used once produced. A further comparison reveals that the distribution of incorrectly used characters parallels that of the previously discussed incorrectly written characters: both show clear concentration patterns and cross-learner regularities. This indicates that learners experience similar instability in mastering certain characters, whether in orthographic processing or character selection. Such systematic, cross-learner errors reflect the shared areas of difficulty faced by Italian learners in the sample in acquiring Chinese characters.

5.3.4 Category of incorrectly used characters

The classification of incorrectly used characters is essential for gaining deeper insight into the underlying Chinese character acquisition mechanisms of learners. Current research usually categorizes these errors based on their purported source, but there is no unified classification standard.

However, for most foreign language learners, particularly at lower proficiency stages, they may not truly grasp the subtle semantic differences between these Chinese characters. In such cases, using “meaning” as a classification criterion risks introducing subjective inference from the researcher, which may not genuinely reflect the learner’s psychological processing. I found that this type of classification is also prone to intersecting and overlapping with other categories. The intervention of too many reference factors leads to blurred boundaries, which is not conducive to establishing a clear and stable classification framework.

I adopted the four-way classification proposed by Jiang & Liu (2004), which does not involve semantic similarity. In addition, for errors caused by phonological similarity, no further distinction is made between homophonous and near-homophonous characters. Based on the actual distribution of incorrectly used characters in MICICL, the specific classification criteria are as follows:

1. Phonologically similar characters (PSCs). Characters that are not graphically similar but

- share identical or similar pronunciation. Identical pronunciation refers to cases where the initials, finals, and tones are all the same, e.g., 他 (*tā*) – 它 (*tā*), 利 (*lì*) – 力 (*lì*). Similar pronunciation in MICICL refers exclusively to cases where the initials and finals are the same but the tones differ, e.g., 良 (*liáng*) – 亮 (*liàng*), 公 (*gōng*) – 共 (*gòng*).
2. Graphically similar characters (GSCs). Characters that are not phonologically similar but are graphically similar. This includes: Independent characters with similar structures and shapes, e.g., 交 (*jiāo*) – 文 (*wén*), 交 (*jiāo*) – 父 (*fù*), 日 (*rì*) – 目 (*mù*); Independent characters and compound characters formed by adding an additional component to the independent character (or vice versa), e.g., 海 (*hǎi*) – 每 (*měi*), 令 (*lìng*) – 冷 (*lěng*); Compound characters with the same main component but different radicals, e.g., 译 (*yì*) – 择 (*zé*), 环 (*huán*) – 坏 (*huài*); Compound characters with the same radical but similar main components, e.g., 饯 (*jiàn*) – 钱 (*qián*); Compound characters where components are symmetrically transformed in position, e.g., 部 (*bù*) – 陪 (*péi*).
 3. Phono-graphically similar characters (PGCs). Characters that are both phonologically and graphically similar, simultaneously meeting the conditions of the above two types, e.g., 担 (*dān*) – 但 (*dàn*), 他 (*tā*) – 她 (*tā*), 采 (*cǎi*) – 菜 (*cài*), 厉 (*lì*) – 历 (*lì*).
 4. Other characters (OCs). Characters that are neither phonologically nor graphically similar, e.g., 心 (*xīn*) – 兴 (*xīng*), 上 (*shàng*) – 说 (*shuō*), 还 (*hái*) – 那 (*nà*).

5.3.5 Incorrectly used characters by category

According to the classification in section 3.4.1, the statistics of incorrectly used characters in MICICL are showed in table 5.9.

Table 5.9

Classification results of incorrectly used characters in MICICL

Category	PSC	GSC	PGC	OC
Number of characters	70	169	120	78
Ratio in total incorrectly used characters	16.02%	38.67%	27.46%	17.85%

The results show that GSCs were the most frequent, accounting for nearly 38.67% of the total

incorrectly used characters. In contrast, the number of PSCs was less than half that of GSCs, accounting for 16.02%. The prominent proportion of GSCs was consistent with previous research findings (Gao & Meng, 2000; Jiang & Liu, 2004; Wu et al., 2006; Deng & Hu, 2022).

Next, I examined the distribution of types within each error category by taking the target characters as the point of reference (see Figure 5.37). This shows that GSCs not only lead in absolute number but also have the broadest range, encompassing nearly one hundred different target characters. This suggests a widespread difficulty among learners in discriminating between overall character forms. In contrast, PSCs involved the fewest unique types, indicating that this error primarily concentrates on a small set of characters.

Figure 5.37

Character type number of four category

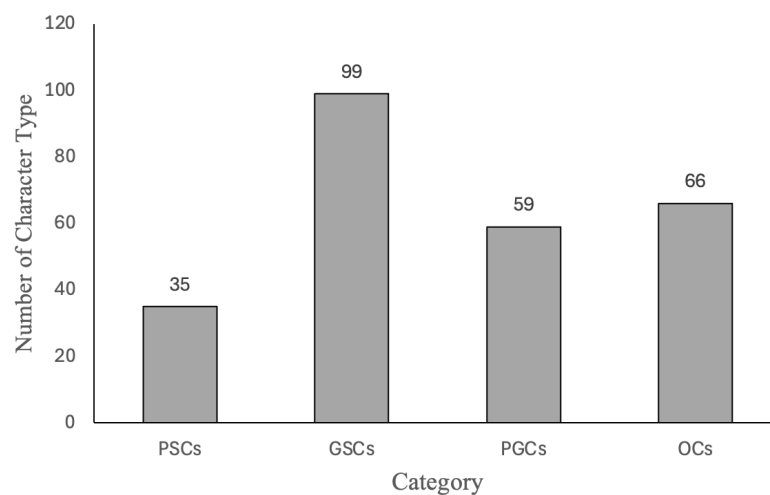


Table 5.10

Statistics of four incorrectly used characters category

Category	PSCs	GSCs	PGCs	OCs
No. of characters	70	169	120	78
No. of character types	35	99	59	66
Top-5 Type Ratio	0.37	0.19	0.25	0.19

To further investigate the distributional characteristics within each error category, I calculated the “Top-5 type ratio”, which measures the proportion of errors that fall into the five most

frequently substituted target characters within that category. This metric assesses whether the error is highly concentrated in a few “high-incidence characters” (calculation results are presented in Table 5.10). The PSCs showed the highest Top-5 type ratio at 0.37, meaning these errors are significantly clustered around a few “high-incidence characters”. In contrast, the GSCs had a Top-5 type ratio of only 0.19. This demonstrates a “broad-coverage, low-concentration” distribution: learners made graphic errors across a large number of distinct Hanzi, rather than focusing the errors on just a few specific characters. PGCs had a Top-5 ratio of 0.25, this suggests that this category involves a relatively high number of target characters while also forming concentrated points of high-frequency errors—a clear reflection of the additive effect of dual phonetic and graphic interference. Finally, OCs shared the low Top-5 ratio of 0.19. This indicates the most uniform distribution, suggesting these errors are more characteristic of individual, sporadic mistakes rather than systemic, high-frequency deviations.

Table 5.11

Top 6 PSCs character types

Target character		Incorrectly used character		
Type	Frequency	Type	Frequency	No. of students
力(<i>lì</i>)	8	利(<i>lì</i>)	8	1
公(<i>gōng</i>)	5	共(<i>gòng</i>)	5	4
园(<i>yuán</i>)	5	院 (<i>yuàn</i>)	5	2
共(<i>gòng</i>)	4	公(<i>gōng</i>)	4	3
它(<i>tā</i>)	4	他(<i>tā</i>)	4	4
		事(<i>shì</i>)	1	1
是(<i>shì</i>)	4	时(<i>shí</i>)	2	2
		失(<i>shī</i>)	1	1

Next, I examined the specific target characters involved in the most frequent PSC types in order to explore the distributional patterns underlying character substitution. However, since the 5th

and 6th most frequent PSC types shared the same frequency, both were included in the analysis to avoid an arbitrary exclusion of equally frequent items. As a result, the analysis for this category is based on the Top-6 PSC types (Table 5.11).

As shown in Table 5.11, the substitution of 力 (*lì*) with 利 (*lì*) was the most frequent, occurring eight times, all produced by a single learner (Student No. 5, Advanced). This strong individual concentration suggests a fossilized phonological–orthographic mapping error that has become internalized within the learner’s interlanguage system. Similarly, the substitution 园 (*yuán*) → 院 (*yuàn*) occurred five times, produced by two learners. This indicates a persistent confusion between characters sharing the same rhyme but differing in semantic scope and word formation function.

By contrast, the reciprocal substitutions 公 (*gōng*) → 共 (*gòng*) and 共 (*gòng*) → 公 (*gōng*) involved a larger number of learners (four and three respectively), revealing a more widely distributed, group-level pattern. Although the two characters differ in form, their phonetic proximity (*gōng/gòng*) causes misselection during writing. In particular, all such errors appeared in the compound word 公共汽车 (*gōng gòng qì chē*, public bus), where the adjacency of two phonologically similar syllables enhanced the likelihood of confusion between holistic word recognition and local character discrimination—an example of cumulative interference driven by both phonological similarity and speech flow continuity.

The substitution 它 (*tā*) → 他 (*tā*), observed in four students, also shows a clear cross-learner tendency. These two characters are fully homophonous but differ in grammatical function and referential domain (inanimate vs. animate). This pattern indicates that learners seem to rely on phonological cues when selecting characters, prioritizing sound-based retrieval over semantic precision. Such cases illustrate a phonological dependency in the learners’ writing strategy, in which the retrieval of a character is primarily sound-driven rather than form- or meaning-based.

In addition, the target character 是 (*shì*) was replaced by multiple phonologically similar characters—事 (*shì*), 时 (*shí*), and 失 (*shī*)—by one or two learners each. This one-to-many substitution pattern reveals a phenomenon of blurred phonological activation: learners can

retrieve the approximate sound but fail to precisely locate the correct orthographic form, leading to variable phonologically motivated substitutions across contexts.

Overall, the distribution of PSC errors exhibits two main patterns: (1) individual-level recurrent deviations such as 力 → 利 and 园 → 院, which reflect fossilized internal phonological mappings, and (2) cross-learner systematic errors such as 公 → 共 and 它 → 他, which highlight the widespread influence of phonological similarity on orthographic selection. These findings suggest that learners rely not only on visual recognition but also heavily on phonological representations when writing Chinese characters.

Table 5.12

Top 5 GSCs character types

Target character		Incorrectly used character		
Type	Frequency	Type	Frequency	No. of students
猫(<i>māo</i>)	8	描(<i>miáo</i>)	8	5
弗(<i>fú</i>)	8	佛(<i>fó</i>)	8	3
夏(<i>xià</i>)	7	复(<i>fù</i>)	7	2
观(<i>guān</i>)	5	欢(<i>huān</i>)	5	5
酒(<i>jiǔ</i>)	5	酉(<i>yǒu</i>)	3	1
		洒(<i>sǎ</i>)	2	2

For GSCs, the most frequent errors (see Table 5.12) involved the target characters 猫 (*māo*) and 弗 (*fú*), each corresponding to a single incorrectly used character type. The substitution 猫 (*māo*) → 描 (*miáo*) occurred eight times, produced by five different learners, which shows clear cross-learner regularity and reflects a group-level confusion between visually similar components such as “犛” and “扌.” In contrast, the substitution 弗 (*fú*) → 佛 (*fó*) also occurred eight times but was produced by only three learners, with the majority of instances coming from one individual (Student No. 09, Advanced), suggesting strong individual concentration rather than group-wide interference.

A similar pattern was found in the substitution 夏 (*xià*) → 复 (*fù*), which occurred seven times across two learners, again indicating individual-level stability errors rather than a general trend. Conversely, 观 (*guān*) → 欢 (*huān*) appeared five times and was made by five different learners, demonstrating a broadly distributed confusion likely driven by visual similarity between the radicals “见” and “欠.”

The character 酒 (*jiǔ*) displayed multiple substitution patterns: three instances of 酒(*jiǔ*) → 酉 (*yǒu*) from one learner and two instances of 酒(*jiǔ*) → 洒 (*sǎ*) from two others. These substitutions show that learners’ errors sometimes extend beyond simple stroke-level confusion, involving semantic association (all three characters relate to liquid or drinking) combined with partial radical overlap.

Taken together, the GSCs reveal two main patterns: 1) group-level misrecognitions driven by common visual similarities among radicals (e.g., 猫 (*māo*) → 描 (*miáo*), 观 (*guān*) → 欢 (*huān*)), and 2) individual fossilized confusions (e.g., 弗 (*fú*) → 佛 (*fó*), 夏 (*xià*) → 复 (*fù*)). These findings indicate that visual proximity between radicals or components creates persistent confusion in orthographic representation, especially when combined with semantic association.

Table 5.13

Top 5 PGCs character types

Target character		Incorrectly used Character		
Type	Frequency	Type	Frequency	No. of students
菜(<i>cài</i>)	11	采(<i>cǎi</i>)	11	5
但(<i>dàn</i>)	6	担(<i>dàn</i>)	6	3
字(<i>zì</i>)	6	子(<i>zǐ</i>)	6	3
她(<i>tā</i>)	6	他(<i>tā</i>)	6	4
气(<i>qì</i>)	5	汽(<i>qì</i>)	5	2

For PGCs (see Table 5.13), which involve both phonological and graphic similarity, the target character 菜 (*cài*) had the highest frequency with 11 errors, all written as 采 (*cǎi*) and

produced by five learners, indicating a clear cross-learner universality. Similarly, 但 (*dàn*) → 担 (*dān*) and 字 (*zì*) → 子 (*zǐ*) each occurred six times and were produced by three learners, showing that the confusion between these phonetic-graphic pairs is widespread rather than isolated.

The substitution 她 (*tā*) → 他 (*tā*) appeared six times and was made by four learners, again confirming the systematicity of this error type. Since the two characters share identical pronunciation but differ in semantic referent and radical (“女” vs. “亻”), this confusion reflects an incomplete integration of form–meaning mapping: learners successfully access the correct phonological form but retrieve the wrong orthographic representation based on radical similarity.

In addition, 气 (*qì*) → 汽 (*qì*) occurred five times, made by two learners, a pattern consistent with reciprocal confusion observed in the opposite direction (气 (*qì*) → 汽 (*qì*)) elsewhere in the corpus. The coexistence of both substitution directions underscores the mutual interference effect of dual phonological and graphical similarity, where two characters become interchangeable for learners.

Overall, PGCs differ from GSCs in that phonological overlap and orthographic resemblance jointly contribute to their co-occurrence as competing forms. While GSCs reflect errors primarily induced by visual similarity, PGCs show how the additive effect of sound–form similarity deepens the instability of learners’ orthographic representations. Together, these two categories illustrate complementary dimensions of interference in character acquisition: visual-perceptual confusion versus phonological–orthographic entanglement, both of which contribute to cross-learner systematicity in Chinese character errors.

In the OCs category the error types were highly scattered. Excluding the most frequent character 或 (*huò*), which appeared 4 times (twice written as 后 (*hòu*) and twice as 这 (*zhè*)), all other target characters occurred only once or twice. Therefore, I limited my further analysis to the case of 或 (*huò*).

Classification Analysis Summary

Across the four categories of incorrectly used characters, distinct yet interrelated error patterns

can be observed. PSCs primarily reflected sound-based confusion, often showing individual concentration but also cross-learner recurrence in cases where phonological proximity blurred the sound–form boundary. GSCs were driven by visual similarity among radicals or components, producing both group-level universality and fossilized individual errors. PGCs (phonographically similar characters) combined phonological and orthographic resemblance, displaying the highest degree of cross-learner regularity and frequent bidirectional substitutions, indicating that dual similarity exerts a strong and systematic influence on character selection. By contrast, OCs appeared highly dispersed and unsystematic, with limited repetition across students, suggesting that they are more likely occasional mistakes rather than systematic interlanguage patterns.

Taken together, the results reveal that except for OCs, the other three categories all exhibit cross-learner commonality, indicating that Chinese character misuse is not random but structured around stable phonological and graphic neighborhoods. I also observed that most of the errors displayed a one-to-one correspondence—a single target character consistently mapped to one main incorrect substitute, regardless of how many learners produced it. This finding indicates that the deviations were not random; instead, they were driven by relatively fixed phonological or graphic neighborhood relationships.

5.4 Character replacement: Pinyin use

5.4.1 Research questions

In the MICICL corpus, a total of 1,737 instances were annotated as Pinyin, indicating that this phenomenon is relatively common among learners. The use of Pinyin is an important feature in learners' written production. It not only reveals their difficulties and strategies in mastering Chinese characters but also provides new insights into their language processing patterns and the relationship between phonology and orthography.

However, few studies have systematically examined Pinyin use in Chinese learner writing. This absence may be related to differences in data collection methods. The MICICL data were collected outside of classroom settings, with participants informed that the collection

was not part of an assessment and would not affect their grades. Students were also told that when they could not recall a particular character or word, they could use Pinyin or another language instead. Consequently, in a relaxed and low-pressure context, learners might have adjusted their writing strategies, resulting in a higher number of Pinyin forms.

This section provides a quantitative and qualitative analysis of the use of Pinyin as a substitution strategy in learners' writing, aiming to explore its distributional features and developmental significance. The analysis revolves around the following two research questions:

- i. What are the distributional characteristics of Pinyin use in learners' writing?

This question examines how Pinyin is distributed across learners and character types, identifying which kinds of Chinese characters are most frequently substituted by Pinyin. It also investigates whether these distributional patterns reflect differences in proficiency levels or cross-learner regularities.

- ii. What characteristics and variations emerge in learners' Pinyin writing behaviors? This question focuses on how learners actually write Pinyin—whether the written forms are complete or partial, accurate or inaccurate, and whether errors occur in spelling or tone marks—thus revealing their developmental trajectory in mastering the sound-form mapping of Chinese.

The following points provide additional clarifications:

- i. When a learner uses Pinyin instead of a Chinese character, the character that ought to have appeared in that context is identified as the target character.
- ii. In the MICICL annotation system, four Pinyin-related tags were designed: /P, /PQ, /PC, and /PZ. Specifically, /P indicates the use of Pinyin; /PQ represents an incorrect Pinyin form (as inferred from contextual clues); /PC refers to Pinyin without tone marks; and /PZ denotes cases where Pinyin and Chinese characters are used together (the character may or may not be correctly written). In practice, however, learners' use of Pinyin was far more complex. In the final transcribed and annotated MICICL corpus, there were 22 different combinations involving Pinyin-related tags and other tag types. For example, /P+W indicates that the learner used both Pinyin and a foreign

word; /PZ+B represents cases where Pinyin co-occurs with a incorrectly used character; and /PZ+C denotes cases where Pinyin co-occurs with an incorrectly written character. All of these cases, including both individual and combined tag occurrences, are included in the statistical analysis presented in this section.

5.4.2 Distribution of Pinyin use among learners

Among the 155 handwritten data samples, 113 contained instances of Pinyin use, with 29 out of 39 participants (74.36%) using Pinyin in their writing. The total number of Pinyin related tags is 1900. This indicates that Pinyin substitution is a relatively common phenomenon among the learners.

However, several extreme cases were identified. Due to limited mastery of Chinese characters, some learners relied almost entirely on Pinyin to complete their writing tasks. For example, Student No.13 (Beginner) participated in nine data collection sessions and used Pinyin in eight of them, except for one instance where a few characters were written in Chinese. Out of 1001 Chinese-character tokens, 953 were Pinyin, accounting for 95.20% of all tokens. Similarly, Student No.17 (Beginner) produced 64 Pinyin tokens out of 65, with only one English word. Since English tokens are not included in the count of Chinese characters in this study, this means that the proportion of Pinyin in this learner's Chinese writing reached 100%.

Because the Pinyin use ratios of these two students are far above the group average, including their data in the overall statistics would significantly distort the results. In particular, when calculating the target characters corresponding to Pinyin, the extensive use of Pinyin in their data leads to an overrepresentation of a small number of highly frequent, basic words, thereby obscuring the distributional patterns observed among other learners with more varied vocabulary use.

As shown in Table 5.14, when Students No.13 and No.17 are included, the target characters corresponding to Pinyin are predominantly concentrated on high-frequency functional or common words such as 我 (*wǒ*, “I”), 很 (*hěn*, “very”), 们 (*mén*, “plural marker”), and 的 (*de*, “possessive particle”). This excessive concentration does not accurately reflect the true distribution of learners' difficulties in Chinese character writing. Therefore, in

the subsequent analyses, the data from Students No.13 and No.17 were excluded to ensure the representativeness and comparability of the quantitative results.

Table 5.14

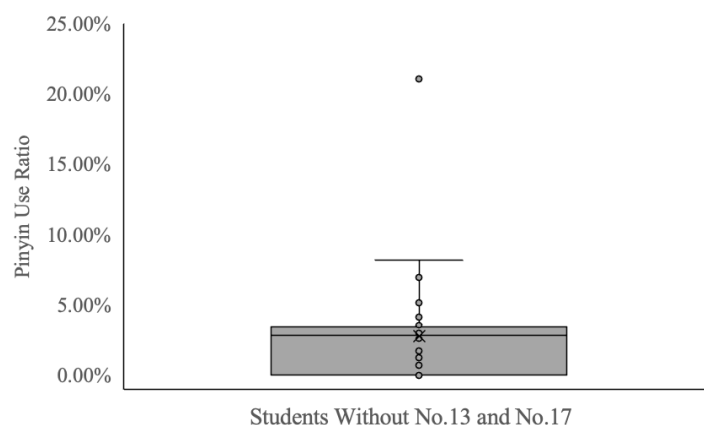
Top 10 target character type of Pinyin (including student No. 13 and No. 17)

Target character type	Frequency
我(wǒ)	105
很(hě)n)	50
们(mén)	32
的(de)	32
你(nǐ)	27
是(shì)	25
以(yǐ)	21
可(kě)	21
个(gè)	20
天(tiān)	20

The total number of Pinyin related tags is 883 after excluding Students No.13 and No.17, taking a ratio of 2.76% in total number of characters (31,966). The proportion of Pinyin use among the remaining 37 students (i.e., the percentage of Pinyin tokens relative to the total number of Chinese-character tokens) is shown in Figure 5.38.

Figure 5.38

Pinyin use ratio by learner (without students No.13 and No.17)



Overall, the distribution appears relatively concentrated, with a median value of approximately 3% and an interquartile range (IQR) between 1.5% and 5%, indicating that most students used Pinyin only occasionally to replace Chinese characters. Pinyin use thus represents a low-frequency yet stable phenomenon within the learner group: although not a dominant pattern, it nonetheless displays a certain degree of cross-learner regularity. The distribution also exhibits a right-skewed pattern, where most students maintained a low proportion of Pinyin use, while a few showed notably higher ratios. The single outlier (approximately 21.09%) represents one student whose Pinyin use was significantly higher than that of the others.

In general, although the overall Pinyin use ratio (around 2%–4%) is not particularly high, its wide presence and patterned distribution make it a meaningful phenomenon worthy of investigation.

5.4.3 Distribution of Pinyin use on different character type

After excluding Students No.13 and No.17, a total of 445 target Chinese characters corresponding to Pinyin substitutions were identified. When examining the target character types corresponding to Pinyin, it becomes evident that learners' use of Pinyin is not entirely idiosyncratic but displays a certain degree of regularity.

Table 5.15

Top 10 Pinyin corresponding Chinese character type

Target character Type	No. of students who wrote the character in Pinyin	No. of students who wrote the character
希(<i>xī</i>)	8	8
望(<i>wàng</i>)	7	8
聊(<i>liáo</i>)	7	5
帮(<i>bāng</i>)	6	11
罗(<i>luó</i>)	6	16
斯(<i>sī</i>)	6	6
园(<i>yuán</i>)	5	12
热(<i>rè</i>)	5	16

择(zé)	5	6
觉(jué)	5	26

As shown in Table 5.15, the Top 10 target characters most widely represented across learners reveal that Pinyin substitution tends to occur on the same set of character types among different students. For instance, 希, 望, 聊, and 帮 were replaced by Pinyin by multiple learners, while many others successfully wrote these characters in their handwritten texts. This contrast between the number of students who used Pinyin and those who produced the correct characters highlights a shared area of difficulty and indicates that certain characters systematically challenge learners at the orthographic production level.

Similarly, Table 5.16 compares the total occurrence frequency of each character with the number of times it was replaced by Pinyin, showing the relative proportion of substitution rather than absolute frequency.

Table 5.16

The 10 target character types with the highest frequency of Pinyin

Target character type	Total frequency	Pinyin-used times	Pinyin ratio	No. of Students
吵(chǎo)	20	10	50.00%	4
架(jià)	19	9	47.37%	3
逛(guàng)	17	8	47.06%	4
铁(tiě)	11	5	45.45%	3
斯(sī)	22	9	40.91%	6
萨(sà)	18	7	38.89%	3
佛(fó)	16	6	37.50%	3
希(xī)	30	11	36.67%	8
聊(liáo)	20	7	35.00%	7
宜(yí)	12	4	33.33%	4

To ensure reliability and avoid noise from rare items, only character types with a total occurrence of 10 or more were included in this analysis. Characters such as 吵(chǎo), 架(jià), 逛(guàng), and 铁(tiě) show high substitution ratios (above 40%), suggesting that even when

these characters appeared relatively frequently in the corpus, learners repeatedly resorted to Pinyin to represent them.

Taken together, these two tables illustrate two complementary perspectives: Table 5.15 emphasizes the cross-learner regularity of Pinyin use—showing that multiple learners substituted the same characters—while Table 16 reflects the frequency-based tendency, showing which characters are most often replaced by Pinyin overall. The convergence of these findings confirms that Pinyin substitution is not a random or individual phenomenon but a patterned and systematic behavior shared across learners.

This suggests that the use of Pinyin in writing reflects a common compensatory strategy among L1 Italian learners when encountering orthographically challenging or less familiar characters. The next section will further examine the qualitative aspects of this phenomenon, focusing on how students write Pinyin—whether complete, partial, or erroneous—to reveal their phonological awareness and the depth of their sound–form mapping in Chinese character acquisition.

5.4.4 Analysis by category

Among the four categories of labels related to pinyin, PZ (using Pinyin and Chinese characters together) and PC (incorrect Pinyin) as these two categories are most informative for describing sound-related patterns in learners’ written use of Pinyin. This section analyzed these two phenomena separately, exploring their writing strategies and their linguistic significance.

In the entire MICICL corpus, there were 1,322 instances of /PQ. After excluding Students No.13 and No.17, who almost entirely used pinyin without tone marks, the number dropped to 404. The large number of /PQ cases in the overall data is mainly attributable to Student No.13’s habitual omission of tone marks. Therefore, /PQ errors likely reflect orthographic habits in pinyin writing rather than actual mispronunciation. To fully assess learners’ mastery of tone and pronunciation, /PQ cases will be further examined in Chapter 6, in conjunction with the reading-aloud data.

PZ

In the MICICL corpus, the tag PZ and its combinations with other tags (e.g., PZ+C, PZ+B)

appeared 161 times, involving 135 different character types, with a Top-5 ratio of 11.11%. This indicates that the phenomenon of using both Pinyin and Chinese characters is relatively dispersed across the learner group. It is not limited to a few individuals, but rather exhibits a certain degree of cross-learner regularity.

Two representative subtypes of PZ use can be identified: 1) Pinyin + incorrectly used characters (PZ+B), and 2) Pinyin + incorrectly written characters (PZ+C).

For the PZ+B type, a total of 23 instances were found. Among these, graphically similar character substitutions (GSCs) and semantically confused substitutions (PGCs) were the most frequent, see Table 5.17.

Table 5.17

Distribution of incorrectly used character types in PZ+B

	PSCs	GSCs	PGCs	OCs
Total number	2	11	7	3
With correct Pinyin	1	6	4	1
Correct Pinyin without tone	0	4	3	2
Correct Pinyin incorrect tone	1	1	0	0

Notably, 13 cases contained entirely correct Pinyin, 8 showed correct initials and finals but omitted tone marks, and only 2 involved tone errors. These cases, in which Pinyin forms are largely accurate despite incorrect character selection, indicate a dissociation between character choice and Pinyin accuracy in learners' written production: even when the target character is not successfully retrieved, the corresponding Pinyin form may still be produced correctly.

Table 5.18

Distribution of incorrectly used character types in PZ+C

	SCs	CCs	STCs	MCs
Total number	3	37	2	9
With correct Pinyin	3	24	0	7
Correct Pinyin without tone	0	12	2	2
Incorrect Pinyin	0	1	0	0

For the PZ+C type, 51 instances were identified, primarily involving component errors (CCs) and multiple-combined errors (MCs), see Table 5.18. In most of these cases, Pinyin was written correctly, with only 16 instances missing tone marks and one case with entirely incorrect Pinyin. This further supports the view that learners rely strongly on phonological cues when writing Chinese characters. When they are unable to recall a character’s exact form, they still tend to recall and write its correct Pinyin representation. Thus, the PZ phenomenon can be regarded as a phonologically driven compensatory strategy—a means by which learners maintain discourse coherence and completeness when the orthographic form is uncertain or inaccessible. This pattern reveals the dominance of the phonological route in the writing process of L1 Italian learners, who, even when processing Chinese characters, tend to activate the phonological representation first before accessing meaning.

In this sense, PZ should not be interpreted merely as a writing “error,” but rather as a functional strategy adopted when learners are unable to retrieve the target character form. The distribution of PZ across proficiency levels indicates that, at the early and intermediate stages of Chinese character learning, pronunciation accuracy is often maintained even when character form retrieval is incomplete, while orthographic control develops more gradually.

PC

Unlike PZ, the PC tag reflects errors that occurred in the production of Pinyin itself. A total of 113 PC instances were identified in the corpus. These can be divided into three major categories: tone errors (65.49%), initial/final errors (24.78%), and multiple errors (9.73%).

Table 5.19

Distribution of different categories of PC

Incorrect tones	Incorrect initials and finals	Multiple errors
74	28	11
65.49%	24.78%	9.73%

Tone errors were the most frequent, indicating that learners found tone marking to be the most challenging aspect of Pinyin writing. Among the 28 cases of initial/final errors, 15 (53.57%) involved confusion between front and back nasals, such as writing “xīn” for *xīng* (兴) or “qián” for *qiáng* (墙). This pattern can be attributed to the absence of similar phonemic contrasts in

the Italian phonological system, demonstrating a clear case of negative transfer from the L1 phonological structure to L2 Chinese sound representation.

Table 5.20

The target tones for tone errors in PC

D1	D2	D3	D4
21	22	13	18
28.38%	29.73%	17.57%	24.32%

Further analysis of tone errors (see Table 5.20) shows that the distribution across the four tones is relatively balanced—28.38% (D1), 29.73% (D2), 17.57% (D3), and 24.32% (D4)—with no strong concentration on any particular tone. This suggests that tone errors occur broadly across all tonal categories rather than clustering around a specific tone. The results imply that learners may rely more on auditory impression or approximate memory rather than on a systematic understanding of tone–phoneme mapping when producing Pinyin.

Overall, PC errors reveal systematic difficulties in learners’ phonological encoding, particularly in tone awareness and the distinction of nasal finals. Nevertheless, it is noteworthy that even in erroneous Pinyin forms, most learners still produced correct initials and finals, indicating that their syllabic awareness is relatively strong and that a basic phonological framework has already been established.

5.5 Foreign words

In the MICICL corpus, a total of 1,184 foreign words were identified, including vocabulary from English and Italian. To ensure the readability and analytical consistency of the corpus, the annotation of foreign words adheres to the following principle:

The annotation tag “/W” is added to the text only when the foreign word has a clear Chinese equivalent translation. If the foreign word does not have a fixed Chinese equivalent in the context or is used purely as a quotation, it is retained in its original form without the tag. Based on this standard, a total of 1,171 “/W” tags were annotated.

The annotated foreign words primarily fall into the following categories:

- Names of Places or Tourist Attractions (93 instances): such as “博洛尼亚 (Bologna),” “弗利 (Forlì),” “佛罗伦萨 (Firenze/Florence),” and “斗兽场 (Colosseo)”.
- Proper Nouns Related to Film or Literature (112 instances): including book titles, movie titles, and names of people or places within them, such as “哈利波特 (Harry Potter),” “霍格沃茨 (Hogwarts),” and “伏地魔 (Voldemort)”.
- Technology-Related Vocabulary (42 instances): such as “科技 (technology)” and “人工智能 (AI)”. Among these, the word “technology” appeared most frequently, totaling 30 instances.

These categories represent the most concentrated types of foreign word occurrences. The remaining foreign words are relatively scattered. Since these words involve lexical level language use analysis, and the focus of this chapter is on the character level writing and error characteristics, they will not be detailed further here. Related lexical level analysis will be explored in subsequent research.

5.6 Summary and discussion of handwritten data analysis

This chapter has examined three major types of character-related deviations observed in the MICICL handwritten data—incorrectly written characters (/C), incorrectly used characters (/B), and Pinyin use (/P, /PC, /PQ, /PZ)—with the aim of uncovering the developmental patterns underlying the orthographic knowledge of L1 Italian learners of Chinese. Table 5.21 presents the overall statistical results of these three error categories.

Table 5.21.

Overall distribution of incorrectly written characters, incorrectly used characters, Pinyin use

	Incorrectly written characters	Incorrectly used characters	Pinyin	Pinyin (excluding outliers)
Absolute frequency	607	437	1900	883
Relative frequency	1.84%	1.32%	5.75%	2.76%
Target character types	245	225	543	445
Samples involved	117	118	113	103
Students involved	32	31	29	27

As shown in Table 5.21, incorrectly written characters and incorrectly used characters display comparable overall frequencies in the handwritten data, whereas Pinyin use clearly stands out as the most frequent phenomenon. This distribution suggests that when learners encounter difficulties in character production, they tend to rely on phonological representations as a compensatory strategy, rather than producing errors confined to orthographic form alone. In terms of character coverage, Pinyin use spans a substantially wider range of target character types, while incorrectly written and incorrectly used characters are concentrated on a more limited set of recurrent problematic forms. This contrast indicates that Pinyin functions as a broadly applicable strategy across the lexicon, whereas the other two deviation types exhibit greater internal convergence.

Although incorrectly written characters account for a relatively small proportion of the total written output, their pedagogical significance should not be underestimated. For incorrectly written character ratios, the majority of students fall within the 1%–3% range, meaning that approximately one to three characters out of every hundred are ill-formed. Quantitatively, this proportion does not constitute the main part of learners' written production; however, once a character form cannot be correctly written, learners are unable to establish a stable sound–form–meaning connection based on that character. Such instability may subsequently affect character recognition and hinder the acquisition of the broader Chinese language system.

A closer examination of internal error structures reveals distinct patterns across different levels of orthographic processing. Incorrectly written characters are dominated by component-level errors, indicating that learners are particularly prone to difficulties in component combination and spatial configuration. Stroke errors occur less frequently but are highly concentrated on a limited set of characters, suggesting that certain stroke patterns pose universal challenges. Structural errors are comparatively rare and tend to cluster around a small number of recurrent forms, while multi-errors, though infrequent, reflect instability in global character construction. Taken together, these patterns point to a lack of automatization in visual character representation and writing procedures, as learners' control over component

positioning and spatial organization has not yet stabilized.

The predominance of component-related errors is closely related to learners' linguistic background and prior writing system experience. Italian orthography belongs to a highly transparent alphabetic system (Baroni, 2011), characterized by a stable one-to-one correspondence between phonemes and letters and a largely linear processing mode. Reading and writing in Italian primarily rely on phoneme blending, reinforcing a strong expectation of sound–form consistency but offering little preparation for processing the non-linear, spatial structure of Chinese characters. In the absence of a transparent sound-to-form mapping, Chinese character reproduction compels learners to rely heavily on visual memory. As a result, L1 Italian learners often adopt holistic memorization strategies, focusing on overall character shape rather than decomposing characters into functional components such as radicals and phonetic elements (Gao, 2000). While holistic strategies may facilitate short-term recognition and reduce immediate learning burden, they hinder the development of stable component awareness, giving rise to persistent errors at the component level.

A different pattern emerges in the distribution of incorrectly used characters. These errors are overwhelmingly dominated by substitutions involving graphically similar characters, far exceeding those involving phonologically similar characters. This tendency aligns with the notion of error persistence in interlanguage development (Selinker, 1972) and suggests unresolved difficulties in distinguishing among phonological, graphic, and semantic representations during character selection. The prominence of graphic similarity interference has also been reported in previous studies on learners from non–Chinese character cultural backgrounds, indicating that visual features play a decisive role in character retrieval during production, while phonological information is comparatively less effectively activated.

In the present data, this form-dominant error pattern appears closely connected to the linguistic background of the L1 Italian learners. The high orthographic transparency of Italian promotes a “sound–form correspondence” strategy, whereby learners may assume that being able to pronounce a word supports its written retrieval. However, because sound–form correspondences in Chinese characters are opaque, attempts to retrieve character forms via

phonology often fail. When this occurs, learners may resort to selecting visually similar characters as substitutes, resulting in a high concentration of graphically similar confusions and a comparatively lower incidence of phonologically similar errors. This interpretation remains provisional and will require further validation through more fine-grained analyses of the writing process, particularly those incorporating visual modality data such as stroke order, component execution, and overall character structure.

Pinyin use displays yet another distinct profile. It is primarily manifested in the use of Pinyin alongside characters and in Pinyin forms with missing or incorrect tone marks. Cases labeled as PZ typically involve largely accurate segmental representations, with only occasional tone-related deviations, suggesting relatively stable phonological representations. In contrast, PC cases exhibit a high incidence of phonological deviations, particularly in tonal realization and segmental structure, with nasal finals emerging as a recurrent source of difficulty. This contrast points to stronger L1 phonological transfer effects in PC than in PZ. Overall, Pinyin use reflects a clear pattern of phonological dominance and sound–form dissociation: when character retrieval fails, learners resort to phonological representations as a compensatory strategy, yet tone production remains the weakest component of their phonological system.

Taken together, the MICICL writing data reveal a characteristic interlanguage orthographic pattern in which the coordination among sound, form, and meaning remains uneven. The coexistence of form-based writing errors, visually driven character misuse, and phonology-based compensation through Pinyin reflects a transitional stage in learners' writing performance. Learners are often able to sustain coherent written output by relying on pronunciation-based strategies, while control over character form remains variable. Overall, this pattern suggests that Chinese character writing by L1 Italian learners does not simply replicate native use, but follows a distinct developmental trajectory shaped by the interaction between alphabetic and logographic writing practices.

These findings provide empirical support for a view of Chinese character acquisition that emphasizes the interaction between visual-structural awareness and phonological processing.

From a pedagogical perspective, character instruction should therefore integrate explicit training in component analysis and spatial structure with sustained phonological support, guiding learners from a phonology-dominant stage toward a more integrated and automatized orthographic system.

Chapter 6 Character Level Reading-aloud data Analysis

6.1 Overview of this chapter

In the previous chapter's analysis of the handwritten data, I examined the orthographic deviations and systematic patterns exhibited by Italian L1 learners in Chinese character production. Building on that foundation, the present chapter shifts its focus to the reading-aloud modality, conducting a systematic investigation of the MICICL reading-aloud data to reveal the types, distributions of learners' phonetic errors in Chinese production. The analysis centers on tonal, segmental (initials and finals), and compound errors, while also incorporating handwritten data to explore, from a cross-modal perspective, the consistency and divergence of the learners' form-sound mappings. Against this background, the chapter proceeds through four main sections.

Section 6.2 focuses on the distribution and types of errors across the five tones (D0–D4), examining the frequency, categories, and concentration of tonal deviations. By analyzing the actual occurrence patterns, substitution relations, and minimal contrast pairs, it identifies which tones are more prone to error, learners' substitution tendencies, and the relative stability of each tone in production. The findings provide empirical insights for tone instruction and reveal key difficulties in tonal perception and control among Italian L1 learners.

Section 6.3 turns to the segmental level, quantifying and comparing the error types found in initials and finals, and further analyzing compound errors involving both segmental and tonal inaccuracies. By examining error ratios, distributions, compositional patterns, and the corresponding target characters, it investigates learners' mastery of Chinese segmental information and discusses possible L1 transfer effects and perception-production asymmetries within the Italian phonological background.

Combining handwritten and reading-aloud data, Section 6.4 explores how learners map orthographic and phonological information across modalities. It compares the pronunciation of incorrectly written characters and a selected subset of incorrectly used characters in the

reading-aloud data, asking: 1) whether the visual load or structural complexity of incorrectly written characters correlates with higher tonal or segmental error rates, and 2) whether graphically similar substitutions show parallel phonetic confusion (e.g., homophonic or near-homophonic misreadings). The analysis aims to reveal the dominant mode of lexical memory—whether learners rely primarily on form or sound—and to evaluate the stability of form–sound integration in their acquisition of Chinese.

Section 6.5 synthesizes the overall results from the reading-aloud modality, integrating the tonal, segmental, and compound analyses as well as the cross-modal comparison. It summarizes the shared patterns, variation, and perceptual–productive mechanisms underlying learners’ phonetic performance and cross-modal relationships. The comparative analysis by proficiency level will be further developed in Chapter 7.

Through a progressive analysis from tones → segments → compound errors → cross-modal mapping, this chapter constructs a comprehensive error profile of the MICICL reading-aloud data, identifying the principal difficulties and developmental patterns in the Chinese phonetic acquisition by L1 Italian learners. By cross-validating form–sound correspondences across modalities, the study provides empirical evidence for the internal consistency of learners’ interlanguage phonological–orthographic system and complements the findings of Chapter 5 on handwriting performance.

6.2 Tone in reading-aloud

6.2.1 Research questions

This study draws on reading-aloud data from the MICICL corpus to conduct an exploratory analysis of tone errors among Italian L1 learners, aiming to reveal their characteristic patterns of tone production and confusion.

Building on the preceding review of tonal difficulties and previous findings in Section 1.3 and Section 1.5, this section focuses on the tonal production patterns of Italian learners and addresses two core research questions:

- i. Which tones are most frequently produced incorrectly by Italian learners of Chinese? In other words, which tonal categories are the most difficult for them to establish as stable perceptual and productive representations within their phonological system? This question concerns the frequency dimension of tone errors and aims to reveal the differences in difficulty among various tones.
- ii. When incorrect tones occur, which tones are most likely to be substituted for the target tones? In other words, do tone errors exhibit certain systematic substitution patterns, indicating fixed directions of confusion between specific tones? This question pertains to the directional dimension of tone errors and helps to analyze the inherent biases within learners' interlanguage phonological system.

The following points provide additional clarification:

- i. The analysis in this study is limited to single-character tones and does not include tone sandhi, tone coarticulation, or intonation phenomena in connected speech, as these belong to higher-level prosodic domains.
- ii. In data processing, D0-D4 represent the tones actually produced by the learners, with the numbers (0–4) indicating the specific tonal category realized in their pronunciation. In contrast, DC0-DC4 are used to mark instances of incorrect tones, where the numerical value following DC denotes the target tone that should have been produced according to standard Chinese pronunciation. The presence of a DC tag therefore indicates that a tone error has occurred: the learner's actual tone corresponds to the transcribed value (D0–D4), while the correct or expected tone is specified by the number following DC. For example, DC1 means that the correct tone for that syllable should be D1, but the learner actually pronounced it as another tone (e.g., D2, D3, or D4). These DC tags were manually added in the annotation layer of the reading-aloud data within the MICICL corpus and serve as the basis for tone error identification and quantitative analysis in this study.

- iii. This study does not involve detailed acoustic analyses such as pitch contour or F0 measurements. Tone classification and error identification are based on categorical distinctions, that is, overall pitch movement patterns (level, rising, falling, or complex), rather than specific numerical values. In other words, the analysis focuses on tonal categories rather than pitch contours in a physical sense. This categorical approach is well suited to describing learners' observable tone use in reading-aloud tasks and allows for consistent comparison of tonal accuracy across proficiency levels.
- iv. In this section, the quantitative analysis of syllable reading includes only the pronunciation of Chinese syllables, excluding any English segments that appeared in the reading-aloud tasks. The count also encompasses Chinese syllables produced during repetitions and self-corrections within the reading-aloud process. It should be noted that the statistical procedures for the reading-aloud and handwritten data are relatively independent: for example, if a learner used Arabic numerals or pinyin forms in writing but pronounced them in Chinese during reading-aloud, the instance was still counted as a Chinese syllable in this analysis. In total, the MICICL corpus contains 35,053 syllables involving Chinese characters in the reading-aloud data.

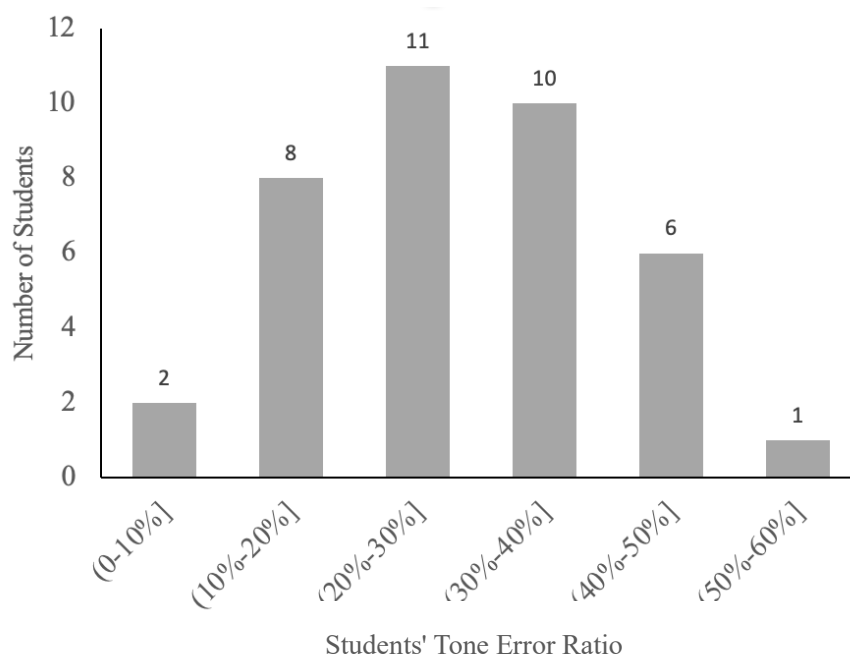
6.2.2 Distribution of incorrect tone types

Among the 153 audio samples collected, 152 contained tone errors, with a total of 9,309 instances distributed across the five tonal categories. These tone errors accounted for 25.27% of the 35,053 Chinese syllables recorded in the MICICL corpus reading-aloud data, indicating that tonal inaccuracy is a widespread phenomenon among the participants. Since Student 16 did not submit any audio recordings, the analysis was conducted on the remaining 38 students. All 38 students produced tone errors to varying degrees, showing that tonal deviation occurred universally within this learner group. Among them, eight students did not produce any DC0 errors (neutral tone errors), while the remaining students exhibited tone errors in all five tonal categories.

The ratio of tone errors to total syllables was calculated for each student, and the resulting distribution is shown in Figure 6.1. The tone error rate of each student was calculated based on the proportion of all tone errors (DC0–DC4) occurring within the total number of Chinese syllables that they read. In other words, the percentage reflects how many tone errors, regardless of type, appeared in each student’s entire set of Chinese syllable readings.

Figure 6.1

Distribution of students’ tone error ratios



The error rates ranged from below 10% to above 50%, indicating considerable individual variability. Two students exhibited a relatively low error rate below 10%, while the majority of learners clustered within the 20%–40% range, with 11 students between 20%–30% and 10 students between 30%–40%. Six students showed higher error rates between 40%–50%, and only one student exceeded 50%.

This distribution suggests that most Italian learners in the MICICL corpus produced tone errors in approximately one quarter to one third of all Chinese syllables they read. Such a substantial proportion of tonal inaccuracies highlights the persistent difficulty of tone acquisition for learners from non-tonal language backgrounds. The relatively wide spread of

error rates also reflects notable inter-learner variability, likely related to differences in proficiency level, tonal awareness, and auditory discrimination ability.

Table 6.1 presents the overall distribution of the five tone error types (DC0–DC4). As shown, among the total 9,309 tone errors identified, D4 occurred most frequently, with 3,810 instances, followed by D3 with 2,366, D2 with 1,503, D1 with 1,425, and D0 with only 205. Overall, the distribution of tone error frequencies appears highly uneven, with D4 and D3 showing notably higher counts than the other categories.

Table 6.1

Total tone error type frequency

Tone error type	DC0	DC1	DC2	DC3	DC4
Frequency	205	1425	1503	2366	3810

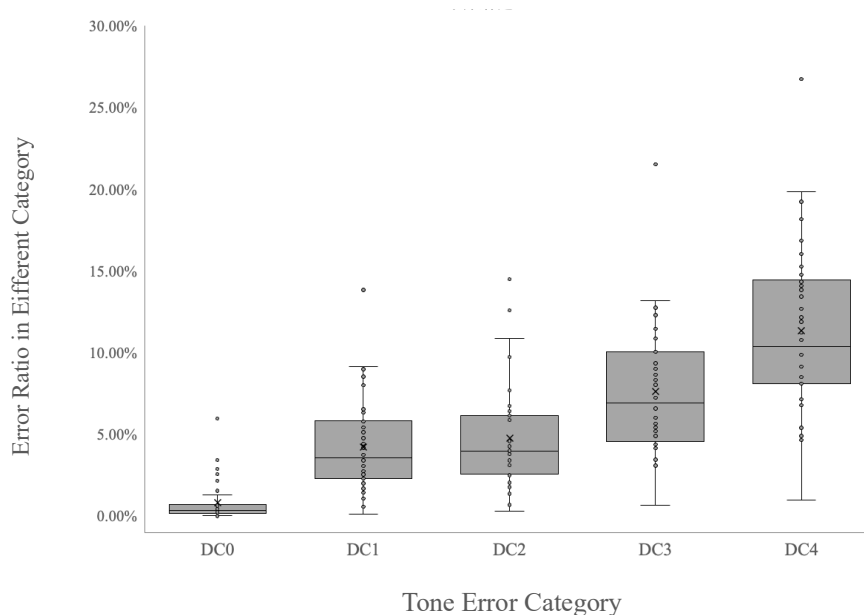
However, it should be noted that such numerical differences do not directly reflect learners' relative difficulty in acquiring each tone, because the natural distribution of tones in Chinese itself is inherently unbalanced—the number of characters for each tone is not evenly distributed. Previous studies (Sun et al., 1996; Liu & Ma, 1986), based on analyses of the National Standard Character Set and lexical databases, indicate that D4 has the highest proportion in Chinese, whereas D0 is extremely rare. Therefore, directly comparing the raw frequencies of tone errors across tones is not meaningful. For this reason, the present study does not conduct a cross-category comparison of tone error proportions. Instead, it focuses on intra-group variation, examining the distributional characteristics and dispersion of each tone error type across individual learners to reveal variability patterns in tone production.

As shown in Figure 6.2, the boxplot illustrates the distribution of the five tone error types (DC0–DC4) across the learner group. The results show clear differences in the degree of dispersion among tones. D4 displays the widest box and the largest number of outliers, indicating the greatest variability among learners. D3 also shows a relatively tall box and multiple outliers, suggesting a higher degree of internal variation. In contrast, D1 and D2 have narrower boxes, implying a more concentrated distribution among learners. D0 exhibits the

smallest box, with data clustered in the lower range and very few outliers. Overall, the five tone error types show distinct patterns of dispersion, suggesting that the distribution of tonal errors across learners is not uniform—some tones display more fluctuation in error proportions, while others remain comparatively consistent.

Figure 6.2

Dispersion of tone error types among learners



6.2.3 Distribution of tone confusion patterns

Based on the overall distribution of tone errors, this section further examines the specific confusion patterns that occurred within the five target tones (D0–D4). In other words, this analysis focuses on the proportion and distribution of instances where a given target tone was misread as another tone. Such statistics reveal the substitutional relationships and systematic confusion patterns among different tonal categories within learners’ interlanguage tone systems.

D0

As shown in Table 6.2, when the target tone was D0, learners most frequently produced D1 (58.54%), followed by D4 (24.87%), D2 (14.15%), and D3 (2.44%). Among these, mispronunciations of the particle *de* (的) accounted for 86 out of the 120 cases where the

neutral tone was read as D1 (approximately 72%), followed by *me* (么) with 12 instances, indicating that tone errors on neutral tones were highly concentrated on high-frequency function words. This pattern suggests that when processing unstressed syllables in speech flow, learners tended to replace the weak neutral tone with a regular full tone. Additionally, 31 out of the 51 cases where D0 was misread as D4 also occurred on *de* (的), showing that certain function words were assigned fixed tonal patterns across different contexts.

Table 6.2

Distribution of tone substitutions when the target tone is D0

Actual tone produced	D1	D2	D3	D4
Frequency (n)	120	29	5	51
Percentage (%)	58.54	14.15	2.44	24.87

D1

As shown in Table 6.3, when the target tone was D1, learners most frequently produced D2 (59.23%), followed by D4 (27.09%) and D3 (13.33%), with almost no substitutions to D0 (0.35%). This distribution indicates a strong confusion between D1 and D2. Since both tones are level tones that differ primarily in pitch height, learners often failed to establish a stable categorical boundary, resulting in frequent substitutions. In addition, about one-quarter of D1 tokens were realized as D4, which may suggest an influence of sentence-final intonation lowering or declination effects on long, high-level tones.

Table 6.3

Distribution of tone substitutions when the target tone is D1

Actual tone produced	D0	D2	D3	D4
Frequency (n)	5	844	190	386
Percentage (%)	0.35%	59.23%	13.33%	27.09%

D2

Table 6.4 shows that when the target tone was D2, the majority of mispronunciations were D1 (62.81%) and D4 (32.60%), with very few substitutions to D3 (3.46%) or D0 (1.13%). This

mirrors the D1 results and confirms a symmetrical confusion between D1 and D2. Because D2 is a rising tone and Italian lacks lexical-level pitch-raising patterns, learners may have tended to approximate D2 as a high-level tone (D1). Furthermore, the one-third of cases where D2 was realized as D4 suggest that learners also struggled to differentiate rising from falling pitch directions.

Table 6.4

Distribution of tone substitutions when the target tone is D2

Actual tone produced	D0	D1	D3	D4
Frequency (n)	17	944	52	490
Percentage (%)	1.13%	62.81%	3.46%	32.60%

D3

As shown in Table 6.5, when the target tone was D3, learners primarily produced D1 (53.97%) and D2 (34.57%), with smaller proportions of D4 (11.33%) and virtually no D0 substitutions. This result indicates a clear trend of flattening for the dipping tone: about 88% of the D3 tokens were simplified into level tones (D1 or D2). Given the complex contour of D3, which involves a fall and rise in pitch, learners often reduced it to a single-level tonal pattern that is easier to produce.

Table 6.5

Distribution of tone substitutions when the target tone is D3

Actual tone produced	D0	D1	D2	D4
Frequency (n)	3	1277	818	268
Percentage (%)	0.13%	53.97%	34.57%	11.33%

D4

Table 6.6 shows that when the target tone was D4, the majority of errors involved substitutions to D1 (42.92%) and D2 (42.62%), while fewer were realized as D3 (13.78%) or D0 (0.68%). This again reflects a strong tendency toward tonal leveling: over 80% of mispronounced D4

tokens were produced as D1 or D2. Considering that D4 is the most frequent tone in Chinese and features a high-to-low pitch contour, this pattern suggests that learners often replaced falling contours with flat or slightly rising ones.

Table 6.6

Distribution of tone substitutions when the target tone is D4

Actual tone produced	D0	D1	D2	D4
Frequency (n)	26	1635	1624	525
Percentage (%)	0.68%	42.92%	42.62%	13.78%

As illustrated in Figure 6.3, the normalized stacked bar chart displays how each target tone (DC0–DC4) was realized as other tones when mispronounced. All tones exhibit a prominent “level-tone concentration” pattern, where the majority of substitution errors fall within D1 and D2. Specifically, D0 was mainly replaced by D1 or D4, D1 and D2 showed strong bidirectional confusion, and D3 and D4 were also most frequently substituted with D1 or D2. Overall, tone errors among Italian learners demonstrate a systematic simplification from contour tones toward level tones, reflecting a general tonal leveling tendency in their Chinese tone production.

Figure 6.3

Distribution of tone substitution patterns

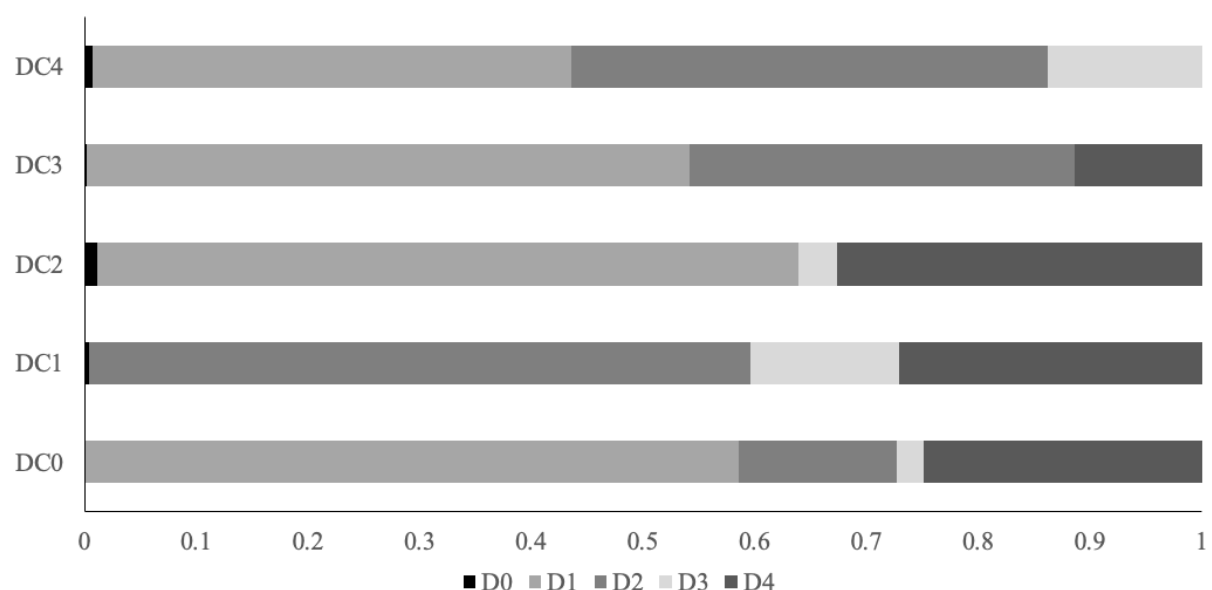


Table 6.7 and Figure 6.4 further illustrate the total number and proportion of correct and incorrect productions for each target tone (D0–D4). Tone errors occurred across all categories but were unevenly distributed. D1 (41.61%) and D2 (37.17%) accounted for nearly three-quarters of all tone errors, followed by D4 (14.79%), D3 (13.28%), and D0 (1.90%). The stacked bar chart reveals a strong concentration of errors in level tones, consistent with the previously identified “level-tone attraction effect,” where learners substitute unstable tones with D1 or D2.

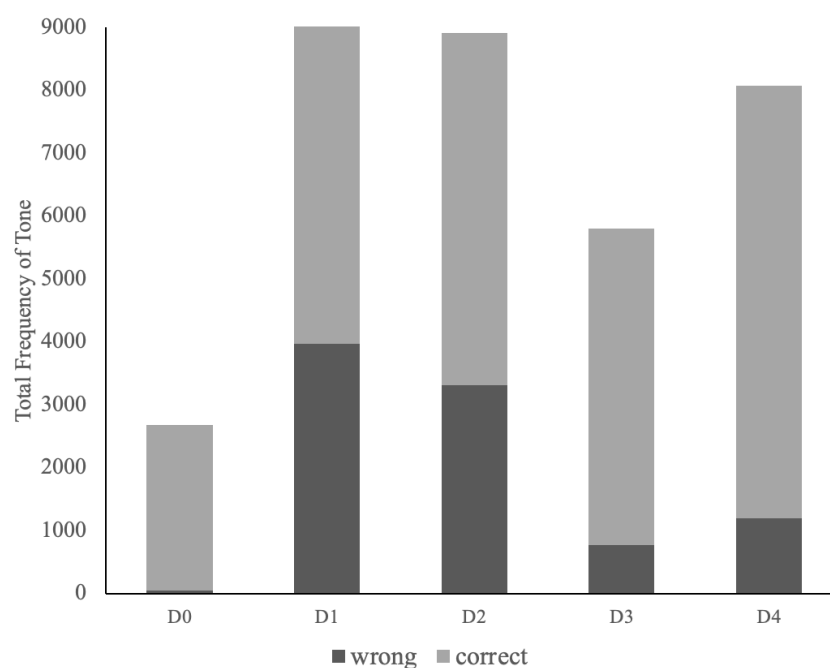
Table 6.7

Overall distribution of correct and incorrect productions for five tones

	DC0	DC1	DC2	DC3	DC4	Total errors	Total number	Ratio
D0	0	5	17	3	26	51	2686	1.90%
D1	120	0	944	1277	1635	3976	9555	41.61%
D2	29	844	0	818	1624	3315	8919	37.17%
D3	5	190	52	0	525	772	5812	13.28%
D4	51	386	490	268	0	1195	8081	14.79%

Figure 6.4

Distribution of correct and incorrect productions across five tones



Notably, this pattern contrasts with the natural tonal distribution in Chinese, where D4 has the highest frequency and D1 only about 25% (Sun et al., 1996; Liu & Ma, 1986). In contrast, in the learners' production, D1 became the most frequently produced tone, indicating a systematic bias toward high-level pitch realization. This "high-level preference" confirms that Italian learners tend to neutralize tonal distinctions by producing most syllables with a high or mid-level pitch, regardless of the intended contour.

6.3 Initials and finals reading

6.3.1 Research questions

This section employs the reading-aloud data from the MICICL corpus to conduct a systematic quantitative analysis and error statistics of initials and finals produced by ItalianL1 learners. The aim is to identify the characteristic regularities and cross-linguistic features underlying their phonetic system. Specifically, this study addresses the following research questions:

- i. Acquisition and error patterns of initials: What are the most frequent types of consonant errors in learners' productions? Which initials are most commonly confused, and do these confusions form systematic patterns?
- ii. Acquisition and error patterns of finals: What are the most frequent error types in vowel production? Are there any shared patterns of deviation among learners?

Three methodological clarifications are in order.

- i. In this section, the analysis and identification of errors in initials and finals are conducted within the framework of the Pinyin system, rather than based on detailed phonetic measurements in the IPA. The Pinyin system and the IPA differ fundamentally in their design principles: the IPA emphasizes fine-grained phonetic distinctions and requires a one-to-one correspondence between symbols and sounds, while Pinyin is a phonemic transcription system in which the relationship between graphemes and phonemes can be one-to-one or one-to-many (Deng, 2020). This means that when certain phonetic distinctions are not contrastive in meaning, a single Pinyin symbol is used to represent multiple phonetic realizations. This "one symbol,

multiple sounds” phenomenon is particularly prominent in Chinese finals—for example, the letter *a* represents /a/, /A/, and /E/; and *i* corresponds to /i/, /ɿ/, and /ʊ/ (Lin & Wang, 2013). Therefore, in this study, all analyses of initials and finals are conducted using the Pinyin system as the operational framework for classification and generalization.

- ii. The analysis in this section is not limited to isolated segmental errors but also includes compound errors involving both segmental and tonal deviations. Specifically, it covers syllables marked as /SC (initial errors), /YC (final errors), and /QC (compound errors combining initial/final with tone errors) in the MICICL corpus, corresponding with Section 6.3.2, Section 6.3.3 and Section 6.3.4.
- iii. In this section, the correct initials and finals that should have been produced are referred to as the target initials and target finals, while the sounds actually produced by the learners are referred to as the incorrect initials and incorrect finals.

6.3.2 Distribution of incorrect initials

Among the 38 students who provided reading-aloud data, 20 produced initial errors, with a total of 73 occurrences, accounting for approximately 0.21% of the 35,053 Chinese syllables in MICICL. Overall, initial errors appeared relatively infrequently among learners and can be considered a minor error type.

As shown in Figure 6.5, the distribution of initial errors across students was highly concentrated: most students exhibited an error ratio below 1%, with only a few exceeding this range. Specifically, 12 students (60%) fell within the (0, 0.5%] interval, 7 students (35%) within (0.5%, 1.0%], and only one student (2.6%) exceeded 1.5%. This indicates that the majority of learners maintained a high level of accuracy in producing Chinese initials, and the overall occurrence of initial errors remained minimal.

At the individual level, the top five students with the highest number of initial errors were as shown in Table 6.8. These results suggest that initial errors were largely concentrated among a few individuals rather than being evenly distributed across the learner group. Overall, Italian L1 learners demonstrated good mastery of Chinese initials, though certain individuals still

showed systematic deviations in the perception or production of specific consonants, warranting further qualitative analysis.

Figure 6.5

Distribution of incorrect initial ratios among learners

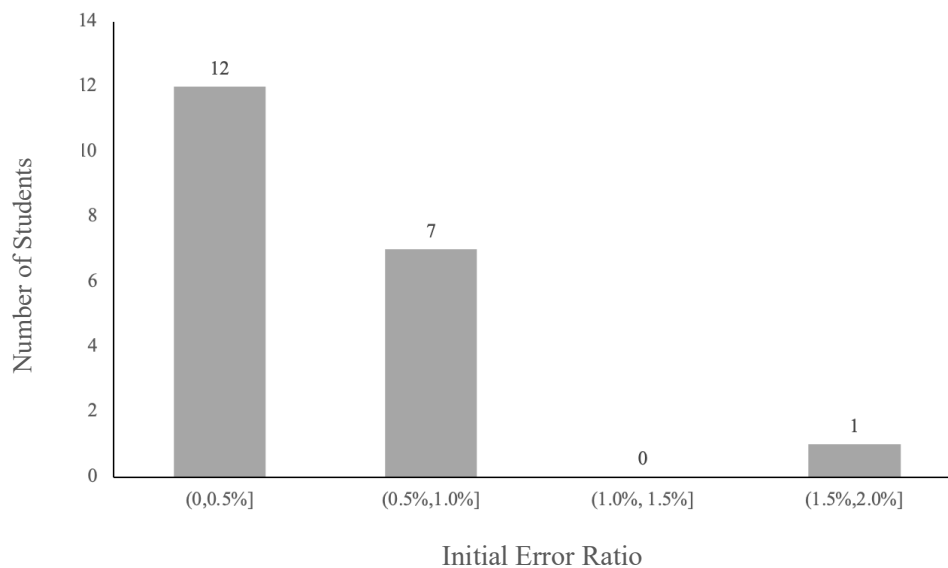


Table 6.8

Top five students with the highest number of incorrect initial

Student code	No. of incorrect initials	Incorrect initial ratio
9	20	0.88%
24	11	0.80%
2	10	0.91%
23	7	0.64%
6	5	0.32%

Of the overall distribution, initial errors were concentrated in a limited number of consonants, showing a clear clustering pattern. As shown in Table 6.9, the six initials with the widest error distribution were q (12 students), ch (6), x (4), b (3), d (3), and zh (3).

Table 6.9*Top 6 target initials with the widest error distribution*

Target initial	Error frequency	No. of students
q	18	12
ch	10	6
x	5	4
b	15	3
d	9	3
zh	3	3

The substitution patterns for these initials, summarized in Table 6.10, reveal several systematic tendencies rather than random errors.

Table 6.10*Substitution patterns of top 6 target initials*

Target initial	Incorrect initial	No. of students
q	ch	8
	x	2
	w	1
ch	c	2
	q	3
	zh	1
	sh	1
	t	1
x	q	1
	sh	2
	zh	1
	s	1
b	p	1

	t	1
	sh	1
d	t	3
	p	1
	-	1
	g	1
zh	j	2

The most striking pattern is the confusion between aspirated and unaspirated pairs, such as q–ch, b–p, and d–t. For instance, q was most frequently pronounced as ch (8 students) or x (2), and d was replaced by t (3 students) or p (1). These substitutions demonstrate that Italian learners often fail to maintain the aspirated–unaspirated distinction that is phonemically contrastive in Chinese. In Italian, plosives and affricates are differentiated primarily by voicing (e.g., /p–b/, /t–d/, /k–g/), not by aspiration. This structural difference leads Italian learners to transfer the voicing contrast from their L1 to Chinese, producing unaspirated stops in contexts where aspiration is required. As a result, Chinese aspirated consonants such as q and ch are often realized with insufficient airflow or shorter voice onset time (VOT), yielding sounds perceptually closer to their unaspirated counterparts.

A second major trend concerns tongue position and place of articulation. The alveolo-palatal sounds (j, q, x) and retroflex sounds (zh, ch, sh) in Chinese are absent from the Italian phonological inventory. Italian consonants are largely articulated in the alveolar region, lacking both palatalized and retroflex categories. Consequently, Italian learners frequently front or flatten their tongue when attempting to produce these Chinese sounds. This can be observed in substitutions such as q → ch/x/w, x → sh/s, and zh → j/g, reflecting a simplification from posterior to anterior and from retroflex to alveolar. The frequent interchange between q and ch—two acoustically similar but articulatorily distinct sounds—further indicates that learners rely on perceptual similarity rather than precise articulatory control when producing unfamiliar phonemes.

Interestingly, some errors (e.g., ch → q or q → ch) involve bidirectional substitution between two aspirated sounds, suggesting that aspiration itself is not entirely absent but inconsistently applied. These patterns reveal that Italian learners may overgeneralize the articulatory features of one aspirated category to another, especially when both involve affrication and high airflow. The confusion between x and sh or zh also indicates perceptual overlap between Chinese's alveolo-palatal and retroflex categories, which Italian learners tend to neutralize into a single alveolar-like articulation similar to /ʃ/ or /tʃ/.

These findings partially align with results from studies on English-speaking learners (Bi, 2001; Shen & Yu, 2009; Huang, 2022; Deng, 2020), which also identified j, q, x and zh, ch, sh as the most difficult initials to acquire. However, the error mechanisms differ. English learners often exhibit voicing of voiceless consonants, a carryover from their NL's voiced–voiceless contrast (Shen & Yu, 2009), whereas Italian learners rarely show voicing errors. Instead, their primary issue lies in insufficient aspiration and unstable tongue positioning, reflecting the absence of an aspiration contrast and the lack of retroflex/palatal categories in Italian phonology.

Overall, Italian learners' initial errors demonstrate both universal and language-specific patterns. Universally, they share the same difficulties as other non-tonal language learners in mastering Chinese's complex palatal and retroflex sounds. Specifically, however, Italian learners' errors are strongly shaped by their L1 phonological structure—most notably, the reliance on voicing rather than aspiration and the confinement of articulation within the alveolar region.

Pedagogically, these findings suggest that pronunciation training for Italian learners should emphasize aspiration control and tongue placement awareness. Techniques such as airflow visualization, spectrogram-based feedback, and articulatory modeling could effectively help learners establish clear phonetic categories for aspirated sounds and palatalized consonants, thereby improving the accuracy and intelligibility of Chinese initial production.

6.3.3 Distribution of incorrect finals

Compared with the acquisition of initials, studies on the acquisition of Chinese finals are

relatively limited. Luo (2007), in a study on Russian-speaking learners, found that the main difficulties in acquiring Chinese finals include insufficient lip rounding in “hekou” and “cuokou” finals (either unrounded or inadequately rounded, e.g., pronouncing $\ddot{u}$ /y/ as /u), difficulty distinguishing between front and back nasal vowels, and unclear articulation of the core vowels in compound finals. Huang (2022), through an analysis of English- and French-speaking learners, further pointed out that errors in monophthongs often involve confusion between a and e, as well as mispronunciations when i is combined with other vowels. For compound finals, learners frequently show incorrect timing in producing ao and iao, and confusion between iu and ui is also common. Regarding nasal finals, errors involving nasalized vowels occur most frequently. Similarly, Bi (2001) reached comparable conclusions based on English-speaking learners, noting that nasal and compound finals pose the greatest challenges in the acquisition of Chinese vowel systems.

Among the 38 students who provided reading-aloud data, 23 made errors in finals, with a total of 44 errors. These errors accounted for only a very small proportion of the 35,053 Chinese syllables in MICICL, indicating that learners generally achieved a high level of accuracy in the production of finals.

Figure 6.6

Distribution of incorrect final ratios among learners

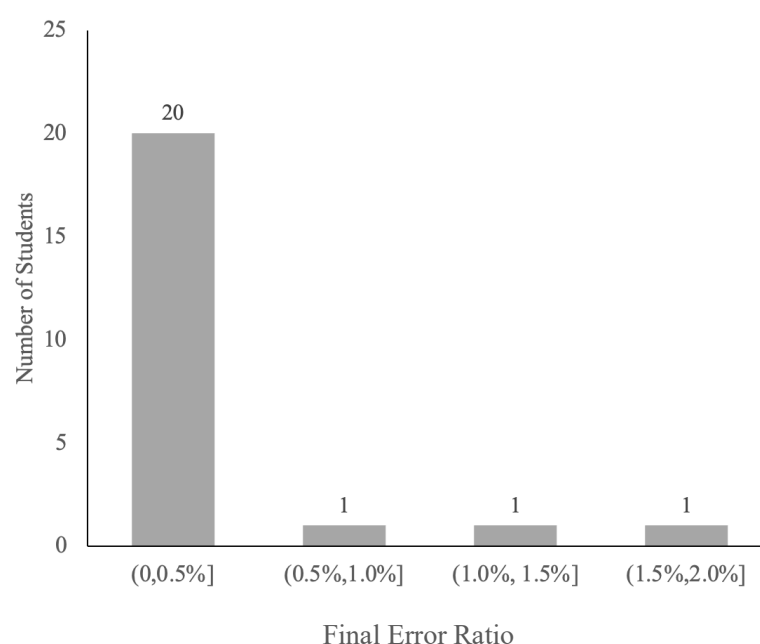


Figure 6.6 shows the distribution of learners' final errors as a proportion of all syllables. As illustrated in the figure, most learners had a final error ratio below 0.5%, with only a few individuals exceeding 1%. Overall, the distribution is highly concentrated in the low-percentage range, suggesting that final errors among Italian learners of Chinese are not widespread but rather occur sporadically on an individual basis.

As shown in Table 6.11, the five students with the highest number of final errors were Student 29 (6 errors, 1.14%), Student 9 (6 errors, 0.26%), Student 2 (4 errors, 0.37%), Student 19 (3 errors, 0.27%), and Student 13 (3 errors, 0.19%). This indicates that although the overall number of errors is small, a few learners still exhibited noticeable instability in their pronunciation of finals.

Table 6.11

Top five students with the highest number of final errors

Student code	No. of incorrect finals	Final error ratio
29	6	1.14%
9	6	0.26%
2	4	0.37%
19	3	0.27%
13	3	0.19%

In terms of overall distribution, errors in finals also show a highly concentrated pattern, with only a few vowel types exhibiting mispronunciations among learners. As shown in Table 6.12, the six finals with the widest error distribution were e (5 students), i (4 students), ang (3 students), eng (3 students), and ian (3 students). Most of these finals involve closely related vowel positions or relatively complex syllable structures, indicating that Italian learners experience difficulties in distinguishing and producing such sounds accurately. As shown in Table 6.13, these final errors are not random but demonstrate systematic patterns.

Table 6.12*Top 6 target finals with the widest error distribution*

Target final	Error frequency	No. of students
/e/	7	5
/i/	4	4
/ang/	3	3
/eng/	3	3
/ian/	4	3
/e/	7	5

Table 6.13*Substitution patterns of top 6 target finals*

Target final	Incorrect final	No. of students
/e/	/i/	4
	/eng/	1
	/ei/	1
/i/	/ie/	2
	/e/	2
/ang/	/an/	2
	/eng/	1
/eng/	/ang/	2
	/ong/	1
/ian/	/an/	1
	/ing/	1

The following observations can be made based on Table 6.13:

- i. The most typical confusion occurred between /e/ and /i/. Four students pronounced /e/ as /i/, and two students pronounced /i/ as /e/ or /ie/. This suggests that learners have

difficulty distinguishing the tongue height and position between mid and front vowels. Although Italian also includes /e/ and /i/, the Italian e is generally produced farther forward than the Chinese /e/, which often leads Italian learners to shift their articulation forward, making the sound resemble /i/. Moreover, Italian lacks a neutralized vowel similar to Chinese /e/, so learners may fail to capture its weakened or transitional role in syllables.

- ii. Among nasalized finals, the most frequent confusion occurred between /ang/–/eng/, /ang/–/an/, and /eng/–/ong/. For example, ang was often pronounced as /an/ (by two students) or /eng/ (by one), while eng was pronounced as /ang/ (by two students) or /ong/ (by one). This finding is consistent with previous studies (Luo, 2007; Huang, 2022; Bi, 2001), which identified nasal finals as one of the most difficult aspects of Chinese pronunciation for foreign learners. Italian has no nasal vowels, and nasal sounds occur only in consonant positions. Because Italian syllable structures are typically simple (C+V), learners tend to weaken or replace nasal resonance with oral sounds when producing Chinese nasal finals, resulting in unclear or missing nasal endings.

Overall, Italian learners' errors in finals mainly concentrate on confusions between closely positioned vowels and instability in nasal finals. This finding partly aligns with research on English and French learners (Huang, 2022; Bi, 2001), confirming that nasal finals remain a common challenge in Chinese phonetic acquisition. However, Italian learners differ in that their errors are more concentrated in the distinction of simple vowels rather than the temporal coordination of compound finals. This may be attributed to the simpler vowel system of Italian, which lacks a hierarchical distribution of tongue positions.

From a pedagogical perspective, pronunciation training for Italian learners should focus on differentiating e and i and stabilizing the production of nasal finals. Targeted instruction such as mouth-shape demonstrations, auditory discrimination exercises, and self-recording comparisons can help learners establish clearer vowel categories at both the perceptual and productive levels, thereby improving the accuracy and naturalness of Chinese pronunciation.

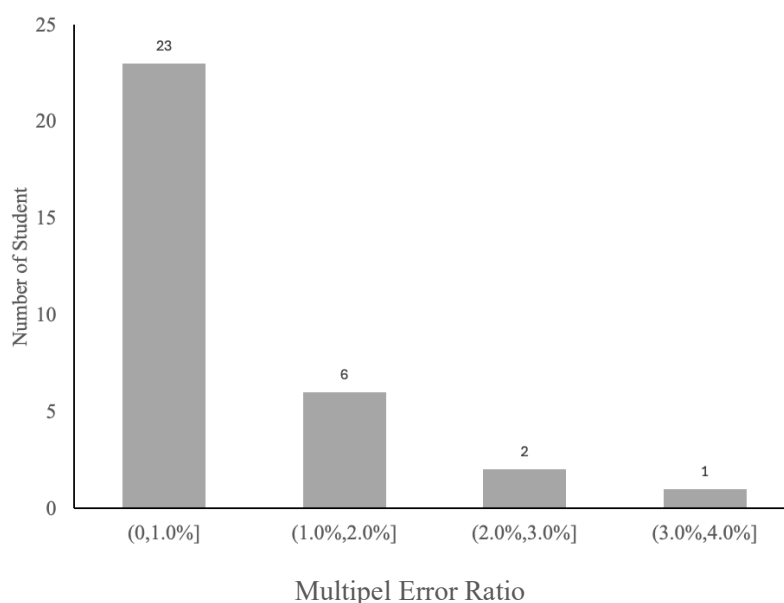
6.3.4 Compound errors involving initials/finals and tones

Among the 38 students who provided audio data, 32 exhibited compound errors involving both initials/finals and tones, with a total of 189 occurrences. These errors accounted for only 0.54% of the 35,053 syllables in the reading-aloud data, indicating that while such compound errors are relatively common across learners, their overall frequency remains low. This suggests that most learners are able to distinguish initials, finals, and tones as separate phonetic dimensions, and that multiple-layered errors occur only in a limited number of cases.

As shown in Figure 6.7, the proportion of compound errors across all syllables is generally concentrated below 1%, with only a few learners exceeding 2%.

Figure 6.7

Distribution of multipel error ratios among learners



Specifically, as presented in Table 6.14, the five students with the highest number of compound errors were Student 19 (34 errors, 3.05%), Student 24 (20 errors, 1.45%), Student 9 (20 errors, 0.88%), Student 2 (15 errors, 1.37%), and Student 29 (12 errors, 2.29%).

Table 6.14

Top five students with the highest number of multiple errors

Student code	No. of multiple errors	Final error ratio
19	34	3.05%

24	20	1.45%
9	20	0.88%
2	15	1.37%
29	12	2.29%

Overall, the distribution of compound errors reveals significant individual variation, suggesting that learners differ considerably in the stability of their phonetic systems. Some students, although performing accurately on isolated initials or finals, still experience coordination difficulties when these are combined with tones during pronunciation.

Compared with isolated initial or final errors, multiple errors—those involving simultaneous deviations in initials/finals and tones—often indicate instability in learners’ overall control of the syllable, rather than localized articulatory mistakes. In other words, such errors reflect syllable-level misproduction rather than isolated segmental deviations. Therefore, this section examines these compound errors from the perspective of their corresponding target characters to explore how they are distributed across different lexical items.

Table 6.15

Top 6 target characters with the widest error distribution

Target character	Error frequency	No. of students
去(<i>qù</i>)	21	8
多(<i>duō</i>)	6	4
边(<i>biān</i>)	4	4
但(<i>dàn</i>)	4	3
用(<i>yòng</i>)	4	3
间(<i>jiān</i>)	4	3

In MICICL, a total of 189 compound errors were identified, involving 103 target characters and corresponding to 100 distinct syllables. The six characters with the widest error distribution can be seen in Table 6.15. Most other character-related errors occurred in only one or two

students, indicating that these compound errors were largely individual-specific rather than cross-learner patterns.

Table 6.16

Substitution patterns of top 4 target characters

Target character	Pronounce multiple errors	No. of student
去(<i>qù</i>)	chi	1
	chi4	1
	chu1	3
	chu2	3
	chu3	2
	chu4	4
多(<i>duō</i>)	tuo2	2
	tuo4	3
边(<i>biān</i>)	bang2	1
	yang2	1
	pian4	1
	guo2	1
但(<i>dàn</i>)	tan2	1
	tan1	1
	dang2	1
用(<i>yòng</i>)	you3	1
	you1	1
	you3	1
	you2	1
间(<i>jiān</i>)	qian2	2
	hou4	1
	wei2	1

A closer examination of the substitution patterns in Table 6.16 reveals three notable tendencies and pedagogical implications:

- i. Many compound errors involve simultaneous shifts in tone and segmental pronunciation. For instance, 去(*qù*) was frequently read as *chi* or *chu*₁–*chu*₄, and 但(*dàn*) as *tan*₁, *tan*₂, or *dang*₂. Such substitutions suggest that some learners have not yet developed a stable tonal category system, and that tonal changes often interfere with the overall structure of the syllable.
- ii. The high-frequency characters listed above show a wide range of substitution types, with error forms varying significantly across learners and lacking shared patterns. This suggests that although learners “recognize” these characters, their phonological representations are still unstable. The phenomenon reflects a typical case of surface familiarity but insufficient phonetic mastery—learners can identify the characters but fail to produce them accurately in speech.
- iii. The character 去(*qù*) exhibited the most concentrated pattern of errors (21 instances), including variants such as *chi*, *chu*, and *chu*₄. Producing *qù* requires both proper aspiration timing and lip rounding with tongue fronting—two articulatory movements absent in Italian phonology. Italian consonant contrasts are based primarily on voicing rather than aspiration, and the language lacks front rounded vowels such as *ü*. As a result, learners tend to retract the tongue position and weaken aspiration, leading to pronunciations like *chu*. This reflects interference at both the articulatory and aerodynamic levels.

In summary, the multiple errors made by Italian learners are neither random nor entirely attributable to L1 transfer; instead, they reveal systematic instability in coordinating tones, syllable structures, and articulatory gestures.

From a pedagogical perspective, addressing such compound errors requires not only separate training for initials, finals, and tones, but also integrated syllable-level pronunciation practice. Techniques such as rhythmic reading, tone–segment coordination drills, and exercises focusing on airflow–tongue synchronization can help learners enhance overall phonetic

stability and achieve a smoother transition from segmental accuracy to syllabic coordination in Chinese pronunciation.

6.4 Incorrectly written/used characters reading-aloud: Cross modal analysis

6.4.1 Research questions

As discussed in Section 1.4.1, Italian and Chinese differ markedly in their degree of orthographic transparency, which has important implications for how learners establish form–sound mappings during language acquisition. However, previous studies have rarely provided empirical evidence on cross-modal consistency during language production processes such as writing and reading-aloud. Most existing findings on the orthography–phonology relationship in Chinese are based on recognition or memory experiments, leaving a gap in understanding how this relationship functions in learners’ spontaneous production behavior. MICICL offers a unique opportunity to explore this issue empirically. In MICICL, the reading-aloud materials were not externally provided, but were instead the texts that learners had just written themselves. Consequently, the pronunciation accuracy observed in their reading-aloud does not reflect the recognition of unfamiliar words, but rather the re-recognition and re-production of their own handwriting output. This allows for a direct examination of how well learners have established orthography–phonology connections in their mental representations of Chinese characters.

Building upon this feature, the present section aims to investigate the cross-modal correspondences between learners’ handwritten and reading-aloud performances, focusing on whether they can accurately pronounce the target syllables when written errors occur. Specifically, it seeks to determine whether learners’ mental representations of Chinese characters are primarily form-based or sound-based, and how this pattern reflects their reliance on orthographic versus phonological processing routes.

6.4.2 Incorrectly written characters reading-aloud

In the MICICL writing data, a total of 607 incorrectly written characters (/C) were identified, among which 215 were mispronounced during reading-aloud, while 392 were pronounced correctly, yielding an overall pronunciation accuracy rate of 64.58%.

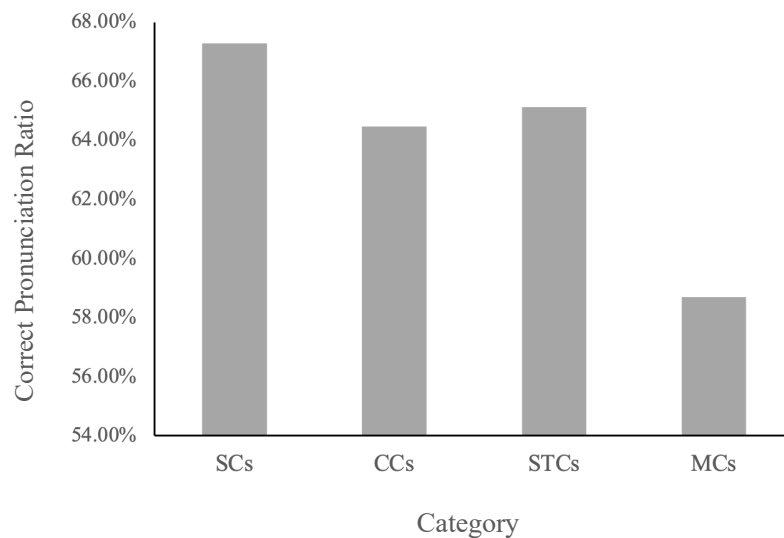
Table 6.17

Pronunciation accuracy of incorrectly written characters across error types

	Total number	Incorrectly pronounced	Correctly pronounced	Correct pronunciation ratio
SCs	107	35	72	67.29%
CCs	411	146	265	64.48%
STCs	43	15	28	65.12%
MCs	46	19	27	58.70%

Figure 6.8

Correct pronunciation ratio of incorrectly written characters by error type



As shown in Table 6.17 and Figure 6.8, characters containing stroke errors (SCs) exhibited the highest pronunciation accuracy: 67.29% of incorrectly written SCs were read correctly, despite deviations in their written form. Pronunciation accuracy was slightly lower for structural and tone-compound errors (STCs, 65.12%) and component errors (CCs, 64.48%), and lowest for

multiple errors (MCs, 58.70%). This difference indicates that the degree of disruption in orthographic form affects how learners retrieve the corresponding phonological representations.

The high accuracy of stroke errors (SCs) suggests that such deviations mainly occur at the motor-execution level of writing rather than reflecting a failure of orthographic knowledge. Learners were still able to access the correct pronunciation despite handwriting inaccuracies, showing that their phonological memory operates relatively independently of surface-level stroke implementation—evidence of a stable orthography–phonology separation.

By contrast, component (CCs) and structural errors (STCs) involve deeper visual disruptions, particularly when phonetic components were incorrectly written or replaced. For native Chinese readers, such alterations would typically impair pronunciation retrieval, as previous studies have shown that native speakers rely more on semantic radicals than on phonetic radicals when recognizing characters (Pine et al., 2003; Shu & Anderson, 1997). In the case of Italian learners, however, even when phonetic components were visually damaged, they were still able to correctly pronounce about 65% of the target syllables. This finding suggests that they tend to rely on a phonological route rather than on semantic or visual cues when recalling and recognizing Chinese characters. This tendency can be interpreted as a transfer effect from their native orthographic experience—Italian, as a highly transparent orthography, promotes a linear sound-to-form processing habit, which these learners continue to apply when acquiring Chinese characters.

The multiple errors (MCs) category, with a pronunciation accuracy of only 58.70%, presents the strongest contrast. Because the learners read aloud their own handwritten texts rather than externally provided materials, these cases do not represent mere recognition failures. Instead, they indicate that the learners had not established a stable orthography–phonology representation. Such instances reflect weaknesses in their form–sound integration and suggest that their interlanguage system remains at an early stage where the orthographic and phonological mappings are not yet firmly consolidated.

To further analyze the nature of mispronunciations, the 215 reading-aloud errors were categorized into tone, initial, final, and combined (multiple) errors. The results are summarized in Table 6.18.

Table 6.18

Proportion of tone, initial, final, and multiple errors in the pronunciation of incorrectly written characters

Tone errors	Multiple errors	Initial errors	Final errors
188	25	2	0
87.44%	11.63%	0.93%	0.00%

The vast majority of pronunciation errors (87.44%) were related to tone, while errors in initials or finals were extremely rare. This indicates that even when the orthographic form was written incorrectly, learners largely retained accurate memory of the syllable’s segmental structure, with errors confined mostly to tonal production.

These findings reinforce the earlier conclusion that Italian learners rely primarily on phonological cues in their memory of Chinese characters, recalling pronunciation through sound–form associations rather than semantic radicals or visual structures. Even when the written form was distorted, they were able to retrieve the correct base pronunciation (initial + final), showing a phonology-dominant processing pattern. In contrast, native Chinese readers tend to rely on semantic radicals and overall visual configuration in character recognition (Pine et al., 2003; Shu & Anderson, 1997), reflecting a form–meaning-dominant route.

This cross-linguistic difference highlights the profound impact of orthographic transfer: while Chinese readers operate primarily through a “form → sound” pathway, Italian learners approach Chinese through a “sound → form” pathway. The MICICL findings thus provide empirical evidence for how orthographic transparency in the native language shapes the phonological reliance of learners in acquiring Chinese characters. Pedagogically, this suggests that instruction for learners with alphabetic L1 backgrounds should leverage their phonological strengths, integrating targeted sound–form mapping exercises to help them build more stable and bidirectional character representations.

6.4.3 *Incorrectly used characters reading-aloud*

In the cross-modal analysis of this study, the category of incorrectly used characters (/B) in the writing modality mainly includes two subtypes: GSCs and OCs. When examining the reanding-aloud performance of these incorrectly used characters, only the graphically similar and other error types were included in the comparison, while PSCs and PGCs were excluded. This decision was based on three methodological and theoretical considerations.

First, from a methodological perspective, PSCs and PGCs often share the same or highly similar pronunciation with their target counterparts—for example, 是(*shì*)/事(*shì*) or 他(*tā*)/她(*tā*). In the reading-aloud data, this makes it impossible to determine whether the learner actually read the target character or the incorrectly used one, as the acoustic output is indistinguishable. Including such cases would therefore introduce ambiguous data and reduce the interpretability of the results. To ensure the reliability and clarity of the statistical analysis, these items were excluded from the cross-modal comparison.

Second, phonologically and graphically similar substitutions differ in their observable characteristics and analytical value. Phonologically similar substitutions typically occur when learners select an incorrect character that shares the same or a similar pronunciation with the target form. In such cases, the pronunciation produced during reading-aloud is often accurate, and additional analysis of reading accuracy provides limited new information, as the deviation lies primarily in character selection rather than pronunciation. By contrast, graphically similar substitutions involve the selection of characters with similar visual form, regardless of phonological overlap. These substitutions are more informative for analysis because they reflect instability in learners' control of character form and the alignment between written form and pronunciation. Examining such cases therefore offers clearer evidence of difficulty in distinguishing visually similar characters during writing and reading-aloud.

Finally, in relation to the goal of the cross-modal analysis, this part of the study aims to examine the consistency and divergence between learners' handwritten and reading-aloud performance. Because PSCs are acoustically identical to their targets, they cannot serve as comparable data points for evaluating sound–form mapping. Graphically similar and other

error types, however, exhibit visible orthographic deviations in writing and may display tonal or segmental variation in reading, thus providing cross-modal evidence of how learners integrate or separate orthographic and phonological information.

In summary, excluding PSCs and PGCs and focusing instead on GSCs and OCs ensures both methodological rigor and interpretive validity. This selective approach aligns with the broader aim of the study—to reveal the mechanisms of sound–form mapping and the degree of cross-modal consistency in the interlanguage system of L1 Italian learners of Chinese.

Table 6.19

Pronunciation accuracy of incorrectly used characters (GSCs and OCs) based on target characters

Category	Total number	No. of correct pronunciations	Correctly reading ratio
GSCs	169	113	66.86%
OCs	78	34	43.59%

As shown in Table 6.19, when the written GSCs and OCs were hypothetically replaced with their corresponding target characters (i.e., assuming the learners wrote the correct forms), the proportion of correct pronunciation during reading-aloud-aloud reached 66.86% for GSCs and 43.59% for OCs. Furthermore, when cases in which the initials and finals were correct but tones were mistaken were also counted as correct, the accuracy rates increased substantially to 92.90% for GSCs and 67.95% for OCs (see Table 6.20).

Table 6.20

Pronunciation accuracy of incorrectly used characters (GSCs and OCs) when tone errors are excluded

	Total number	No. of correct pronunciations	Correctly reading ratio
GSCs	169	157	92.90%
OCs	78	53	67.95%

The results reveal two important features of learners' orthography–phonology connections in Chinese character processing:

First, the high pronunciation accuracy for graphically similar characters (GSCs) indicates that, although learners frequently confused similar-looking characters in writing, they retained a stable memory of the correct pronunciation. In other words, during the writing process, they appear to activate the pronunciation first and then search for a corresponding character shape based on the sound, with the mapping between sound and form being the point of failure. This “sound-before-form” production pattern clearly demonstrates the phonology-centered processing tendency of Italian L1 learners.

Second, although other confusions (OCs) deviate from the target characters in both form and sound, their pronunciation accuracy still exceeded 40%, and rose to nearly 68% when tone errors were disregarded. This finding is highly revealing: even when lacking both phonological and visual cues, learners were still able to recall the correct pronunciation of the intended target characters in a considerable number of cases. This suggests that their phonological representations remained stable in long-term memory, and that they primarily relied on phonological routes rather than visual or semantic cues when retrieving character forms.

Overall, the reading-aloud performance of incorrectly used characters corroborates the findings from the analysis of incorrectly written characters: Italian learners display a generally phonology-oriented pattern in Chinese character processing. Their pronunciation tends to be more accurate than their orthographic recall, indicating that their mental representation of Chinese characters is first established through syllabic phonological encoding, with orthographic information being integrated later. This pattern stands in sharp contrast to that of native Chinese speakers, who rely primarily on semantic radicals and overall visual configuration during character recognition (Pine et al., 2003; Shu & Anderson, 1997). The findings therefore highlight the deep impact of orthographic transfer on cross-linguistic character processing mechanisms and provide pedagogical implications: for learners from alphabetic backgrounds, teaching strategies should leverage their phonological strengths by integrating sound–form mapping exercises that help consolidate a bidirectional “sound → form” connection in character learning.

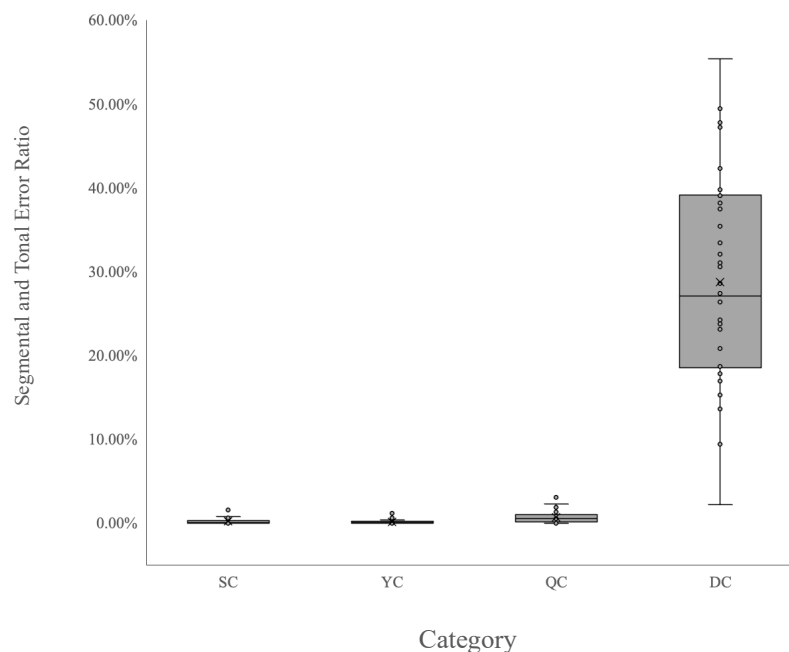
6.5 Summary and discussion of reading-aloud data analysis

An analysis of the MICICL reading-aloud data shows that tone errors far outnumber errors in initials, finals, or compound forms. In the overall distribution, tonal mistakes constitute the overwhelming majority of reading-aloud errors, while segmental deviations account for only a small fraction (see Figure 6.9). This indicates that Italian L1 learners of Chinese are generally able to maintain stable control over the segmental skeleton of the syllable (initial + final), yet still experience substantial difficulty with pitch control and tonal categorization.

The causes of this phenomenon are primarily structural in nature. As Ma (1994) noted, the range of selectable initials and finals in a Chinese syllable is relatively wide, whereas the tonal inventory is limited. As a result, individual tones carry a higher likelihood of deviation than initials or finals. In other words, the tonal system of Chinese constitutes a relatively compact category space, with fine-grained distinctions between tonal categories. This structural property makes tonal contrasts more vulnerable to overlap in actual language use, leading to frequent confusion in pitch height and contour direction.

Figure 6.9

Boxplot of the distribution of segmental and tonal errors (SC, YC, QC, DC)



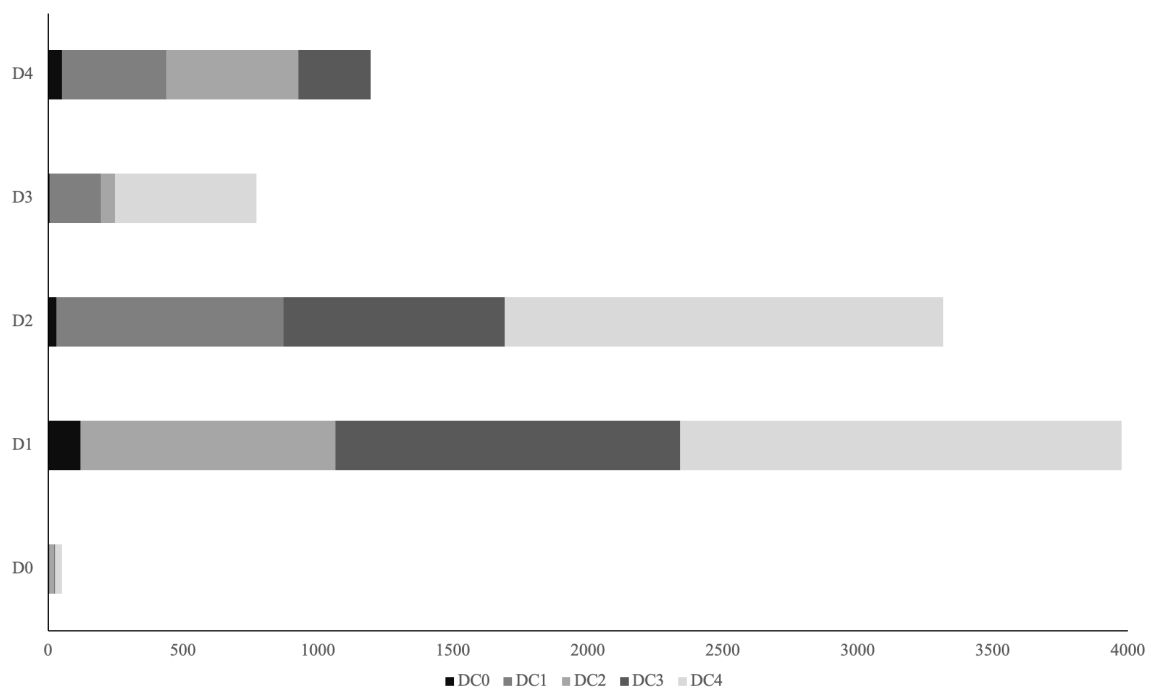
Note. DC indicates tone error in all categories (DC0-DC4).

Further analysis reveals that tone errors are not random but systematic (See Figure 6.10) . Learners exhibit a clear level-tone concentration effect and a tendency toward tonal leveling—the dipping (D3) and falling (D4) tones are often simplified into the high-level (D1) or rising (D2) tones, while D1 and D2 are frequently confused bidirectionally. Even when the overall pronunciation is correct, errors are concentrated at the tonal level, showing that segmental representations are relatively stable whereas tonal information has not yet been fully integrated into the mental representation of the syllable.

Compared with previous studies, these results partly align with but also diverge from earlier findings. While many studies (Mao, 2007; Zhang, 2012; Guo & Tao, 2008) identified D3 as the most difficult tone and D1 as the easiest, Italian learners in this study displayed a distinctive pattern in which the main challenge lies not in producing complex contour tones (D3, D4), but in distinguishing and stabilizing the two level tones (D1, D2).

Figure 6.10

Stacked bar chart of tone error distribution across five tone categories



From a pedagogical perspective, these findings suggest that tone instruction for Italian learners should place greater emphasis on differentiating the two level tones rather than focusing solely

on the more salient contour tones (D3, D4). Targeted perception and imitation exercises could help learners establish clear pitch height and range distinctions between D1 and D2. Additionally, incorporating contextualized reading and contrastive listening tasks may improve learners' sensitivity to subtle pitch differences in level tones, thereby enhancing the overall stability of their tonal system.

From a cross-linguistic perspective, this tendency is closely related to the prosodic system of Italian, a non-tonal language in which pitch serves grammatical or intonational functions rather than lexical contrasts. When producing Chinese, Italian learners often transfer their native intonation patterns to the tonal system, resulting in intonationalization of tone, that is, realizing most syllables with high or mid-level pitch and neutralizing subtle rising–falling distinctions. This preference for high-level tones suggests that their speech processing relies more on a general pitch impression than on precise tonal contours.

This tonal instability is further supported by the cross-modal comparison between handwritten and reading-aloud data. The writing–reading correspondence analysis shows that Italian learners exhibit a predominantly phonology-oriented memory pattern in Chinese character processing.

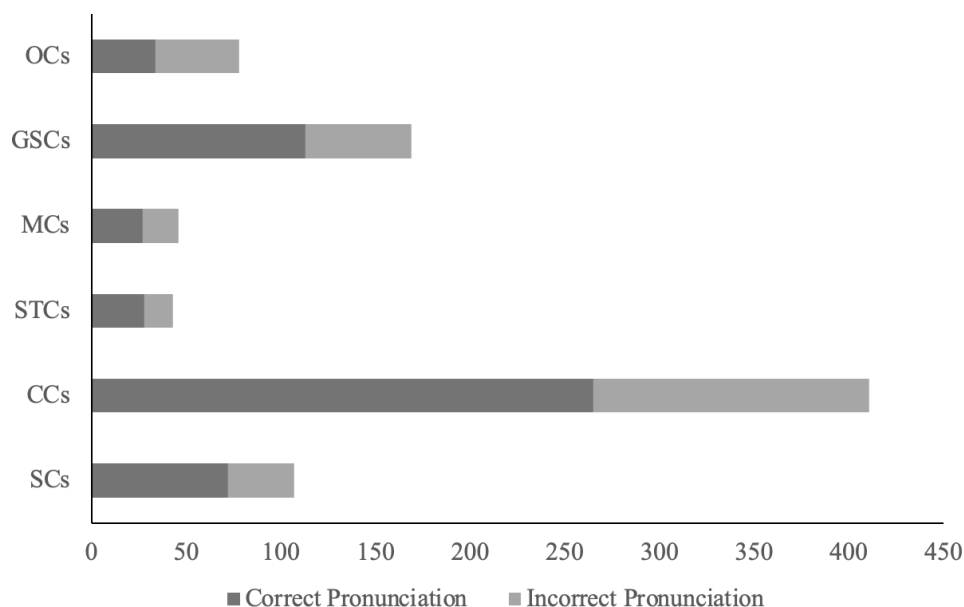
Table 6.21

Pronunciation accuracy of incorrectly written and incorrectly used characters across error types

	Error category	Total number	Correctly pronounced	Correct pronunciation ratio
Incorrectly written characters	SCs	107	72	67.29%
	CCs	411	265	64.48%
	STCs	43	28	65.12%
	MCs	46	27	58.70%
Incorrectly used characters	GSCs	169	113	66.86%
	OCs	78	34	43.59%

Figure 6.11

Comparison of correct and incorrect pronunciations across error types of incorrectly written and incorrectly used (part) characters



As summarized in Table 6.21 and Figure 6.11, both incorrectly written (/C) and incorrectly used (/B) characters display moderately high pronunciation accuracy during reading-aloud-aloud:

- For incorrectly written characters, the overall pronunciation accuracy reached 64.58%. Stroke errors (SCs) showed the highest accuracy (67.29%), indicating that minor execution errors had little effect on phonological recall; component (CCs) and structural–tone-compound (STCs) errors ($\approx 65\%$) show that even when the orthographic structure was distorted, learners could still retrieve the target sound fairly stably; even in multiple errors (MCs), accuracy remained 58.70%, meaning that more than half of the words were still read correctly. Overall, these results suggest that orthographic errors exert only limited interference on pronunciation retrieval, and that phonological representations remain relatively independent and stable in memory.
- For incorrectly used characters, pronunciation accuracy reached 66.86% for graphically similar characters (GSCs) and 43.59% for other confusions (OCs). Even when the written forms differed substantially from the targets both visually and

phonetically, learners could still pronounce the intended syllables correctly in nearly half of the cases, revealing a relatively autonomous and robust phonological retrieval pathway.

Taken together, these results indicate that Italian learners display a phonology-dominant processing orientation in Chinese character acquisition: when retrieving a character, they tend to activate its phonological representation first and then search for the corresponding graphic form. Even when the orthography is inaccurate or partially damaged, the phonological route remains largely functional, suggesting that phonological memory operates independently and robustly. Only when visual information is severely distorted does the sound–form mapping begin to break down. This demonstrates that while learners have begun to differentiate sound and form in their mental representations, the integration between them remains incomplete—reflecting an interlanguage stage in the transition toward a stable orthography–phonology system.

Overall, the instability observed in tone control and the predominantly one-way pattern of sound–form mapping represent two closely related features of Italian learners’ Chinese language performance. Segmental accuracy is generally stable, whereas tones and character forms show weaker alignment with pronunciation. This pattern is closely related to the characteristics of learners’ native writing system. Italian, as a highly transparent alphabetic language, is organized around consistent sound-to-letter correspondences, which encourages a sound-driven approach to written production. When this strategy is extended to Chinese, learners tend to rely on pronunciation to support writing and reading, while tonal contrasts and character form distinctions receive comparatively less attention. As a result, tone accuracy and orthographic stability improve more slowly and remain less consistently aligned with phonological output.

Pedagogical implications emerge accordingly:

- i. Leverage learners’ phonological strengths by implementing sound–form mapping activities (e.g., tone-character association, visualized syllable reading) to build bidirectional connections;

- ii. In tone instruction, shift focus from contour imitation to pitch-height discrimination and category-boundary reinforcement, particularly between D1–D2 and D2–D3;
- iii. Adopt integrated syllable-level training, combining initials, finals, and tones as unified practice units to facilitate transition from segmental accuracy to holistic coordination;
- iv. Incorporate multimodal approaches—acoustic feedback, rhythmic reading, and tone-curve visualization—to help learners phonologize pitch contrasts, transforming them from sentence-level intonation to stable lexical tone categories.

In conclusion, the MICICL cross-modal findings demonstrate that Italian L1 learners follow a distinct “sound-before-form, segment-before-tone” developmental pathway in Chinese acquisition. Even when orthographic forms are impaired, they can still correctly pronounce most syllables, but their control of tones and complex structures remains fragile. This phonology-centered pattern reveals the deep influence of native orthographic transfer on Chinese learning and offers valuable guidance for pedagogy: by integrating sound–form association, perceptual feedback, and multimodal training, teachers can help learners establish a more stable, bidirectional, and hierarchically organized phonological–orthographic system for Chinese characters.

Chapter 7. Comparison Across Proficiency Levels

7.1 Chapter overview

Building upon the analyses of handwritten and reading-aloud modalities in the previous chapters, this chapter adopts a comparison across proficiency levels to examine how Italian L1 learners develop their orthographic and phonological competence in Chinese character acquisition. By comparing beginner, intermediate, and advanced learners, it explores how their performance in writing, reading, and cross-modal processing evolves with increased exposure and linguistic experience.

The classification of proficiency levels (beginner, intermediate, and advanced) in this chapter follows the criteria established in Section 3.5. Against this background, the chapter proceeds through four main sections:

Section 7.2 focuses on the writing modality, comparing the distributions of three major error types—incorrectly written characters (/C), incorrectly used characters (/B), and pinyin use (/P)—across proficiency levels. It examines how the frequency, variability, and concentration of each error type change with learners' progress, revealing the developmental shift from high variability and phonological dependence at the beginner stage to greater orthographic stability and structural awareness at advanced levels. The analysis aims to illustrate how Italian learners' Chinese writing develops from phonology-based substitution toward visual–orthographic mastery.

Section 7.3 turns to the reading-aloud modality, investigating the developmental differences in tone, segmental (initial/final), and compound (segmental + tonal) errors among the three proficiency groups. It quantifies error ratios, dispersion, and within-group distributions to uncover how learners' phonological system evolves from partial instability to system-level integration. The section further highlights stage-specific difficulties, such as the mid-level fluctuation of tone 3 control and the transient rise of multi-errors, providing insight into the non-linear progression of Chinese phonetic acquisition.

Section 7.4 integrates the writing and reading-aloud data to conduct a cross-modal proficiency comparison. Focusing on incorrectly written and incorrectly used characters, it examines whether learners at different proficiency levels can correctly pronounce the characters they previously miswrote or incorrectly used. By comparing cross-modal consistency rates, the analysis explores the interaction between orthographic and phonological representations, revealing how learners' form–sound mapping transforms across stages—from phonological reliance in early learning to orthographic dominance and partial decoupling at advanced levels.

Section 7.5 summarizes the overall findings of the cross-proficiency comparison, integrating results from the writing, reading, and cross-modal analyses. It outlines the shared and stage-specific developmental tendencies observed in Italian learners' Chinese character acquisition, highlighting the dynamic balance between accuracy, fluency, and cross-modal integration. This synthesis not only clarifies the developmental trajectory of orthographic and phonological learning but also provides pedagogical implications for stage-based and multimodal instruction in Chinese as a foreign language.

Through a stepwise analysis from handwriting → reading-aloud → cross-modal integration, this chapter constructs a comprehensive developmental profile of L1 Italian learners' Chinese character acquisition. By tracing both intra-modal progress and inter-modal connections, it demonstrates that Chinese character learning is a dynamic, non-linear process of progressive stabilization and system restructuring, in which learners gradually transform from phonology-dependent beginners into orthographically competent and integrated users of written Chinese.

7.2 Writing data comparison

7.2.1 Research questions

Same as Chapter 5, the analysis about writing data in this section focuses on three major questions designed to trace the developmental trajectory of Chinese character writing among Italian L1 learners and to reveal stage-specific patterns in their orthographic acquisition:

- i. How are incorrectly written characters distributed across learners of different proficiency levels? This question examines both the number of learners who produced incorrectly written characters and the overall frequency and variability of such errors across the three groups. It seeks to determine how the proportion and internal dispersion of incorrectly written characters change as proficiency increases, and how different error subtypes (e.g., stroke-, component-, or structure-level errors) evolve throughout the learning process.
- ii. How are incorrectly used characters distributed across proficiency levels? This question explores the developmental differences in learners' misuse of characters, focusing on the frequency, type composition (e.g., graphically similar, phonologically similar, or semantically confused substitutions), and group-level variability. The goal is to identify whether the incidence of character misuse decreases linearly with proficiency or shows non-linear patterns, reflecting learners' shifting reliance from phonological to semantic and orthographic cues.
- iii. How does Pinyin use vary among learners at different proficiency levels, and what does this reveal about their writing strategies? This question investigates both the proportion of learners who used Pinyin and the ratio of pinyin tokens among total written output. It aims to uncover how the function of Pinyin evolves—from a compensatory phonological tool at the beginner stage, to a transitional scaffold at the intermediate stage, and finally to an occasional strategic aid at the advanced stage—thereby revealing the functional transformation of learners' writing behavior as their orthographic competence develops.

Together, these three research questions address the developmental mechanisms underlying Italian learners' Chinese writing, highlighting how their orthographic performance shifts from phonological dependence to visual-orthographic control, and from variability to stability, as proficiency increases.

7.2.2 Incorrectly written character comparison

In terms of learner distribution, I examined the share of students who produced at least one

incorrectly written character in each group. As shown in Table 7.1, it indicates that the prevalence of incorrectly written characters remains high across levels, with a slight uptick from Beginner to Intermediate.

Table 7.1

Proportion of learners with incorrectly written characters across proficiency levels

Proficiency level	No. of learners with errors	Total number of learners	Percentage of learners with errors
Beginner	5	6	83.33%
Intermediate	18	21	85.71%
Advanced	9	12	75.00%

Intermediate learners are less likely to avoid characters via pinyin substitution and are exposed to more task vocabulary, which increases the chance of making at least one writing error. The prevalence then drops at the Advanced level to 75.00%, suggesting greater stability of orthographic knowledge even though errors have not been eliminated. The total numbers of incorrectly written characters and overall written characters were calculated separately for each proficiency group, as summarized in Table 7.2.

Table 7.2

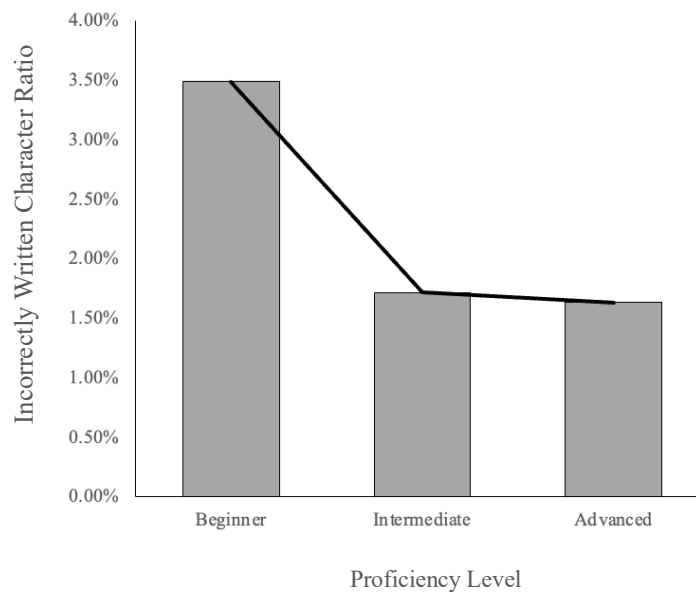
Incorrectly written character ratios across proficiency levels

Proficiency level	No. of incorrectly written characters	No. of total characters	Incorrectly written character ratio
Beginner	99	2833	3.49%
Intermediate	296	17238	1.72%
Advanced	212	12960	1.64%

The resulting trend in error proportions across the three groups is visualized in Figure 7.1, showing the developmental trajectory of learners' writing accuracy as proficiency increases. The error ratio across the different proficiency groups showed an overall trend of decreasing progressively with the increase in proficiency. The Beginner group had the highest proportion of errors (3.49%), the Intermediate group dropped significantly to 1.72%, and the Advanced group further decreased slightly to 1.64%.

Figure 7.1

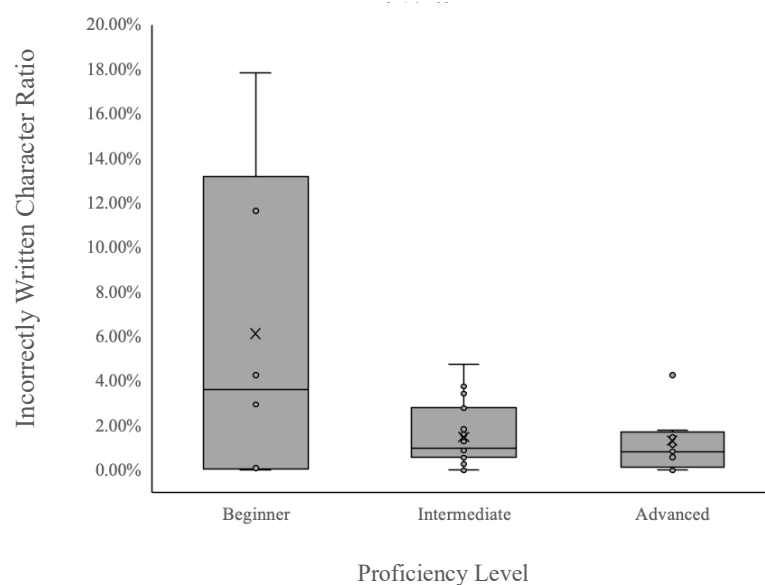
Trend of incorrectly written character ratios across proficiency levels



This result indicated that as learners' Chinese proficiency improved, their Chinese character writing accuracy increased overall, and the incidence of writing errors decreased accordingly. At the same time, this change presented a “fast-to-slow” decline pattern: the reduction in the error proportion was substantial from the Beginner to the Intermediate stage, but the rate of decrease slowed from the Intermediate to the Advanced stage, suggesting that learners' writing performance gradually stabilized after reaching the mid-to-high proficiency levels.

Figure 7.2

Boxplot of incorrectly written character ratios across proficiency levels



Then as shown in Figure 7.2, I observed that the distribution of incorrectly written characters varied significantly across the groups:

The Beginner group exhibited a median incorrectly written character ratio of 3.61% and an average of 6.47%. The interquartile range (IQR) reached 10.05%, with an overall distribution spanning 0.00%–17.86%, indicating a wide degree of spread. This result showed significant internal disparity among beginner learners: some students had virtually no incorrectly written characters (0.00%), while others showed a very high proportion of errors. Notably, these zero values did not necessarily represent writing accuracy, as they were often associated with learners substituting Chinese characters with Pinyin during the writing task, thereby masking potential writing errors at the statistical level. Overall, the learners in this group had not yet formed stable character knowledge, and their control over strokes, components, and structure was still in the establishment phase, resulting in highly volatile writing performance.

The Intermediate group's ratio of incorrectly written characters was more concentrated, with a median of 1.30%, an average of 1.72%, and an IQR of 2.14%. Most students' error proportions fell between 0.00% and 4.75%. Compared to the Beginner group, this indicated that intermediate learners had achieved stable mastery of common character forms, and the overall number of errors had dropped significantly. The smaller dispersion suggested that the writing knowledge structure at this stage was becoming more systematic. However, a few individuals still had error proportions exceeding 4.00%, which I concluded might be related to the complexity of specific characters or the difficulty of the task vocabulary, rather than a universal writing problem.

The Advanced group's ratio of incorrectly written characters was the lowest overall and the most concentrated, with a median of 0.81%, an average of 1.27%, an IQR of 1.13%, and a range between 0.00% and 4.28%. The box plot was very compact, with only a few outliers. This suggested that the writing performance of advanced learners had largely stabilized, and errors were primarily confined to individual instances (such as complex characters), rather than systematic issues.

In summary, the ratio of incorrectly written characters across the three proficiency groups showed a clear trend of gradual decrease as language ability improved. The average error proportion dropped from 6.47% in the Beginner group to 1.72% in the Intermediate group, and finally to 1.27% in the Advanced group. Simultaneously, the internal group variability (dispersion) narrowed significantly. This change reflected the progressive consolidation of learners' Chinese character orthographic knowledge and the stabilization of their writing performance. The Beginner stage was characterized by high volatility and strategy dependence (e.g., using Pinyin or over-generalization), the Intermediate stage represented a transition toward structural refinement and automation, and the Advanced stage stabilized, with errors being primarily sporadic and limited to particularly complex characters.

I then analyzed the distribution and developmental trends of different categories of incorrectly written characters across the three proficiency groups, the result can be found in Table 7.3.

Table 7.3

Number of four categories in different proficiency levels

Proficiency level	No. of SCs	No. of CCs	No. of STCs	No. of MCs	Total number
Beginner	20	58	10	11	99
Intermediate	68	193	15	20	296
Advanced	19	160	23	10	212

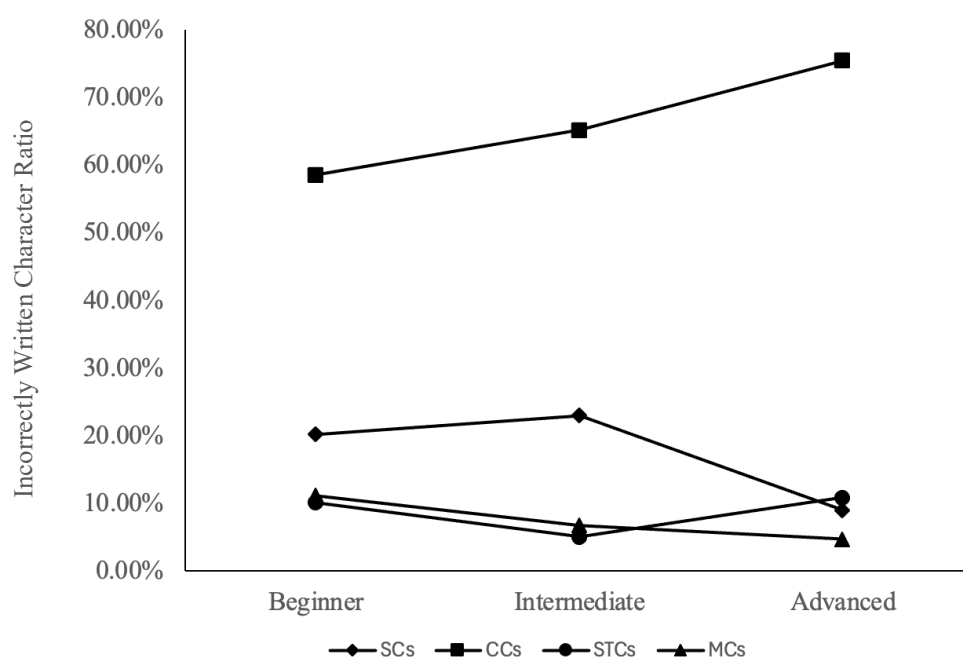
Figure 7.3 shows the categorical ratios of incorrectly written characters across proficiency levels, visualizing how stroke-level (SC), component-level (CC), structural (STC), and multiple (MC) errors evolved as learners' proficiency increased.

The detailed proportions are as follows: in the Beginner group, SCs accounted for 20.20%, CCs for 58.59%, STCs for 10.10%, and MCs for 11.11%; in the Intermediate group, SCs rose slightly to 22.97%, CCs increased to 65.20%, STCs decreased to 5.07%, and MCs remained relatively stable at 6.76%; in the Advanced group, SCs dropped to 8.96%, CCs continued to rise to 75.47%, STCs rebounded to 10.85%, and MCs decreased further to 4.72%. This distribution indicates that the development of Chinese character writing among L1 Italian

learners follows a hierarchical and non-linear pattern, with different error types showing distinct trajectories across proficiency levels.

Figure 7.3

Categorical ratio of incorrectly written characters in different proficiency level



Stroke-level errors (SCs) display a “high at Beginner → slight rise at Intermediate → sharp drop at Advanced” pattern. The temporary increase at the intermediate stage coincides with learners’ engagement with a larger number of characters and greater structural complexity, while the subsequent decline at the advanced level suggests increased stability in stroke execution in routine writing tasks.

Component-level errors (CCs) show a steady and pronounced increase across proficiency levels, rising from 58.59% to 75.47%. This trend indicates that as learners progress, an increasing proportion of their writing deviations involve component selection, combination, or configuration, rather than isolated stroke inaccuracies. Component-related difficulties therefore remain a persistent feature of learner writing, even at higher proficiency levels.

Structural errors (STCs) follow a U-shaped distribution, decreasing at the intermediate stage and rebounding at the advanced stage. This pattern is associated with learners’ expanding character repertoire and increased exposure to structurally complex or low-frequency characters, for which global spatial organization poses renewed challenges.

Multiple errors (MCs) exhibit a consistent downward trend across proficiency levels (11.11% → 6.76% → 4.72%), indicating a gradual reduction in compounded deviations. As proficiency increases, errors tend to involve fewer interacting dimensions, with deviations becoming more localized rather than affecting multiple structural levels simultaneously.

In summary, these categorical distributions confirm that progress in Chinese character writing is uneven across orthographic dimensions. Stroke execution stabilizes relatively early, whereas component and structural accuracy continue to fluctuate as learners encounter greater lexical and formal complexity. This layered pattern highlights the incremental and non-uniform nature of Chinese character writing development among L1 Italian learners.

7.2.3 Incorrectly used character comparison

Then I explored the distributional characteristics of incorrectly used characters across different stages of interlanguage development. As shown in Table 7.4, the proportion of learners who produced incorrectly used characters increased steadily across proficiency levels: 50.00% (3/6) in the Beginner group, 80.95% (17/21) in the Intermediate group, and 91.67% (11/12) in the Advanced group.

Table 7.4

Proportion of student with incorrectly used characters across proficiency levels

Proficiency level	No. of student with errors	Total number of student	Percentage of student with errors
Beginner	3	6	50.00%
Intermediate	17	21	80.95%
Advanced	11	12	91.67%

This upward trend contrasts with the pattern of incorrectly written characters and highlights a different developmental trajectory in character use. Beginners showed the lowest proportion of such errors, as many avoided uncertainty by using Pinyin or limiting output to familiar characters, which concealed potential use mistakes. At the intermediate level, the ratio rose sharply to 80.95%, reflecting expanding vocabulary and the challenges of distinguishing

similar forms and meanings during semantic restructuring. By the advanced stage, nearly all learners (91.67%) produced at least one incorrectly used character, suggesting that while orthographic control had stabilized, fine semantic and stylistic precision remained difficult.

By calculating the total incorrectly used character ratio of each proficiency level group (see Table 7.5 for specific data), a comparative trend chart of the ratio in the three groups was obtained.

Table 7.5

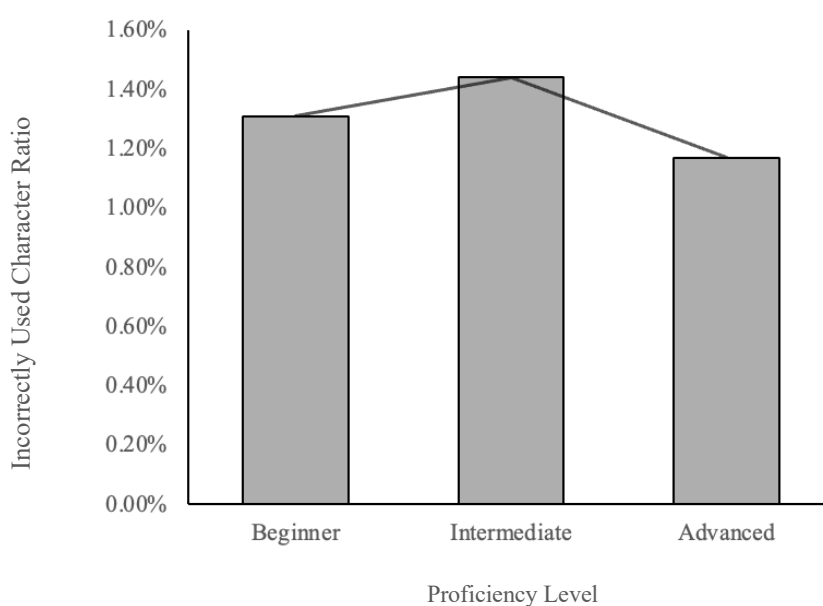
Number of incorrectly used characters across proficiency level

Proficiency level	No. of incorrectly used characters	No. of total characters
Beginner	37	2833
Intermediate	248	17238
Advanced	152	12960

As visually demonstrated in Figure 7.4, the statistical results of the present study display a “higher-in-the-middle” distribution pattern, with the Intermediate group exhibiting higher ratios than both the Beginner and Advanced groups.

Figure 7.4

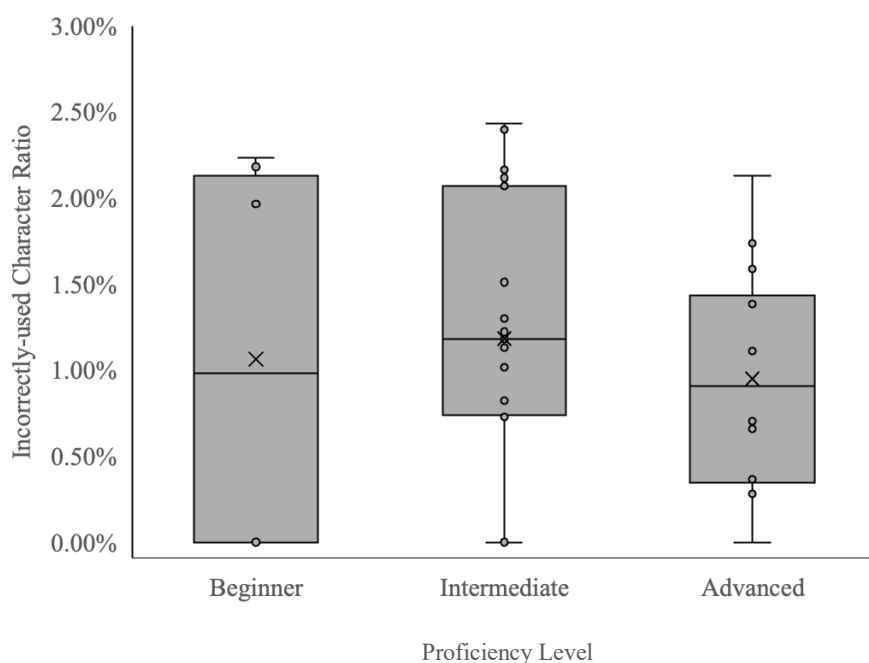
Incorrectly used character ratio across three proficiency level



This result is not entirely consistent with existing research. From Figure 7.4, the statistical results of this study clearly presented a distributional pattern where the Intermediate level was higher than both the Beginner and Advanced levels. Figure 7.5 illustrates the distribution of incorrectly used character ratios among the three groups. Although the median ratio of all groups remains below 1.50%, distinct distributional differences can be observed.

Figure 7.5

Boxplot of incorrectly used character ratios across proficiency levels



The Beginner group shows a median of 0.99%, an average of 1.40%, and the widest interquartile range (IQR = 1.72%), with values spanning from 0.00% to 2.23%. This wide dispersion indicates significant internal variation among beginners, suggesting a “bimodal differentiation” pattern. Nearly half of the learners in this group produced no incorrectly used characters (0%), while the rest exhibited relatively high error ratios approaching 2%. This contrast suggests divergent learning strategies and production behaviors: some beginners heavily rely on avoidance (e.g., omitting uncertain characters or replacing them with Pinyin), while others attempt full character production despite limited accuracy.

The Intermediate group demonstrates a median of 1.18%, an average of 1.26%, and a narrower IQR of 1.26%, with a range between 0.00% and 2.43%. Compared with the Beginner group, the overall error distribution becomes more centralized, but the upper outliers remain.

Approximately one-fifth of students in this group (19.0%) reached or exceeded a 2% error ratio, indicating that while most intermediate learners achieved relatively stable control of character use, some individuals still experienced difficulty in accurately mapping the relationships among form, sound, and meaning. This stage thus reflects a transitional phase in which the expansion of vocabulary and character inventory temporarily increases the likelihood of semantic confusion and inappropriate substitution.

The Advanced group displays the lowest median (1.00%) and the most compact distribution (IQR = 0.78%), with a range of 0.00%–2.13%. The smaller dispersion indicates that incorrectly used characters became less frequent and more uniform across learners. Nevertheless, a few outliers still exceed the 2% threshold, suggesting that certain residual errors persist even at higher proficiency levels. These stable, fossilized errors may reflect overgeneralization or entrenched interlanguage patterns that are resistant to correction through general instruction.

The developmental trajectory of incorrectly used characters across proficiency levels exhibits a non-linear progression—characterized by wide variability at the Beginner stage, concentration with persistent high outliers at the Intermediate stage, and partial fossilization at the Advanced stage.

I further investigated the distributional patterns of incorrectly used characters across the three proficiency levels to observe how learners’ morpheme awareness evolved with increasing language skill. First, I looked at the composition within each level (see table 7.6).

Table 7.6

Categorical statistics of incorrectly used characters in different proficiency level

Proficiency level	PSCs		GSCs		PGCs		OCs	
	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio
Beginner	2	5.41%	10	27.03%	12	32.43%	13	35.13%
Intermediate	38	15.32%	103	41.53%	64	25.81%	43	17.34%
Advanced	30	19.74%	56	37.50%	44	28.29%	22	14.47%

In the Beginner group, OCs and PGCs accounted for the highest percentages (35.13% and 32.43%, respectively), followed by GSCs at 27.03%, with PSCs being the lowest at 5.41%. This indicated that the orthographic representations were highly unstable in the introductory stage; students frequently wrote based on context or instinct—if a character “looked or sounded similar,” they used it—leading to a scattered distribution of error types. In the Intermediate stage, GSCs rose significantly to the first rank (41.53%), while OCs dropped to 17.34%, and PGCs also fell to 25.81%. This suggested that basic sound-form correspondence had become relatively stable, reducing sporadic and context-driven errors. The real hurdle had shifted to the fine-grained discrimination of components and structures. By the Advanced stage, GSCs remained the highest proportion (37.50%), but PSCs continued to climb to 19.74%, and PGCs rebounded to 28.29%. Meanwhile, OCs dropped further to 14.47%. In other words, errors in advanced writing were more easily triggered by homophonous competition and dual sound-and-form interference.

Figure 7.6

Categorical ratio of incorrectly used characters in different proficiency level

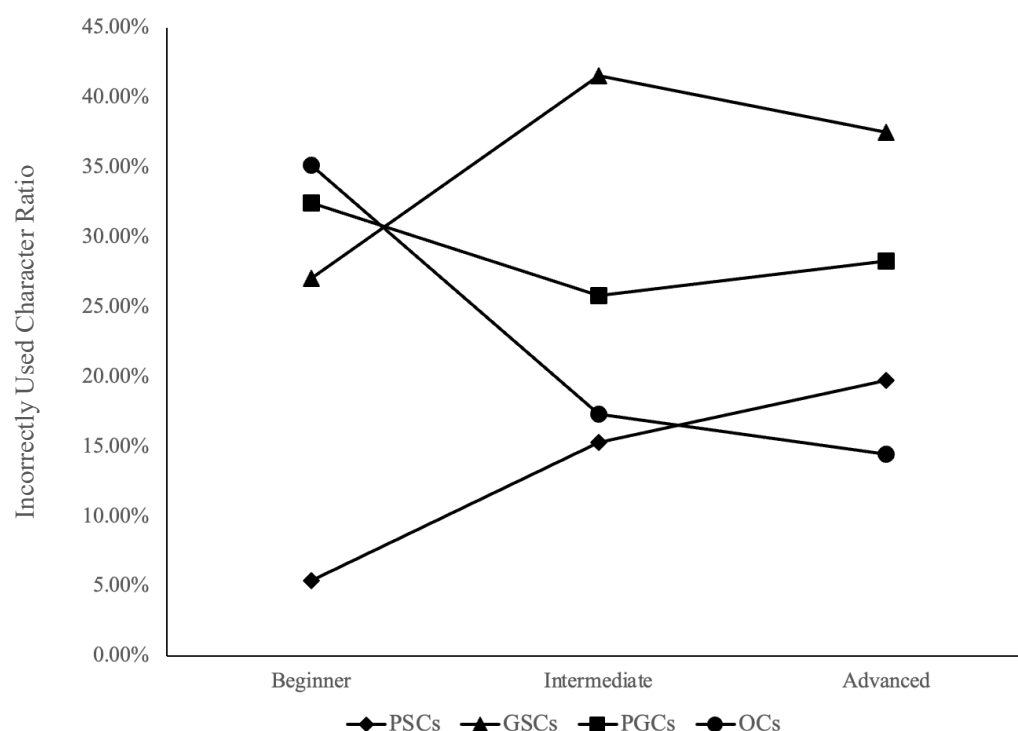


Figure 7.6 shows the evolution in the type of incorrectly used characters produced: First, the percentage of OCs steadily declined as proficiency increased (35.13%→17.34%→14.47%).

This suggests that the early, more sporadic errors—those driven by context or immediate judgment—gradually recede, shifting the error focus toward more structured orthographic challenges. Second, GSCs remained the leading error type in both the intermediate and advanced stages (41.53%, 37.50%). This indicates that the subtle discrimination of components and structures is a persistent bottleneck that runs throughout the mid-to-high proficiency levels and does not automatically resolve with overall skill improvement. Third, the proportion of PSCs increased with proficiency (5.41%→15.32%→19.74%). I interpret this as being due to the expansion of vocabulary and the increased exposure to lower-frequency items, which more frequently activates sets of homophonous or near-homophonous characters; consequently, the “retrieval of form via sound” is more susceptible to competing phonological neighbors. Finally, PGCs followed a trajectory of “initial drop, then rebound” (32.43%→25.81%→28.29%). The dip at the intermediate stage might be the result of component and structure training temporarily suppressing the dual interference, but this interference re-emerged during the advanced stage, likely due to increased material complexity and task demands in advanced writing.

Overall, the errors transitioned from a pattern of “scattered/context-driven mistakes” in the beginner stage to “structured orthographic challenges” in the intermediate and advanced stages. GSCs clearly represent the long-term main challenge, while the proportion of PSCs and PGCs is amplified in the advanced stage by increased vocabulary and task complexity.

7.2.4 Pinyin use

Table 7.7 presents the proportion of learners who used pinyin in their written data. The results show that pinyin use was widespread across all proficiency levels, but followed a non-linear developmental trend: 100.00% (6/6) of learners in the Beginner group used pinyin, compared to 66.67% (14/21) in the Intermediate group, and 75.00% (9/12) in the Advanced group.

Table 7.7*Proportion of student with Pinyin use across proficiency levels*

Proficiency level	No. of student with errors	Total number of students	Percentage of student with errors
Beginner	6	6	100.00%
Intermediate	14	21	66.67%
Advanced	9	12	75.00%

At the Beginner stage, every learner used pinyin, indicating that it served as an indispensable writing support during the early phase of Chinese acquisition. Since orthographic knowledge had not yet been established, learners relied heavily on pinyin to maintain fluency and completeness in expression. Pinyin thus functioned as a phonological intermediary, compensating for the lack of visual-orthographic representation and facilitating early written output.

In the Intermediate stage, the proportion of learners using pinyin dropped markedly to about two-thirds. Most learners were able to write primarily in Chinese characters and used pinyin only as a compensatory tool when they forgot a specific character or faced uncertainty in production. The role of pinyin shifted from a dominant phonological representation system to an auxiliary support strategy, reflecting a transitional stage from phonology-based to orthography-based processing in learners' writing development.

At the Advanced stage, the proportion of learners using pinyin rose slightly again to 75.00%, even though the writing tasks were identical across proficiency levels. This modest increase does not indicate renewed reliance on pinyin as a primary writing system. Rather, it reflects the continued use of pinyin as an optional writing strategy among some advanced learners. When encountering rare or temporarily unavailable characters, a small number of learners resorted to pinyin as an immediate placeholder to maintain written output. Although such instances were infrequent, they show that pinyin remains available as a supplementary resource in advanced writing.

Overall, the changing proportion of learners using pinyin across proficiency levels

demonstrates the stage-specific and functional nature of pinyin use in the development of Chinese writing. At the beginner stage, pinyin largely substitutes for unknown characters; at the intermediate stage, its use becomes more selective and limited; and at the advanced stage, it persists mainly as a backup strategy in isolated cases rather than as a systematic writing practice.

This non-linear pattern suggests that pinyin should not be treated simply as a writing error. Instead, it functions as a flexible auxiliary resource whose role diminishes but does not disappear as learners' character writing ability develops. In advanced learner writing, occasional pinyin use reflects strategic choice under specific task conditions rather than a general deficiency in orthographic competence.

Table 7.8

Proportion of student with Pinyin use across proficiency levels

Proficiency level	No. of Pinyin	Ratio of Pinyin
Beginner	1129	28.06%
Intermediate	538	3.00%
Advanced	234	1.77%

Table 7.8 presents the ratio of pinyin use among all learners (the ratio of Pinyin in total characters). The results show a clear downward trend across proficiency levels: 28.06% for the Beginner group, 3.00% for the Intermediate group, and 1.77% for the Advanced group (see Figure 7.7). Figure 7.7 visually highlights this sharp decline, showing a steep drop from beginners to intermediate learners and only a slight decrease thereafter.

Meanwhile, the boxplot in Figure 7.8 illustrates the internal variability across groups, revealing extremely high dispersion among beginners, mainly due to two outlier learners who wrote almost entirely in pinyin. While the intermediate and advanced groups show much narrower distributions, indicating more consistent orthographic performance.

Figure 7.7

Pinyin use ratio across three proficiency level

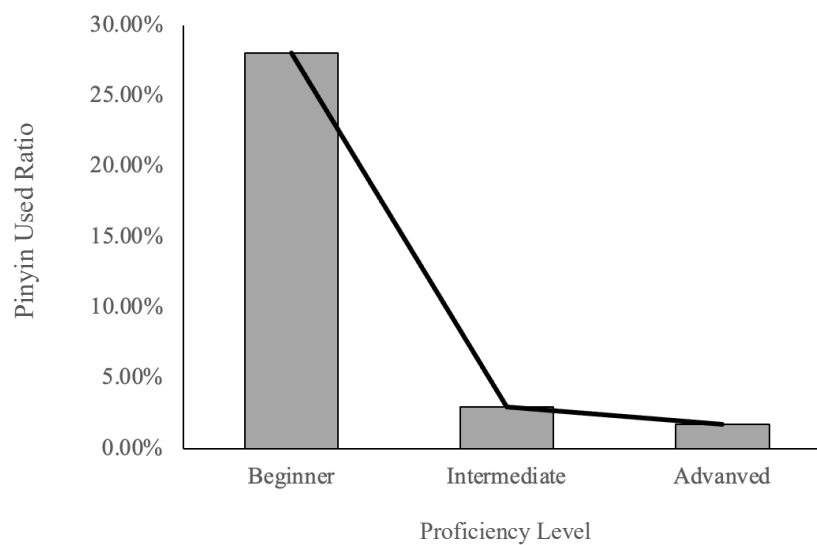
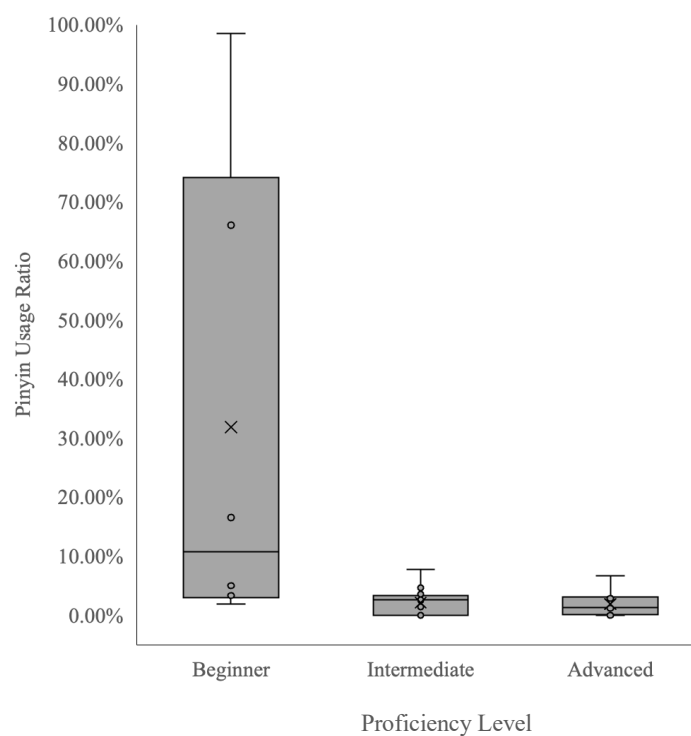


Figure 7.8

Boxplot of Pinyin use ratios across proficiency levels



The remarkably high proportion in the Beginner group is largely attributable to these two extreme learners (Students No. 13 and No. 17), who wrote almost entirely in pinyin, producing an artificially elevated aggregate ratio. However, their extreme behavior should not be viewed as an isolated anomaly; rather, it represents a typical developmental pathway in early Chinese character acquisition. At the initial stage, many learners adopt a phonological writing strategy,

relying on the familiar pinyin system to replace unfamiliar Chinese characters in order to maintain message completeness and communicative coherence. This phonology-based strategy serves as a transitional mechanism in cross-linguistic transfer, reflecting learners' natural tendency to depend on sound representation before orthographic knowledge is established and marking a necessary step from phonological mediation toward orthographic mastery. Consequently, the extreme cases of Students 13 and 17 are illustrative: although their output deviates from standard forms, it reveals a fundamental psychological mechanism in early Chinese learning—using pinyin to maintain consistency between input and output while orthographic knowledge is still under construction.

Table 7.9 excludes these two students and thus more accurately represents the writing behavior of the majority of learners.

Table 7.9

Number of Pinyin use across proficiency level (excluding No.13 and No. 17)

Proficiency level	No. of Pinyin	Total number	Pinyin ratio
Beginner	111	1768	6.28%
Intermediate	538	17238	3.12%
Advanced	234	12960	1.81%

After their removal, the proportion of pinyin use drops sharply to 6.28% in the Beginner group, 3.12% in the Intermediate group, and 1.81% in the Advanced group (also can be seen in Figure 7.9). Compared with the previous results, this reduction confirms that pinyin use was heavily concentrated among a small number of individuals, yet the overall declining trend remains consistent. In other words, even after excluding extreme cases, pinyin use continues to decrease with rising proficiency, indicating a gradual weakening of phonological dependence and a steady increase in orthographic automatization throughout the learning process. After excluding Students 13 and 17, the boxplot (Figure 7.10) still shows a clear downward trend in Pinyin use across proficiency levels, but the internal variability has markedly decreased, especially within the beginner group, where the previous extreme dispersion has been largely eliminated.

Figure 7.9

Pinyin use ratio across three proficiency level (excluding Student No.13 and No.17)

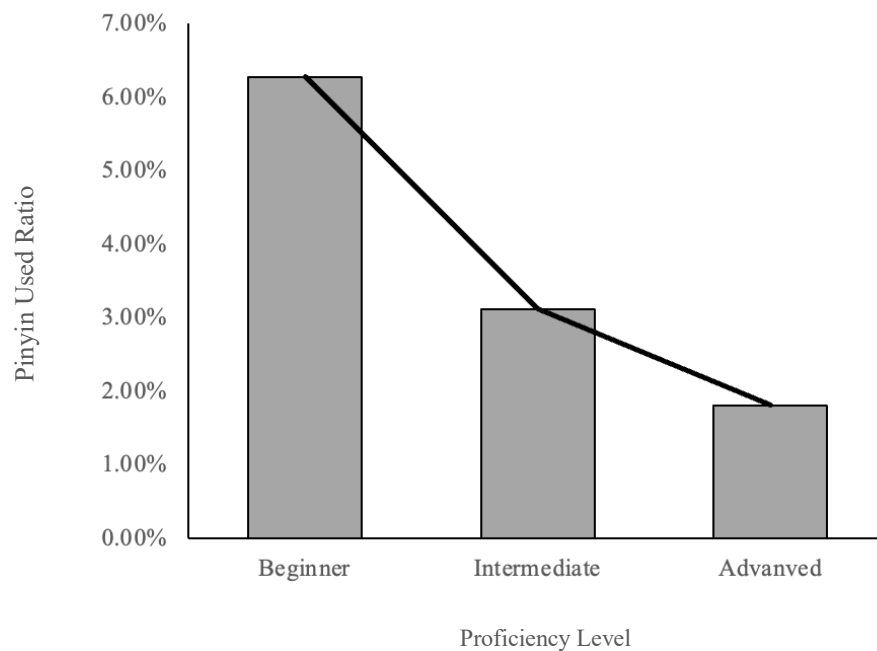
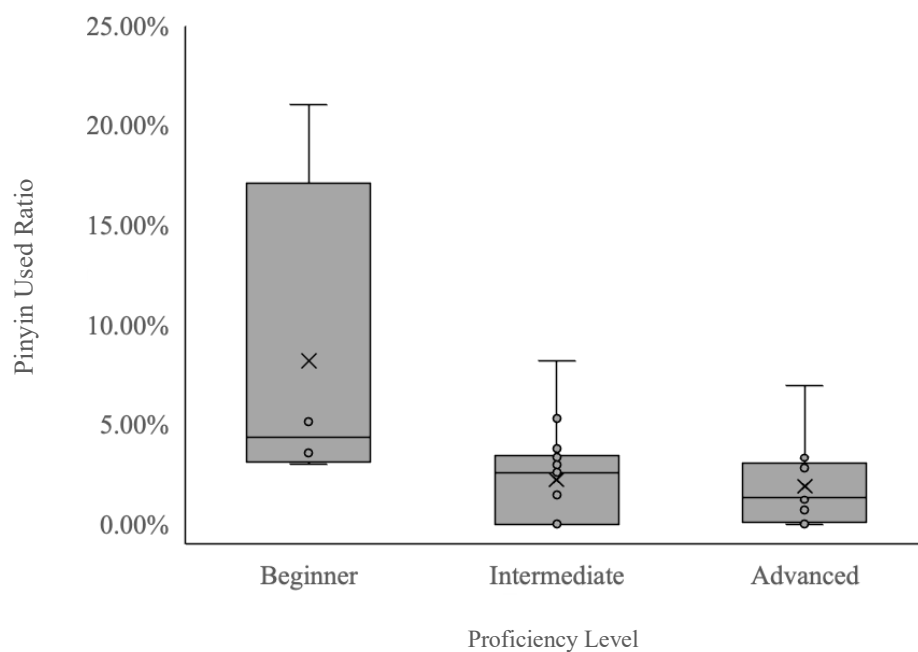


Figure 7.10

Boxplot of Pinyin use ratios across proficiency levels (excluding Student No.13 and No.17)



For beginner group, the Pinyin use ratios range from 3.01% to 21.09%, with a median of about 5% and an average of 8.21%. This indicates that while most beginners used Pinyin at a moderate level, one learner still showed relatively high reliance (over 20%). The dispersion within the group remains visible, reflecting individual differences in orthographic mastery—

some learners still depend on Pinyin for lexical gaps, while others have begun to rely mainly on Chinese characters.

For intermediate group, ratios vary from 0.00% to 8.19%, with most learners clustering between 2% and 4% (mean $\approx$ 2.87%). The box is narrower and lower than that of the beginner group, and several learners (about one-fourth) no longer used any Pinyin at all (0%). This concentration indicates that Pinyin use at this stage has become relatively stable and selective—mainly as a temporary aid when recalling forgotten characters, rather than a dominant writing medium.

For advanced group, ratios range from 0.00% to 6.97%, with a median around 2% and an average of 1.93%. The box is compact and close to the baseline, suggesting high stability of orthographic control. Only a few individuals occasionally resort to Pinyin as a compensatory mechanism (below 7%), typically when facing rare or low-frequency characters.

The three groups display a “high start – rapid decline – stable convergence” developmental pattern. As proficiency increases, the overall average Pinyin use drops from approximately 8% (Beginner) to 3% (Intermediate) and further to 2% (Advanced), while internal variability steadily narrows.

Overall, at the Beginner stage, pinyin accounts for the highest proportion, reflecting learners’ heavy dependence on it to sustain written production. Since orthographic knowledge is not yet established, learners tend to substitute pinyin for characters. This dependence, though manifesting as “pinyin-based writing,” is actually a constructive transitional strategy. Learners use the phonological system to support expression, building connections between sound and meaning that later facilitate the integration of form, sound, and meaning.

At the Intermediate stage, the pinyin ratio drops substantially. Most learners are able to produce complete texts in Chinese characters, and pinyin appears only in isolated cases, typically involving forgotten or low-frequency characters. This stage marks a clear shift in writing practice: pinyin moves from being a primary means of written expression to a temporary auxiliary device used in specific situations.

At the Advanced stage, pinyin use decreases further to 1.77% (1.81% after excluding outliers), indicating a high level of stability in character-based writing. At this level, pinyin no longer signals insufficient character knowledge but instead functions as an occasional fallback strategy, for example when a character cannot be immediately retrieved during writing. Such infrequent and context-dependent use shows that pinyin remains available as a supplementary option, rather than forming part of regular writing practice.

When these frequency data are considered together with Table 7.8, which shows that the proportion of learners who used pinyin remains relatively high across stages (100% for Beginners, 66.67% for Intermediates, and 75% for Advanced learners), a more nuanced pattern emerges. Although the overall amount of pinyin use declines sharply with proficiency, its distribution across learners remains broad. In other words, fewer characters are written in pinyin, but many learners continue to use it occasionally. At the Beginner stage, both high frequency and universal participation indicate widespread reliance on pinyin as a writing medium. At the Intermediate stage, reductions in both user proportion and use frequency reflect increased consistency in character writing. At the Advanced stage, while three-quarters of learners still used pinyin at least once, its occurrence was minimal and highly individualized.

Taken together, this combination of wide participation and low frequency illustrates a functional reorientation of pinyin use—from an explicit and dominant writing tool to an implicit and auxiliary one. Rather than disappearing entirely, pinyin is retained as a flexible support option that learners draw on selectively under specific writing conditions. This pattern highlights the non-linear nature of Chinese orthographic development: progress is characterized not by the complete abandonment of pinyin, but by its gradual repositioning from central to peripheral status within written production.

7.2.5 Summary and discussion of writing data comparison

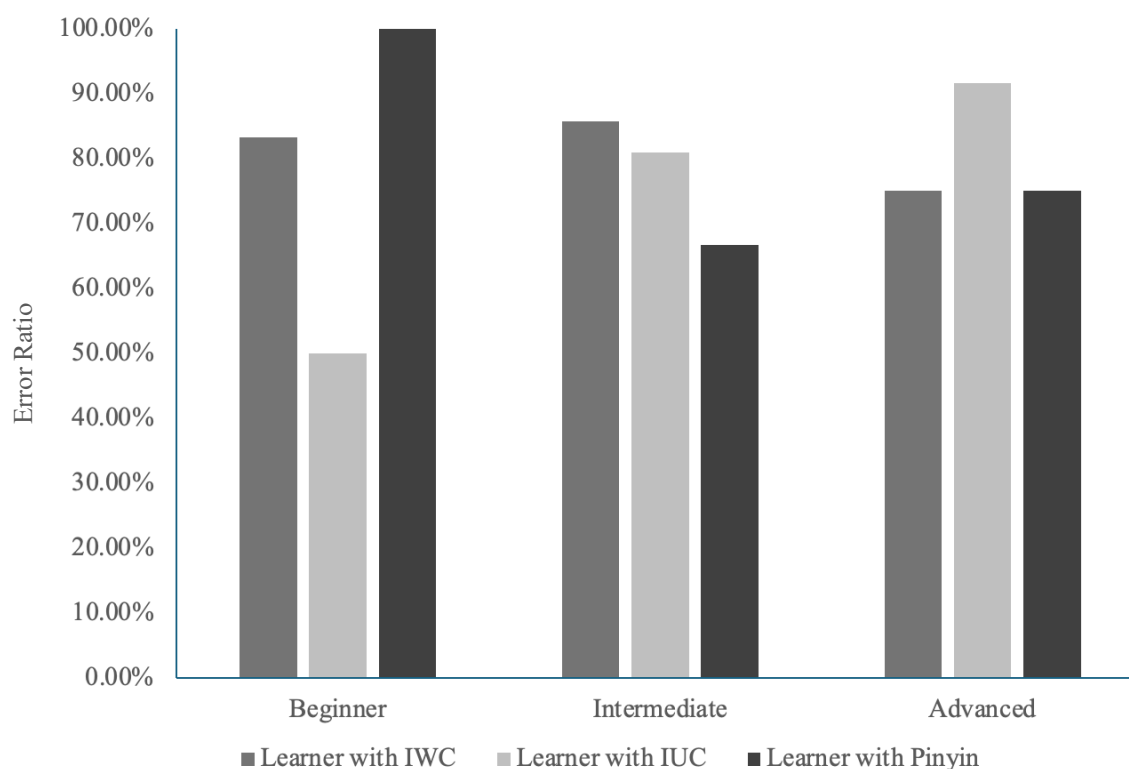
To better capture learners' overall progress in Chinese character acquisition, this section compares the results of incorrectly written characters, incorrectly used characters, and Pinyin use. The aim is to reveal how different error types reflect the developmental shift from phonological reliance to orthographic stabilization across proficiency levels.

Table 7.10*Proportion of student with three errors across proficiency levels*

Proficiency level	Students with IWC	Students with IUC	Students with Pinyin	Total number of students
Beginner	5	3	6	6
Intermediate	18	17	14	21
Advanced	9	11	9	12

Note. IWC = Incorrectly written characters, IUC = Incorrectly used characters

Table 7.10 lists the number and proportion of learners who exhibited each error type within the three groups, Figure 7.11 shows the proportions of learners who has made at least ont time of each errors across profivency levels. By combining the data on error proportions and learner distribution, a comprehensive picture emerges of how the three error types evolve throughout the process of Chinese character acquisition.

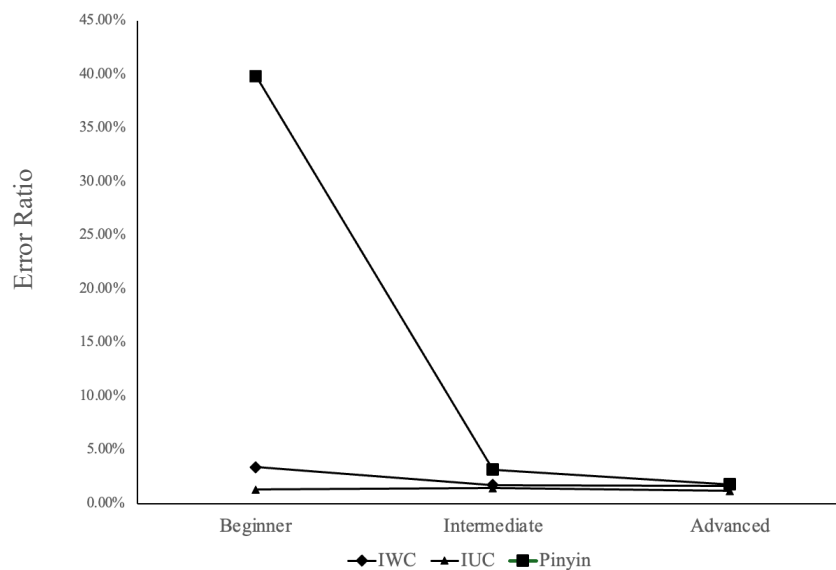
Figure 7.11*Proportion of student with three errors across proficiency levels*

Note. IWC = Incorrectly written characters, IUC = Incorrectly used characters

Figures 7.12 and 7.13 illustrate the distribution of three major types of writing errors across three proficiency levels. The line chart shows the proportion of erroneous characters in relation to the total number of written characters, while the boxplot visualizes the within-group dispersion of these error ratios.

Figure 7.12

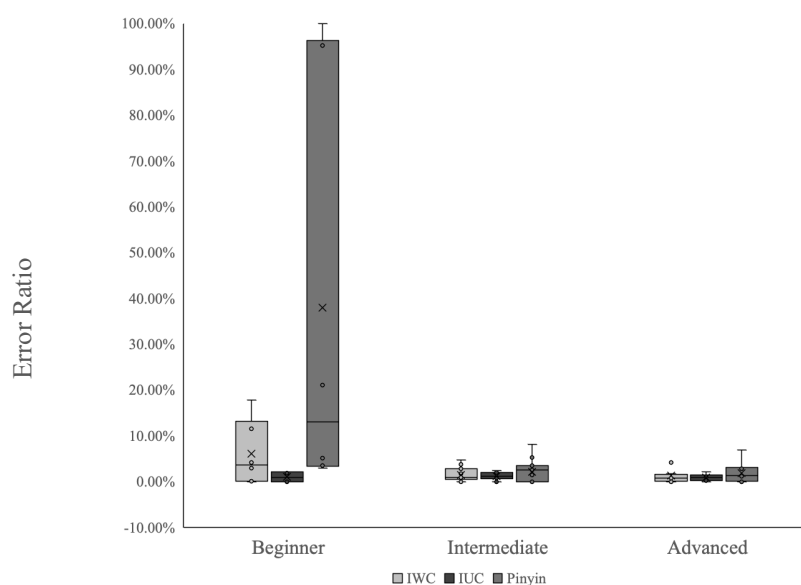
Comparative ratios of three errors across three proficiency levels



Note. IWC = Incorrectly written characters, IUC = Incorrectly used characters

Figure 7.13

Boxplot of three errors ratios across proficiency levels



Note. IWC = Incorrectly written characters, IUC = Incorrectly used characters

At the beginner level, all three types of errors were highly prevalent. Among the six learners, 83.33% produced incorrectly written characters, 50.00% produced incorrectly used characters, and all learners (100%) resorted to Pinyin. Pinyin use accounted for the largest proportion of total written output—approximately 40%—far exceeding the proportions of IWC (3–4%) and IUC (1–2%). The boxplot further indicates substantial variability among learners, particularly with respect to Pinyin use, suggesting large individual differences in strategy selection and output control. Overall, beginner learners tended to avoid uncertain character forms by using Pinyin, which reduced the surface proportion of character-level errors but resulted in extremely high phonological reliance. This pattern supports the interpretation of Pinyin as a transitional support mechanism that allows learners to sustain written production before orthographic knowledge is sufficiently developed.

At the intermediate level, the overall proportion of errors decreased markedly, but the three categories followed distinct developmental trajectories. The proportion of learners using Pinyin dropped to 66.67%, and its ratio in total writing fell sharply to around 3%. In contrast, the proportion of learners producing IWCs rose slightly to 85.71%, while the proportion producing IUCs increased substantially to 80.95%. This stage thus represents a critical restructuring phase. As learners increasingly relied on characters rather than Pinyin, instability in sound–form mapping became more salient, giving rise to frequent phonographic confusions. These incorrectly used characters were not primarily caused by semantic misunderstanding, but rather by incomplete consolidation of orthographic representations. At the same time, the narrowing of boxplots across all three categories indicates reduced variability and greater internal consistency among learners, suggesting that interlanguage systems were becoming more systematically organized despite local instability.

At the advanced level, participation rates remained relatively high—75.00% of learners produced IWCs, 91.67% produced IUCs, and 75.00% used Pinyin—but the actual error ratios were the lowest across all stages. Pinyin use fell below 2%, IWCs averaged around 1.5%, and IUCs slightly above 1%. The boxplots were the most compact, indicating minimal dispersion and a high degree of stabilization. These results suggest that advanced learners had largely

automatized character writing procedures, with very few errors in strokes or structure. However, sound–form linkages remained partially unstable. As learners’ character inventories expanded, the increased density of visually and phonologically similar forms led to cross-character interference, resulting in occasional substitutions. Such errors reflect competition within the orthography–phonology mapping network rather than semantic confusion. The slightly higher proportion of advanced learners using Pinyin compared to intermediate learners likely reflects strategic use rather than dependence, with Pinyin functioning as a temporary placeholder when encountering low-frequency or momentarily inaccessible characters.

Viewed longitudinally, the proportion of incorrectly written characters consistently decreased with increasing proficiency and stabilized at advanced levels. This trend aligns closely with the findings of Deng and Hu (2022), who similarly reported a steady reduction of writing form errors as learners progressed. Together, the two studies point to a staged regularity in the development of Chinese character writing ability: high error rates and variability at the beginner stage, substantial improvement at the intermediate stage, and relative stability at advanced levels. The convergence of findings across studies provides empirical support for the generalizability of this developmental trajectory.

In contrast, the developmental pattern observed for incorrectly used characters differs from some previous findings. While Deng and Hu (2022) reported a monotonic decline in misuse errors with increasing proficiency, the present results are more consistent with those reported by Zhang (2025; see Section 1.3), showing a temporary rise at the intermediate stage. A comparison across studies suggests that this discrepancy can be attributed to two main factors. First, differences in corpus type play a decisive role. Deng and Hu (2022) relied on examination essays produced under time pressure and evaluative constraints, conditions that tend to suppress spontaneous errors and foreground formal accuracy. In contrast, both Zhang (2025) and the present study draw on natural writing tasks such as classroom essays and routine writing, which allow learners’ habitual writing strategies and underlying orthographic knowledge to surface more fully. As a result, exam-based corpora more closely reflect performance under controlled conditions, whereas natural writing corpora provide a better

window into learners' underlying competence.

Second, differences in proficiency classification criteria affect comparability. Although all three studies adopt a three-level division, Deng and Hu (2022) base their grouping on a unified textbook-linked testing system, whereas Zhang (2025) and the present study define levels in relation to teaching progress and learning stage. Because these systems do not align in terms of vocabulary size, grammatical mastery, or writing experience, learners labeled as “intermediate” may differ substantially in absolute proficiency across studies. These differences in grouping criteria likely contribute to the divergent trends observed in incorrectly used character ratios.

Despite these discrepancies, the studies converge in highlighting the staged nature of Chinese character writing development. In natural writing contexts, the intermediate stage appears to be a critical period marked by temporary instability, during which learners transition from reliance on visual memory toward the internalization of orthographic rules. The “intermediate peak” in misuse errors observed in this study does not contradict earlier findings, but rather complements them by revealing a developmental pattern that becomes visible only when learners' writing is examined outside examination settings.

Further insight emerges when examining the internal composition of incorrectly used characters. The overall trajectory observed here broadly aligns with the findings of Jiang and Liu (2004), who noted that as character knowledge expands, errors related to graphic form tend to decrease while phonologically driven errors increase. Their explanation emphasized the growing role of phonological competition as learners' lexicons expand. Based on the present data, two additional points can be made. First, the peak of graphically similar confusions at the intermediate stage appears to reflect an “intra-form” bottleneck, in which component and structural distinctions have not yet been fully consolidated, rather than a simple linear shift from graphic to phonological interference. Second, at the advanced stage, vocabulary expansion introduces more low-frequency items, making retrieval via sound increasingly vulnerable to homophonous or near-homophonous competition. This accounts for the rebound observed in phonologically related substitutions, even as graphic confusions remain prominent.

Some previous studies have reported different distributions of error types (e.g., Wu et al., 2006; Gao & Meng, 2000; see Section 1.3). However, these studies were largely based on elicited or test-oriented data. In contrast, the present study draws on natural writing corpora in which learners produce language in genuine expressive contexts without external prompts. This design better reflects the automated state of learners' orthographic knowledge and the internal dynamics of writing processes. From this perspective, the observed pattern—a gradual reduction of graphic errors accompanied by increased phonological competition—reflects changes in the kinds of difficulties learners encounter at different stages of development.

Overall, the comparative analysis across proficiency levels confirms that Chinese character writing development follows a non-linear trajectory. Beginners rely heavily on phonological scaffolding, intermediate learners undergo a restructuring phase marked by heightened instability, and advanced learners achieve a high degree of orthographic automatization while continuing to negotiate fine-grained sound-form competition. These findings underscore the importance of stage-sensitive pedagogical intervention and caution against interpreting error reduction as a uniform or linear process.

7.3 Reading-aloud data comparison

7.3.1 Research questions

Similar to the analysis of reading-aloud data in Chapter 6, this section investigates the developmental patterns of Italian L1 learners' Chinese pronunciation across different proficiency levels. It focuses on three key research questions designed to trace how learners' phonological competence evolves and to identify stage-specific characteristics in tone and segmental production:

- i. How do tonal errors vary across proficiency levels? This question examines the frequency, distribution, and trend of tone errors among beginner, intermediate, and advanced learners. It aims to reveal how tonal accuracy improves (or fluctuates) with proficiency and whether specific tones remain persistently problematic across stages. Which tones do learners tend to favor in actual pronunciation? By analyzing the

relative proportions of the five Chinese tones in learners' spoken output, this question investigates whether learners show a bias toward certain tonal patterns (e.g., overuse of high-level tones) and what such tendencies reveal about their tonal perception and production strategies.

- ii. How do segmental errors distribute across proficiency levels? This part focuses on initial (consonant), final (vowel), and compound (segmental + tonal) errors, exploring how the accuracy of phonemic and syllabic coordination develops as learners progress. The goal is to uncover whether segmental mastery improves linearly or undergoes temporary instability during mid-level fluency development.

Together, these questions aim to map the developmental trajectory of phonological acquisition in NL-Italian learners of Chinese, revealing how their tonal and segmental systems interact, reorganize, and gradually stabilize across proficiency levels.

7.3.2 *Incorrect tone*

From the distribution of learners across the three proficiency levels, tonal errors appear to be highly prevalent at all stages (see Table 7.11).

Table 7.11

Proportion of student with five tone errors across proficiency levels

Proficiency level	Beginner	Intermediate	Advanced
No. of student with DC0	5	15	11
No. of student with DC1	5	21	12
No. of student with DC2	5	21	12
No. of student with DC3	5	21	12
No. of student with DC4	5	21	12
No. of total students	5	21	12

In the beginner group, all five learners produced tonal errors in each of the five tone categories, indicating that tonal control had not yet stabilized at this stage. In the intermediate group, the number of students producing tonal errors increased substantially: all 21 learners made errors

in DC1–DC4 tones, and 15 of them also mispronounced the neutral tone (DC0). This suggests that as vocabulary size and speech rate increase, intermediate learners face greater load in phonological processing, leading to a wider range of tonal instability. In the advanced group, the number of students with tonal errors decreased (11–12 students), implying a gradual improvement in tonal mastery with proficiency, although full control had not yet been achieved.

Overall, all five tone categories exhibited errors across all proficiency groups, indicating that tonal inaccuracies are not limited to specific tones but reflect a system-wide weakness in tonal acquisition. The peak error distribution observed in the intermediate group highlights that tonal learning difficulties persist beyond the early stages and tend to re-emerge during the developmental phase of interlanguage consolidation. Therefore, tone instruction should be emphasized throughout all learning stages, with a particular focus on the intermediate level, where training should target prosodic control and dynamic pitch modulation through multimodal pedagogical approaches such as tone contour visualization, prosodic imitation, and connected-speech shadowing exercises.

As shown in Table 7.12 and Figure 7.14, the overall proportion of tonal errors decreases progressively across proficiency levels. The beginner group exhibited the highest tone error ratio (30.73%), indicating an unstable mapping between tonal perception and production at the early learning stage. The intermediate group showed a slightly lower ratio (28.09%), yet tonal control remained inconsistent even with increased linguistic knowledge and oral experience. In the advanced group, the error ratio dropped to 23.43%, suggesting that learners' tonal mastery improves gradually with proficiency.

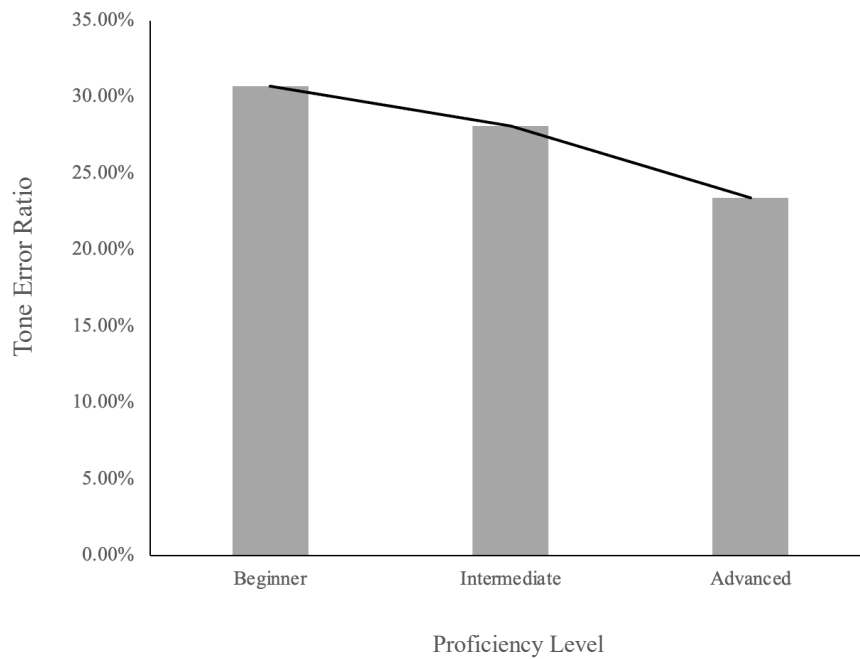
Table 7.12

Number of tone errors across proficiency level

Proficiency level	No. of errors	Total number of 5 tones	Error ratio
Beginner	1118	3638	30.73%
Intermediate	5102	18160	28.09%
Advanced	3105	13255	23.43%

Figure 7.14

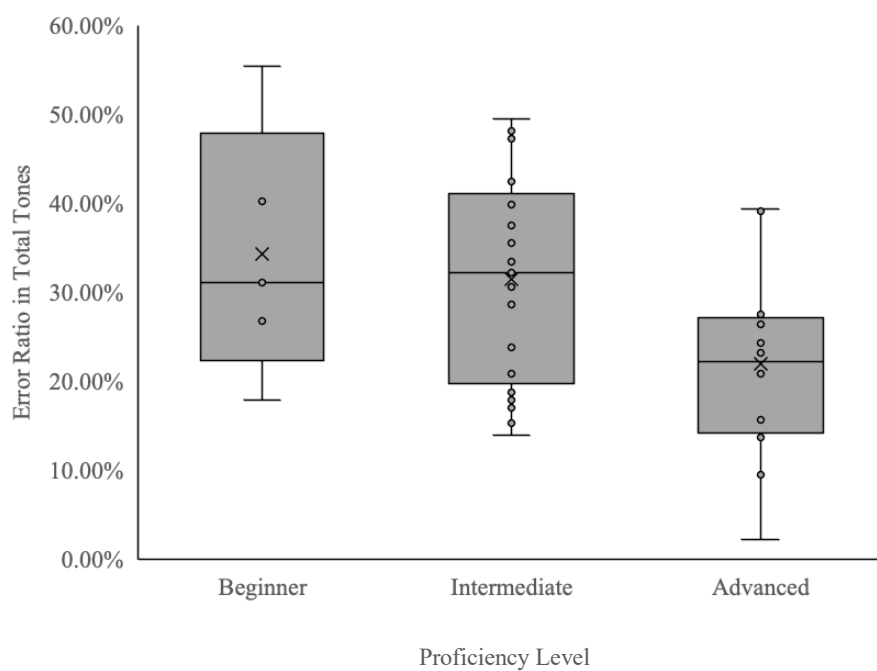
Tone errors ratios across proficiency level



The boxplot provides a clearer picture of the dispersion and central tendency of tone error ratios across proficiency levels (see Figure 7.15).

Figure 7.15

Boxplot of tone error ratios across proficiency levels



In the beginner group, the median error ratio is around 31%, but the overall range is wide (approximately 18%–55%), with a visibly large interquartile range. This indicates substantial individual variation: while some learners exhibit severe tonal inaccuracy, others demonstrate emerging tonal control, reflecting high heterogeneity within the group.

The intermediate group shows a slightly higher median (around 33%) but with a narrower spread, suggesting that although tonal errors remain frequent, individual performance becomes more consistent. This may imply that most intermediate learners have acquired the basic tonal system, but their control in continuous speech remains unstable.

In the advanced group, the median drops to about 24%, and the box narrows significantly, with a smaller overall range (around 2%–39%). This pattern indicates that tonal production becomes more stable at higher proficiency levels, as most learners cluster around lower error ratios. However, a few outliers with high error rates still exist, suggesting that tone stabilization is a prolonged process influenced by individual perceptual sensitivity or phonological awareness.

Overall, the boxplot reveals a clear developmental trend of decreasing variability and lowering median error ratios across proficiency levels, supporting the view that tonal acquisition is both gradual and subject to diminishing individual differences with increased proficiency.

Table 7.13

Total frequency of each tone across different proficiency levels

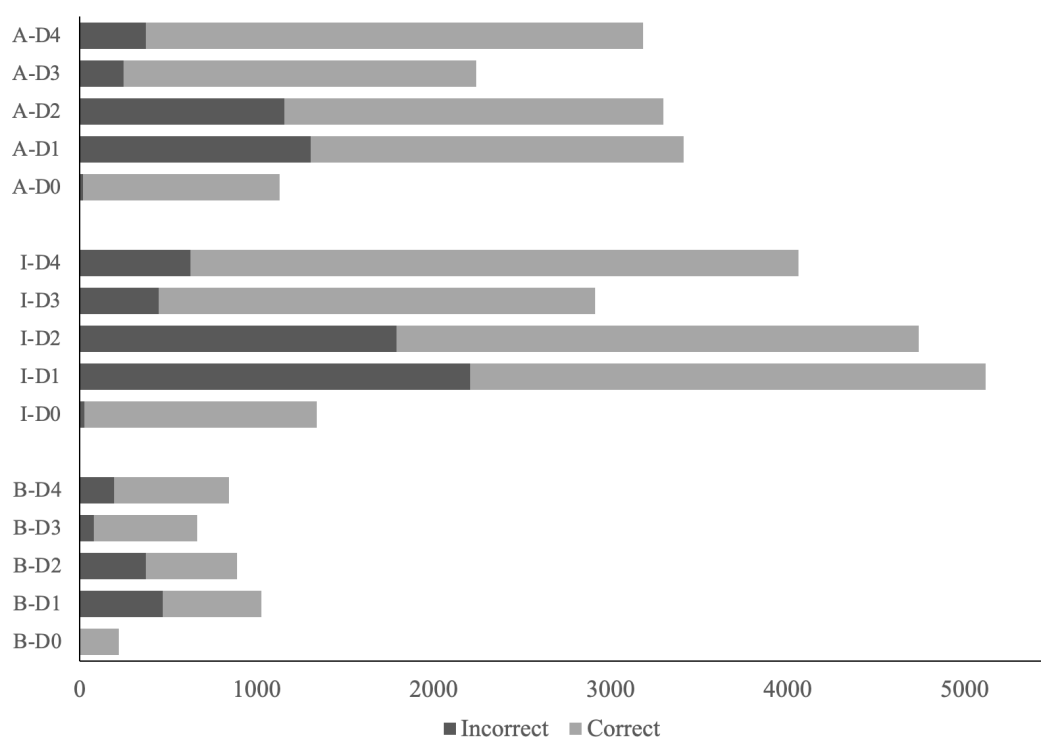
Tone types	Beginner	Intermediate	Advanced
D0	219	1337	1130
D1	1027	5117	3411
D2	888	4736	3295
D3	662	2912	2238
D4	842	4058	3181

The distribution of the five tones in learners' oral reading shows clear differences from the typical tonal distribution in native Chinese speech (see Table 7.13). In natural Chinese, D4

syllables are generally the most frequent, followed by D1 and D2, while D3 syllables occur less often and D0 (neutral tone) is the least frequent. However, in MICICL, learners of all three proficiency levels display a marked bias toward D1, with D1 being the most frequently produced tone across all groups, followed by D2 and D4, while D3 and D0 occur least often. Specifically, in the beginner group, D1 accounts for 28.3%, D2 for 24.4%, D4 for 23.2%, and D3 for 18.2%; in the intermediate group, D1 reaches 28.2%, D2 26.1%, D4 22.3%, and D3 16.0%; in the advanced group, D1 remains highest (25.8%), followed by D2 (24.9%) and D4 (24.0%). This “high-level tone defaulting” pattern suggests that learners tend to adopt the high-level pitch as their baseline tonal form, likely because it is articulatorily simple, acoustically stable, and perceptually salient, thereby weakening the contrast between other tonal categories.

Figure 7.16

Accuracy ratio of five tones across different proficiency levels



In terms of tonal accuracy (see Figure 7.16), the five tones exhibit clear stratified differences in error rates. Overall, D1 and D2 not only occur most frequently but also show the highest error ratios. In the beginner group, the error rates of D1 and D2 are 45.67% and 42.00%; in the intermediate group they remain high at 43.05% and 37.73%; and in the advanced group they

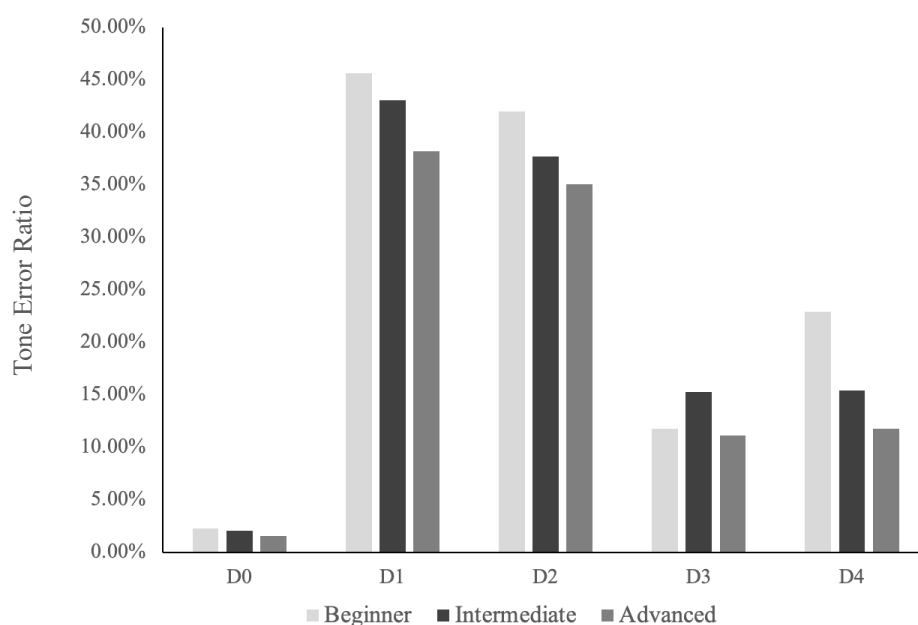
decrease slightly but still reach 38.23% and 35.05%. This indicates that, although tonal control over D1 and D2 improves somewhat with proficiency, these two tones remain the most persistent difficulties in tone acquisition. In contrast, D4 and D0 show the lowest error ratios, both steadily decreasing with proficiency (D4: 22.92% → 15.45% → 11.79%; D0: 2.28% → 2.09% → 1.59%), suggesting that falling and neutral tones are comparatively easier to perceive and maintain in connected speech.

Of particular interest is the non-linear developmental trajectory of D3, whose error ratio rises at the intermediate level but drops again at the advanced level (11.78% → 15.28% → 11.13%). Since the current statistics exclude tone-sandhi contexts (e.g., 3 → 2), this fluctuation cannot be attributed to contextual tone change; rather, it reflects shifts in the articulatory control mechanisms involved in producing D3 during the learning process.

It should also be noted that although D3 shows fluctuations at the intermediate stage, its overall error ratio remains comparatively low (see Figure 7.17).

Figure 7.17

Error ratio of five tones at each proficiency level



Across all proficiency levels, D3 errors stay within the range of 11%–15%, far lower than those of D1 and D2 (about 35%–46%) and even below the D4 error rate of beginners (22.92%). Hence, the fluctuation of D3 reflects temporary instability rather than systematic weakness—

it is not the most difficult tone for learners, but the tone most susceptible to task complexity, fluency demands, and attentional shifts at a specific developmental stage.

Overall, the five tones exhibit a general downward trend in error rates with increasing proficiency, though the degree of improvement varies. D1 and D2 remain persistent bottlenecks, D3 shows stage-specific fluctuation, and D4 and D0 are the most stable. This pattern demonstrates the non-linear nature of tonal acquisition: learners often experience temporary regression and reorganization in tonal control during the mid-developmental stage. Pedagogically, training should focus on strengthening D3 pitch control at the intermediate level through tone-contour visualization, delayed imitation, graded reading, and rhythmic prosody practice, helping learners preserve full pitch movement and avoid insufficient rise or contour drift. At the same time, instruction should consistently reinforce high-low contrast and pitch differentiation for D1 and D2 across all levels to promote overall stabilization of the tonal system.

Further quantitative results show that D1 and D2 together account for approximately 80% of all tonal errors across proficiency levels (75.4% for beginners, 78.4% for intermediates, and 79.3% for advanced learners), confirming that these two tones are the primary sources of tonal inaccuracy. By contrast, D4 and D3 contribute only about 10%–15% each, and D0 less than 1%, posing virtually no problem. In other words, learners' tonal systems display a structural asymmetry of “high-frequency, high-error” versus “low-frequency, low-error.” The higher a tone's frequency of occurrence, the greater its likelihood of being produced incorrectly—a phenomenon known as the “frequency–accuracy paradox.” This paradox suggests that frequent exposure or use does not guarantee better mastery; rather, tones that occur most often, with subtle pitch contrasts and lower attentional salience, can remain the most persistent learning challenges.

Table 7.14 shows the specific tone substitution patterns (i.e., how each of the five tones was mispronounced across different proficiency levels), the data reveal clear and systematic regularities in all three groups.

Table 7.14*Frequency of tone substitutions across proficiency levels*

Beginner	DC0	DC1	DC2	DC3	DC4
D0	0	0	1	0	4
D1	16	0	104	194	155
D2	9	114	0	95	155
D3	1	26	4	0	47
D4	4	76	63	50	0
Intermediate	DC0	DC1	DC2	DC3	DC4
D0	0	4	11	1	12
D1	84	0	512	697	910
D2	10	465	0	460	852
D3	2	104	37	0	302
D4	30	192	265	140	0
Advanced	DC0	DC1	DC2	DC3	DC4
D0	0	1	5	2	10
D1	20	0	328	386	570
D2	10	265	0	263	617
D3	2	60	11	0	176
D4	17	118	162	78	0

Overall, learners at all stages show an overreliance on the high-level tone (D1) — that is, a strong tendency to “flatten” or “raise” other tones toward D1. In addition, there is frequent confusion between rising and falling tones, especially the bidirectional substitutions among D1, D2, and D4.

In the beginner group, learners have not yet established clear categorical boundaries between tones. The data show that among syllables actually pronounced as D1, most were originally third-tone targets (194 instances) or fourth-tone targets (155 instances), followed by second tones (104 instances), indicating that third and fourth tones are most often “flattened” into a high-level tone. Syllables pronounced as D2 mainly came from target D4 (155) and D1 (114), suggesting that learners had difficulty switching between rising and falling pitch

movements. As for target D4, it was frequently mispronounced as D1 (76), D2 (63), or D3 (50), showing that its falling contour was not yet mastered. Overall, beginners tend to pronounce a variety of tones as high-level or falling tones, reflecting blurred tonal contrasts and insufficient pitch control.

In the intermediate group, the total number of tonal errors increased, with target D4 showing the highest instability. Fourth tones were most often pronounced as D1 (910) or D2 (852), followed by D3 (140), suggesting that learners still struggled to control pitch declination. Target D1 was frequently mispronounced as D3 (697) or D4 (910), indicating instability in maintaining high-level pitch during connected speech. Target D2 was often produced as D4 (852) or D3 (460), showing that the rising portion of the contour remained insufficient. Overall, intermediate learners exhibited blurred boundaries between rising and falling tones and unstable control over high versus low pitch levels, reflecting a stage of adjustment in prosodic control during fluent speech production.

In the advanced group, tonal errors decreased overall, but the main confusion still occurred between the high-level and falling tones. Target D4 continued to be frequently mispronounced as D1 (570) and D2 (617), suggesting that learners tended to “raise” the falling contour in spontaneous reading. Target D3 was also often read as D1 (386), showing that the low-rising contour was easily “flattened” in rapid speech. Although overall accuracy improved, advanced learners showed subtle post-automatization degradation—in pursuit of fluency, they often sacrificed pitch contrast precision and tonal distinctiveness.

In summary, all three proficiency levels display a clear “high-level tone default” tendency: when uncertain, learners are most likely to pronounce a target tone as D1. Furthermore, target D4 (falling tone) and target D3 (low-rising tone) are the most error-prone tones, frequently shifting toward D1 or D2. Although the total number of errors decreases and the distribution becomes more concentrated with proficiency, the distinction between high and low tones (D1/D4, D2/D4) remains a long-term challenge for Italian learners in Chinese tone acquisition.

Building upon these findings, the tone-substitution data further confirm that the high-level tone (D1) functions as a default pitch target across all proficiency levels. Learners

frequently flatten or raise other tonal contours toward D1, reflecting a tendency to neutralize tonal contrasts through pitch simplification. This phenomenon, combined with the frequent confusion between rising (D2) and falling (D4) tones, suggests that learners' tonal representations are anchored around a high and perceptually salient pitch range, while low or complex contours (such as D3's low-rising shape) remain less stable.

Taken together, the overall results demonstrate that tonal acquisition among L1-Italian learners is characterized by gradual stabilization accompanied by structural asymmetry: tones with higher pitch (D1, D2) dominate the phonetic space, whereas lower or dynamically changing tones (D3, D4) are more vulnerable to deviation or simplification. This imbalance explains the persistent D1 defaulting and the recurrent D2–D4 substitution patterns observed across all stages.

7.3.3 Segmental errors (SC, YC and QC)

The distribution of learners who produced initial (SC), final (YC), and multi-error (QC) mistakes across the three proficiency levels (see Table 7.15) reveals clear stage-specific differences. It should be emphasized that the “ratio” discussed here does not refer to the proportion of erroneous syllables among all syllables, but rather to the proportion of students within each group who exhibited a given error type. This index therefore reflects the prevalence of error types at the learner level, rather than the density of errors within individual performances.

Table 7.15

Proportion of student with segmental errors across proficiency levels

Proficiency level	Students with SC	Students with YC	Students with QC	Total number of students
Beginner	3	4	3	5
Intermediate	10	13	20	21
Advanced	5	6	9	12

Overall, segmental errors are highly prevalent across all proficiency levels. Even in the

advanced group, more than 40% of students still displayed problems in producing initials or finals. This finding indicates that segmental deviations are not confined to the early stages of Chinese learning but represent a persistent and dynamic feature throughout the acquisition process.

Among the three categories, QC show the highest and most variable proportions: 60.0% in the beginner group, surging to 95.2% in the intermediate group, and then slightly decreasing to 75.0% in the advanced group. Because QC includes not only co-occurring errors of initials and finals but also combined segmental–tonal errors, this type reflects learners’ difficulty in coordinating multiple phonological subsystems—namely, the integration of segmental (initial–final) and suprasegmental (tone) dimensions. The near-universal occurrence of QC among intermediate learners suggests that while they can control single phonemic segments, they still struggle to maintain stable synchronization among initials, finals, and tones within a coherent syllable structure. The subsequent reduction of QC in the advanced group indicates gradual improvement in inter-system coordination, though fluctuations persist at the system level.

By contrast, the proportions of students showing single-segment errors (SC and YC) decline steadily with increasing proficiency. The proportion of learners with initial errors dropped from 60.0% (beginner) to 47.6% (intermediate) and 41.7% (advanced), while those with final errors fell from 80.0% to 61.9% and 50.0%. This pattern suggests that control over individual segmental units stabilizes relatively early, yet difficulties remain when segmental and tonal elements must be integrated simultaneously in fluent speech.

Taken together, the distribution of students with each error type exhibits three key tendencies: QC show the highest prevalence and the sharpest fluctuation, following a “mid-level peak” pattern (60.0% → 95.2% → 75.0%), indicating temporary instability during phonological integration; SC and YC decrease progressively, reflecting linear improvement in single-segment mastery; The intermediate stage constitutes the critical phase of phonological instability, when learners’ fluency increases but precision in coordinating segmental and tonal control temporarily weakens.

This trend parallels the non-linear patterns observed in tonal errors: learners often

experience a rebalancing phase in mid-level development, during which gains in fluency temporarily come at the cost of fine-grained phonetic control.

Based on Table 7.16, Figure 7.18 and Figure 7.19, clear stage-specific patterns and non-linear trends can be observed in the three types of segmental errors across the three proficiency levels. Although all three error ratios remain relatively low (below 1%), the variation in their magnitude and distribution reveals how learners' phonological systems develop from local segmental control toward global coordination.

Table 7.16

Segmental error ratios across proficiency level

Proficiency level	SC ratio	YC ratio	QC ratio
Beginner	0.08%	0.22%	0.99%
Intermediate	0.15%	0.12%	0.52%
Advanced	0.26%	0.11%	0.45%

Figure 7.18

Comparative ratios of three errors across proficiency levels

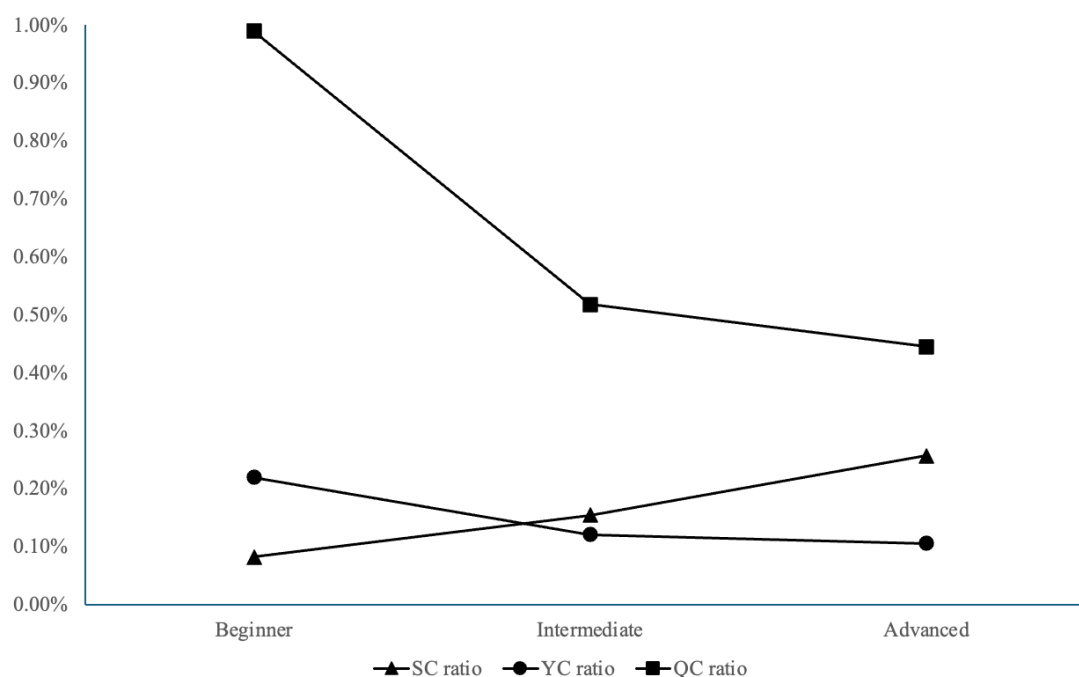
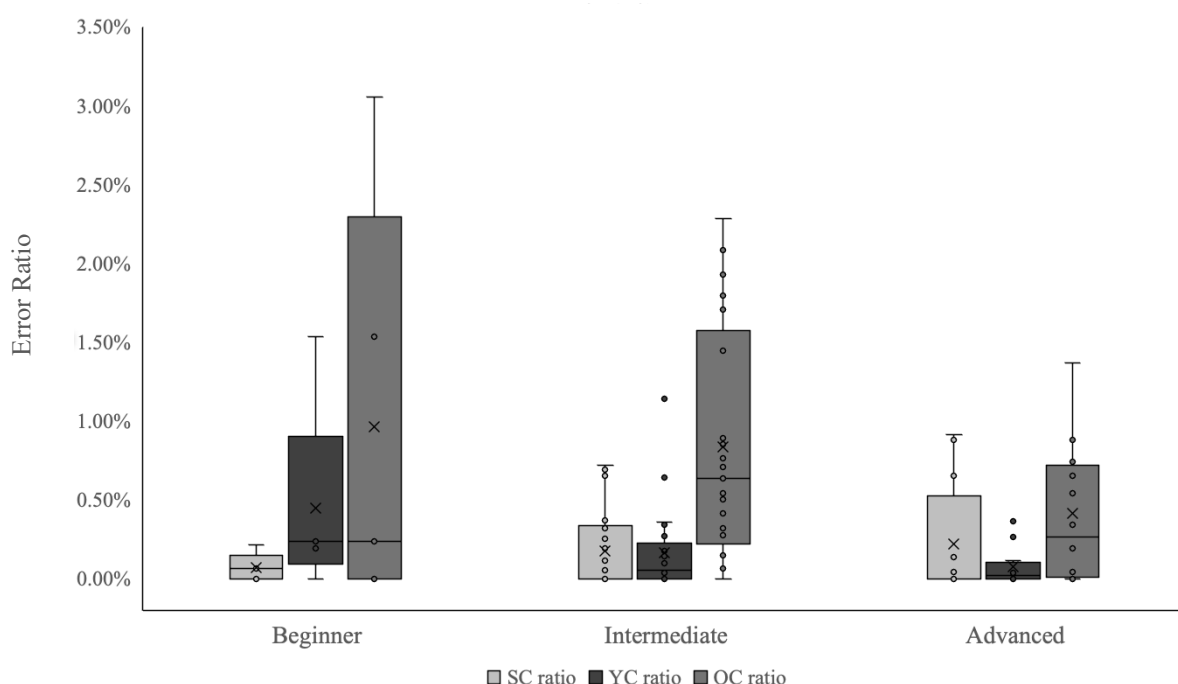


Figure 7.19

Boxplot of three segmental errors ratios across proficiency levels



Overall, the beginner group shows the highest error ratios (SC = 0.08%, YC = 0.22%, QC = 0.99%), indicating that learners at the early stage have not yet established stable coordination among initials, finals, and tones. The dominance of QC in this group suggests that most pronunciation deviations involve compound or co-occurring errors spanning multiple phonological subsystems (segmental + tonal). In the intermediate group, the overall error ratios decrease notably (SC = 0.15%, YC = 0.12%, QC = 0.52%), showing that learners gain better segmental control; however, QC remains relatively high, reflecting that integration between segmental and tonal features is still unstable. The advanced group demonstrates further stabilization (SC = 0.26%, YC = 0.11%, QC = 0.45%), with the ratio of multi-errors reduced to below 0.5%, indicating that the phonological system has become largely coordinated.

The line chart highlights differentiated developmental trajectories among the three error types. QC show a steady downward trend (0.99% → 0.52% → 0.45%), reflecting continuous improvement in segmental–tonal integration and overall phonological coordination. By contrast, SC exhibit a slight increase (0.08% → 0.15% → 0.26%), while YC decrease consistently (0.22% → 0.12% → 0.11%). This suggests that learners gain more stable control

over finals than over initials as proficiency rises. The modest increase in SC at the advanced stage may be linked to the fluency effect: as learners achieve greater naturalness and speech rate increases, certain segmental boundaries become less distinct, resulting in minor reductions in articulation clarity rather than genuine regression in phonological competence.

The boxplot further illustrates the distributional characteristics and individual variation within each group. The beginner group shows the highest median and greatest dispersion, especially for QC ratios, where some individuals exceed 3%, indicating pronounced inter-learner variability. The intermediate group displays a noticeably smaller box but still includes several high outliers (mainly in QC), suggesting that while most learners cluster around lower error ratios, a few still struggle with phonological integration. The advanced group presents the smallest and most compact boxes for all three categories, demonstrating that learners' pronunciation accuracy and stability converge markedly as proficiency increases.

Taken together, the quantitative and distributional evidence reveals a general developmental pattern of “high at the beginning – integration in the middle – stabilization at the advanced level.” QC remains the largest and most variable category, serving as a key indicator of overall system stabilization. SC and YC ratios are consistently low and converge over time, reflecting the early consolidation of single-segment control.

7.3.4 Summary and discussion of reading-aloud data comparison

In this section, syllable-initial errors (SC) and final errors (YC) are merged and analyzed as segmental errors. This decision is motivated by both empirical and theoretical considerations. Quantitatively, SC and YC each occur at extremely low frequencies in the reading-aloud data, remaining below 0.5% across all proficiency levels. Analyzing them separately would therefore yield minimal variation and provide limited insight for cross-level comparison. From a phonological perspective, initials and finals jointly constitute the segmental tier of the Chinese syllable and are processed through similar articulatory and perceptual mechanisms, whereas tones belong to the suprasegmental tier and operate at a higher level of phonological organization. Grouping SC and YC as segmental errors thus aligns with the hierarchical structure of the phonological system and allows for a clearer contrast between segmental and

tonal development.

Table 7.17

Errors ratios in reading-aloud data across proficiency level

Proficiency level	DC ratio	SC+YC ratio	QC ratio
Beginner	30.73%	0.30%	0.99%
Intermediate	28.09%	0.28%	0.52%
Advanced	23.43%	0.36%	0.45%

Figure 7.20

Errors ratios in reading-aloud data across proficiency level

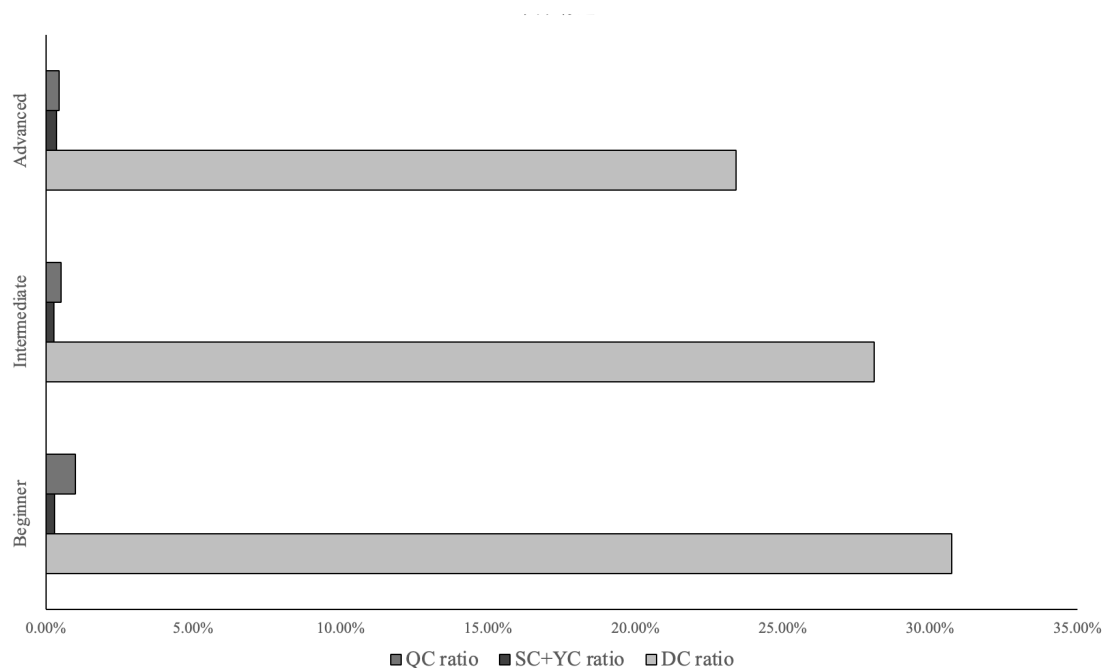


Table 7.17 and Figure 7.20 present the distribution of tonal errors (DC), segmental errors (SC+YC), and compound errors (QC) across proficiency levels. A clear hierarchical pattern emerges. Tonal errors dominate at all stages, with proportions far exceeding those of segmental and compound errors. The DC ratio reaches 30.73% at the beginner level, decreases to 28.09% at the intermediate level, and further drops to 23.43% at the advanced level, indicating a gradual but steady improvement in tonal control with increasing proficiency. By contrast, segmental errors remain consistently low overall, but display a non-linear pattern: their ratio decreases

slightly from the beginner level to the intermediate level (0.30% → 0.28%) and then rises at the advanced level (0.36%). Compound errors show a steady decline across levels (0.99% → 0.52% → 0.45%), remaining marginal throughout.

Within-group comparisons further underscore the dominance of tonal errors, which account for more than 95% of all phonological deviations at every proficiency level. This confirms that tone control constitutes the primary bottleneck in Chinese phonological acquisition for L1 Italian learners. In contrast, the consistently low proportion of segmental errors suggests that learners establish relatively stable articulatory control over initials and finals at an early stage. The slight rebound of segmental errors at the advanced level, however, interrupts the otherwise downward trend and calls for interpretation.

From a developmental perspective, this rebound does not indicate regression. Rather, it reflects a dynamic rebalancing between automation and precision. As learners reach advanced proficiency, they increasingly pursue naturalness, fluency, and discourse-level coherence. Under conditions of faster speech rate and greater prosodic complexity, articulation becomes highly automated, and fine-grained segmental monitoring may temporarily weaken, resulting in minor imprecision. At the same time, attentional resources are redistributed from syllable-level accuracy toward higher-level prosodic organization, such as intonation, rhythm, and information structure. The resulting increase in segmental errors therefore reflects an adjustment phase accompanying advanced automatization rather than a decline in phonological competence.

The developmental trajectory of tonal errors reveals a clearer and more linear pattern. Tonal perception and production are weakest at the beginner level, moderately stabilized at the intermediate level, and significantly improved at the advanced level. Nevertheless, the relatively small gap between proficiency groups—approximately seven percentage points—suggests that tone acquisition is a slow and cumulative process rather than one characterized by rapid improvement through exposure alone. This finding highlights the persistent difficulty of tone mastery and underscores the need for sustained tonal training throughout the learning process.

A closer examination of tonal subtypes further illuminates this pattern. The increase in third-tone (D3) errors observed at the intermediate level can be attributed to several interacting factors. Acoustically, D3 has the widest pitch span, involving a low onset followed by a rising contour. While intermediate learners are generally able to perceive tonal contrasts, they often struggle to control both the starting pitch and the extent of pitch rise in fluent speech, leading to flattened or incomplete contours. In addition, intermediate learners typically read longer and more complex texts than beginners, increasing phonological load and prompting a shift of attention toward fluency and semantic coherence. Under such conditions, tones with longer duration and greater contour complexity, such as D3, are particularly vulnerable to simplification. Furthermore, tonal category boundaries remain fuzzy at this stage, especially between D2 and D3, resulting in over-rising or prematurely initiated pitch movement and consequent deviation.

The subsequent decline in D3 error rates at the advanced level indicates that learners gradually establish more stable tonal control through prolonged exposure and practice. This improvement can be attributed to enhanced speech automatization, improved perceptual precision in distinguishing low-rising and high-rising contours, and more mature phonological awareness that supports active monitoring and self-correction. Together, these factors facilitate positive transfer from tonal perception to production.

From a longitudinal perspective, the reading-aloud data thus reveal differentiated developmental trajectories across phonological tiers. Tonal errors exhibit the largest magnitude and the most persistent difficulty, segmental errors remain minimal but show a temporary rebound associated with advanced automatization, and compound errors steadily decline, reflecting improved coordination between segmental and suprasegmental subsystems. This pattern illustrates the hierarchical nature of phonological development: learners first achieve segmental stability, then progressively integrate tonal control, and finally undergo fine-tuning as fluency increases.

Pedagogically, these findings suggest that pronunciation instruction should be stage-sensitive and tier-aware. For beginners, instruction should focus on establishing basic syllable

structure awareness and clear segmental articulation. At the intermediate level, training should reinforce the coordination between segmental and tonal elements, helping learners maintain tonal accuracy under increasing phonological load. At the advanced level, instruction should not concentrate exclusively on tone accuracy, but also on preserving segmental clarity in fluent, connected speech. Techniques such as graded speech-rate training, slow–natural speed contrast reading, tone contour visualization, and recording-based self-analysis can support learners in balancing fluency and precision, thereby promoting stable and natural spoken Chinese.

7.4 Cross modal comparison

7.4.1 *Incorrectly written characters reading-aloud*

From a cross-modal perspective (see Table 7.18), learners of different proficiency levels show a generally high degree of consistency in their reading-aloud performance of incorrectly written characters. The results indicate that although many learners produced orthographic deviations during the writing task, they were still able to pronounce the corresponding characters correctly when reading aloud, suggesting a relatively strong dissociation between orthographic and phonological representations.

Table 7.18

Number of student who read incorrectly written characters correctly

Proficiency level	No. of correctly read students	No. of total students
Beginner	4	4
Intermediate	16	18
Advanced	9	9

Specifically, in the beginner group, all four students who produced incorrectly written characters were able to pronounce at least one of those characters correctly during reading-aloud-aloud, resulting in a 100% success rate. In the intermediate group, 18 students produced incorrectly written characters, among whom 16 were able to read at least one of their incorrectly written forms correctly (88.9%). In the advanced group, all nine students showed at least one

instance of “incorrectly written but correctly read” correspondence, again yielding a 100% success rate.

It should be noted that the “correctly read” here refers to learners who have shown at least one instance of correct pronunciation for an incorrectly written character, rather than those who correctly pronounced all of their incorrectly written characters. In other words, the data reflect learners’ potential ability to connect orthographic and phonological forms across modalities, rather than the overall percentage of full reading–writing consistency.

Among the intermediate learners, the two students who did not show any “incorrectly written but correctly read” instances had written only one and three incorrect characters respectively, none of which were pronounced correctly. These isolated cases may indicate that they had not yet formed stable phonological–orthographic associations for those specific items—representing lexical-level gaps in form–sound mapping. Apart from these two exceptions, all other learners exhibited at least one case of “incorrectly written but correctly read” suggesting that they were able to rely on phonological compensation in cross-modal processing. Even when their visual representations of characters were inaccurate, they could still retrieve the correct pronunciation from their phonological memory.

Overall, students at all proficiency levels demonstrated the “incorrectly written but correctly read” phenomenon, indicating that the memory of Chinese characters is predominantly phonologically mediated—learners tend to retain and access the sound of a character even when its visual form is unstable or inaccurate. This finding highlights a fundamental characteristic of Chinese character learning among TL learners: the primacy of phonological memory over orthographic precision in early and intermediate stages of acquisition.

From the perspective of cross-modal correspondence between writing and reading-aloud (see Table 7.19 and Figure 7.21), clear differences emerge among the three proficiency groups in the reading-aloud accuracy of incorrectly written characters.

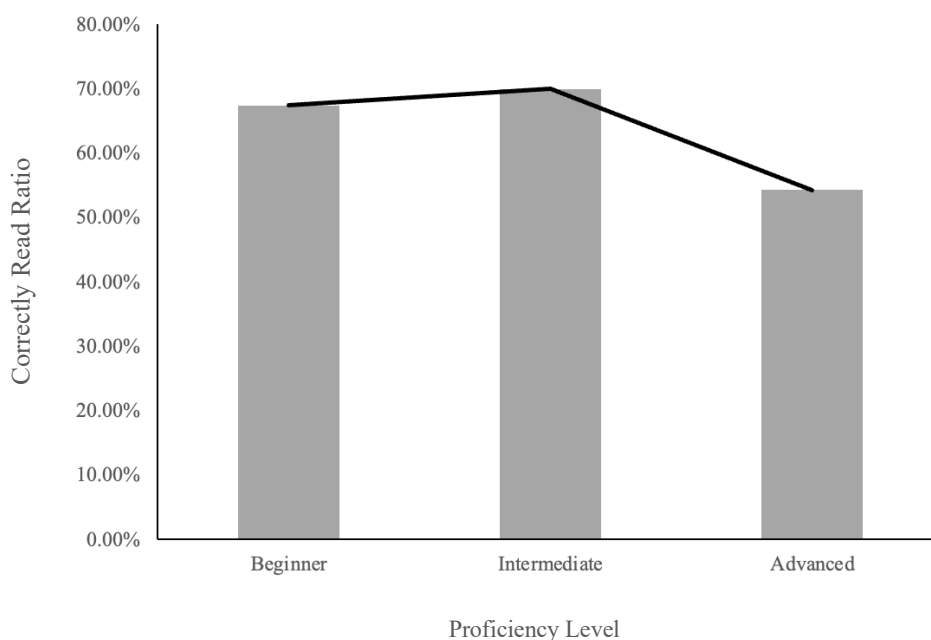
Table 7.19

Number of correctly read incorrectly written characters across proficiency levels

Proficiency level	No. of correctly read incorrectly written characters	Total number of incorrectly written characters
Beginner	60	89
Intermediate	207	296
Advanced	115	212

Figure 7.21

Ratio of incorrectly written characters correctly read across proficiency levels



Overall, although learners at all levels show instances of “incorrectly written but correctly read” characters, the accuracy ratio exhibits a non-linear pattern—rising from the beginner to the intermediate stage and then declining at the advanced level.

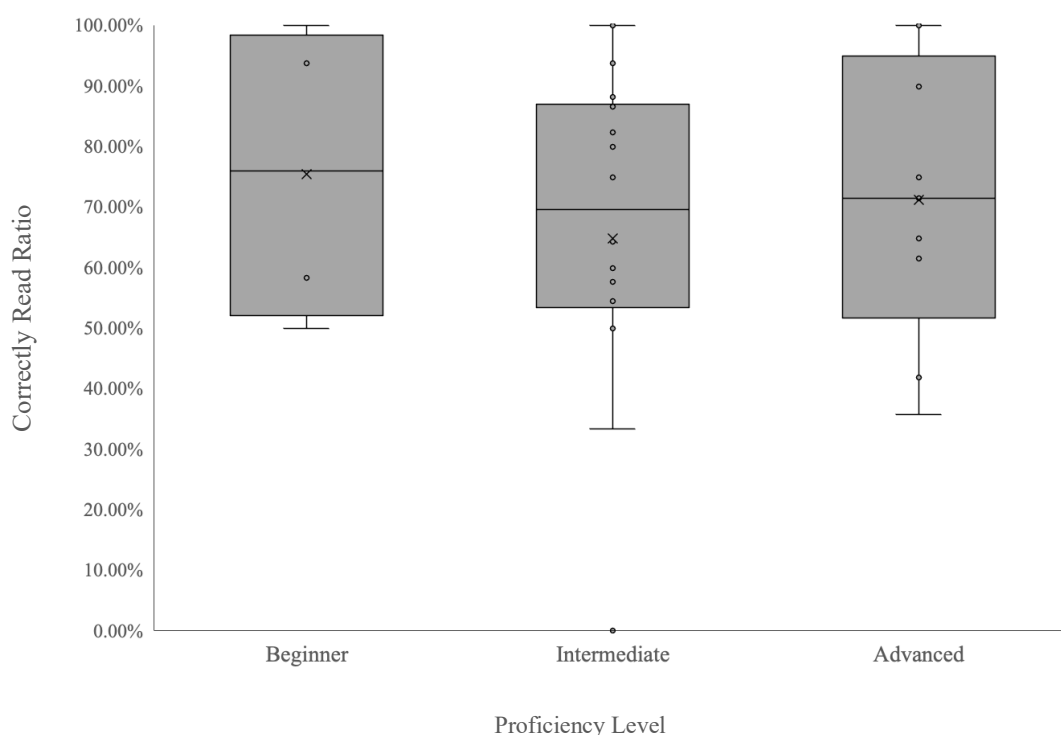
Specifically, the beginner group produced a total of 89 incorrectly written characters, of which 60 were pronounced correctly during reading-aloud-aloud, yielding a correct-reading ratio of 67.42%. The intermediate group produced 296 incorrectly written characters, 207 of which were read correctly (69.93%). In contrast, the advanced group produced 212 incorrectly written characters, with only 115 correctly read (54.25%). In other words, learners at the

intermediate level achieved the highest rate of cross-modal consistency, whereas the advanced group showed a notable decline.

As shown in Figure 7.22, the three proficiency groups display distinct patterns of intra-group variability in the correct-reading ratios of incorrectly written characters. Overall, the results reveal that the beginner group shows relatively high and consistent performance, the intermediate group exhibits the largest degree of dispersion, and the advanced group becomes more stable again though with a slightly lower mean.

Figure 7.22

Boxplot of incorrectly written characters correctly read across proficiency levels



Specifically, in the beginner group, individual correct-reading ratios range from 50% to 100%, with three of the four learners exceeding 90%, indicating strong internal consistency and stability. Because beginners produced fewer written samples and the tasks involved more basic lexical items, they tended to rely on direct phonological recall to access word pronunciations. As a result, “incorrectly written but correctly read” cases were highly frequent among this group.

The intermediate group demonstrates the greatest degree of fluctuation, with correct-reading ratios ranging from 0% to 100%. Some learners achieved very high accuracy (above

90%), whereas others showed significant phonological deviations or complete reading-aloud errors (0%–30%). This polarized distribution suggests that the intermediate stage represents a restructuring and instability phase in the orthography–phonology integration process. At this stage, learners begin to recognize the importance of orthographic accuracy and attempt to integrate visual and phonological representations; however, during this integration, some learners experience temporary form–sound conflicts or processing imbalances, leading to marked individual variability. In other words, intermediate learners do not progress synchronously but follow diverse developmental trajectories of integration.

In the advanced group, the correct-reading ratios range from 35% to 100%, with a more compact distribution and a slightly lower mean than that of the intermediate group. The reduced variability indicates that most advanced learners, after experiencing mid-level fluctuations, have gradually established more stable form–sound correspondences. Nevertheless, a few advanced learners still remain in the 40%–60% range, suggesting that even at high proficiency levels, some individuals have not achieved full stabilization of their phonological–orthographic system. This may be attributed to localized fossilization or a decline in fine-grained monitoring as pronunciation becomes more automated.

Overall, the distribution across proficiency levels forms a “low dispersion – high dispersion – re-convergence” developmental trajectory, confirming the non-linear pattern of cross-modal consistency in second language learning. In the early stage, learners rely heavily on direct phonological memory for pronunciation; in the intermediate stage, they enter a restructuring period of phonological–orthographic integration; and in the advanced stage, they reach system-level stabilization and partial automatization, though some fossilized deviations may persist.

Learners at all proficiency levels demonstrated a high degree of consistency in reading incorrectly written characters, indicating that even when the orthographic form was inaccurate, they were still able to retrieve the correct pronunciation through phonological memory. This reflects a sound-based memory pattern in Chinese character acquisition.

At the character level, the correct-reading ratio followed a non-linear trend, peaking at

the intermediate stage and slightly decreasing at the advanced level. The intermediate stage thus represents the period of strongest orthography–phonology integration, while the slight decline in advanced learners may be related to increased pronunciation automatization and reduced attention to individual characters.

Within-group distributions further revealed a “low–high–reconverging” trajectory of variability, reflecting the restructuring and rebalancing process of cross-modal mapping during development. In sum, learners’ orthography–phonology integration evolves dynamically—from early reliance on phonological memory to the gradual establishment of coordinated sound–form connections

7.4.2 Incorrectly used character reading-aloud

This section analyzes learners’ reading-aloud performance of incorrectly used characters, focusing specifically on two subtypes identified in the corpus—GSCs (graphically similar character substitutions) and OCs (other substitutions). Here, the analysis examines whether learners at different proficiency levels were able to pronounce the target sound correctly when reading characters they had incorrectly used in writing, thereby revealing the mechanisms of cross-modal activation and phonological compensation in their orthography–phonology processing.

Table 7.20

Number of students who read incorrectly used characters correctly

Proficiency level	No. of correctly read student	No. of total student
Beginner	3	3
Intermediate	14	17
Advanced	7	11

Learners of different proficiency levels also exhibited a generally high degree of consistency in their reading-aloud performance of incorrectly used characters (see Table 7.20). In this context, a “correct reading” refers to instances where a learner was supposed to write character A but mistakenly wrote character B, yet pronounced the sound of A correctly during the

reading-aloud task. Such cases indicate that although learners made orthographic selection errors during writing, they were still able to accurately activate the target phonological representation of character A through the semantic–phonological channel when reading aloud, demonstrating strong phonological retrieval ability and cross-modal compensation mechanisms.

Specifically, in the beginner group, three students produced incorrectly used characters, and all three correctly pronounced at least one target sound during reading-aloud, resulting in a 100% success rate. In the intermediate group, 17 learners made such errors, among whom 14 correctly pronounced at least one target sound (82.35%). In the advanced group, 11 learners produced incorrectly used characters, and seven of them demonstrated at least one instance of “incorrectly written but correctly read,” yielding a success rate of 63.64%. These results show that all three groups exhibited the “incorrectly written–correctly read” phenomenon, but the correct-reading ratio decreased with proficiency, suggesting a growing tendency toward decoupling between the orthographic and phonological channels at higher levels.

As shown in Table 7.21 and Figure 7.23, the correct-reading ratio of incorrectly used characters exhibits a clear developmental pattern across proficiency levels, following a “mid-level peak” trend. Beginners correctly read 10 out of 23 incorrectly used characters (43.48%), intermediates achieved the highest rate with 92 out of 146 (63.01%), and advanced learners read 45 out of 78 correctly (57.69%). This non-linear change indicates that the integration between orthographic and phonological representations reaches its strongest and most flexible point at the intermediate stage.

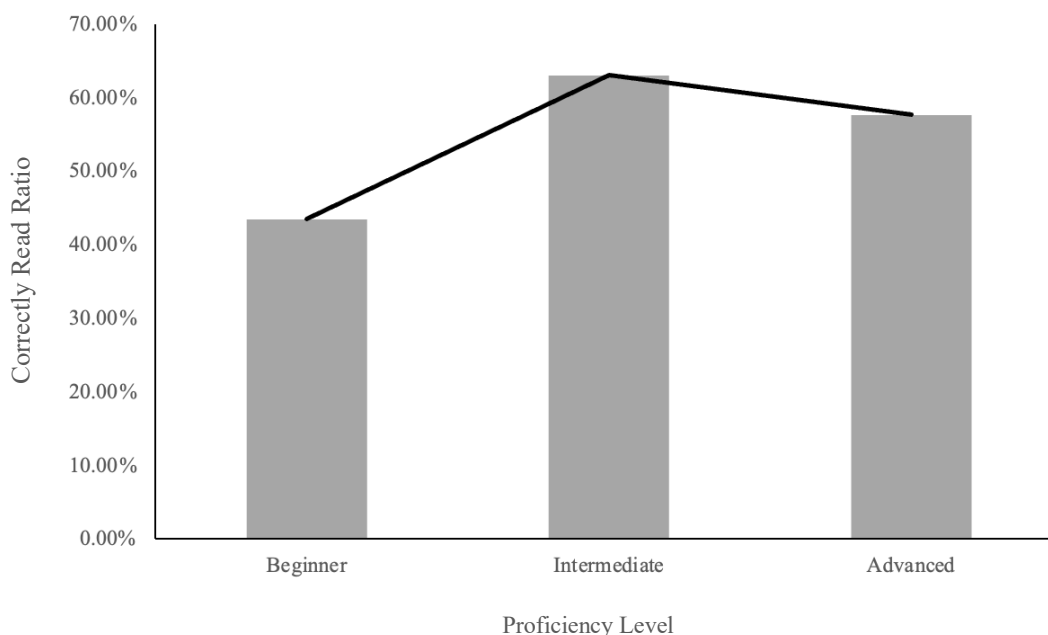
Table 7.21

Number of correctly read incorrectly written characters across proficiency levels

Proficiency level	No. of correctly read incorrectly written characters	Total number of incorrectly written characters
Beginner	10	23
Intermediate	92	146
Advanced	45	78

Figure 7.23

Ratio of incorrectly written character correctly read across proficiency levels



At the beginner stage, incorrectly used characters are largely attributable to learners' limited orthographic knowledge. With a relatively small written vocabulary, beginners rely heavily on phonological memory; thus, even when they select an incorrect character in writing, they are often able to revert to the familiar phonological representation when reading.

The performance of intermediate learners is the most representative: at this stage, their vocabulary expands substantially, and the mapping system between sound and form is being integrated. Consequently, even when semantic competition or orthographic selection difficulty occurs during writing, they can still activate the correct phonological representation in reading. This indicates that the phonological and semantic systems are most complementary at this stage—when visual representations are inaccurate, the phonological channel effectively compensates.

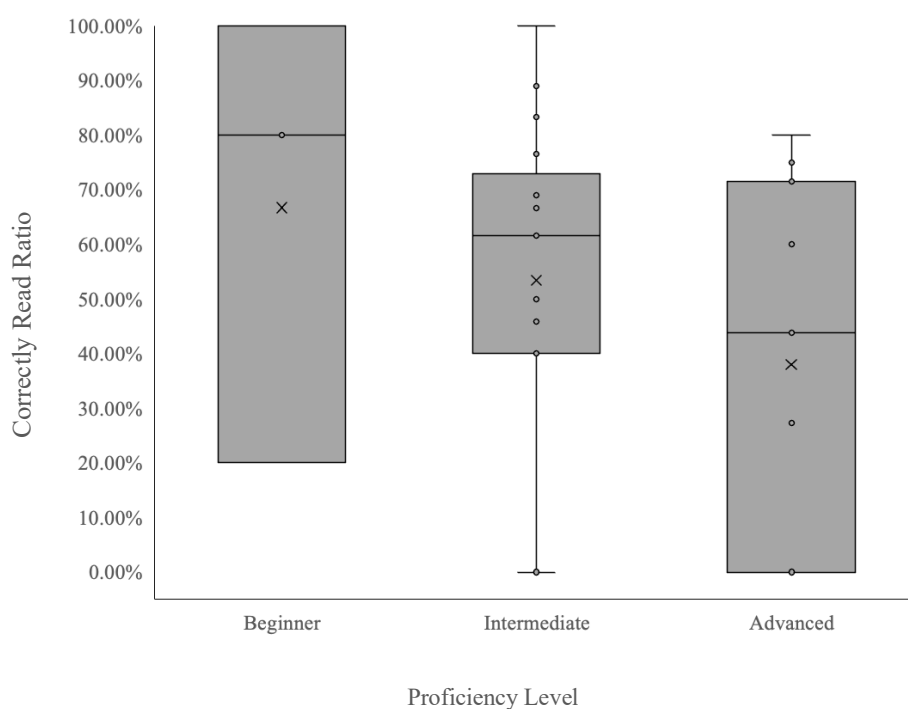
The slight decrease in the correct-reading ratio at the advanced stage (from 63.01% to 57.69%) suggests that with increased proficiency, learners become more reliant on the orthographic system of written Chinese. Semantic processing gradually gains dominance in cross-modal tasks, while the efficiency of phonological activation declines. This semantic-prior

and phonology-secondary processing mode reflects a trend of form–sound decoupling during the automatization of written language cognition.

Overall, learners across all proficiency levels demonstrate that phonological representations take precedence over orthographic ones in the “incorrectly used character reading” task. Even when they select the wrong character, they tend to retrieve and produce the correct pronunciation through the phonological channel. As proficiency increases, the direction of system integration shifts: the beginner stage relies on phonological memory for compensation, the intermediate stage enters a period of unstable phonological–semantic integration, and the advanced stage shows a tendency toward semantic dominance in cross-modal processing. This result suggests that the production of incorrectly used characters mainly stems from lexical interference at the visual–semantic level, while learners’ ability to pronounce the target sound correctly during reading-aloud depends on whether they can rapidly retrieve the phonological representation of the intended word from memory.

Figure 7.24

Boxplot of incorrectly used character correctly read across proficiency levels



As shown in Figure 7.24, the intra-group distribution of correct-reading ratios for incorrectly

used characters reveals stage-specific patterns in learners' mastery of sound-form mapping during Chinese character acquisition. Overall, the results indicate that the beginner group shows relatively high but fluctuating performance, the intermediate group exhibits the greatest instability, and the advanced group becomes more convergent but with a lower mean accuracy.

Specifically, the correct-reading ratios in the beginner group range from 20% to 100%, with a generally high average. Although their orthographic representations were inaccurate, these learners were still able to pronounce the target sounds correctly during reading-aloud. This suggests that their phonological system develops relatively early and serves a cross-modal compensatory function—even when orthographic representations are unstable, learners can retrieve accurate pronunciations through phonological memory.

The intermediate group displays the largest fluctuation (0%–100%) and the widest range of individual variability. This indicates that their sound-form system is undergoing a restructuring and reorganization phase. Some learners are able to activate the correct pronunciation even when the character form is incorrect, demonstrating stronger sound-form integration; others, however, experience temporary instability under the dual burden of increasing orthographic complexity and expanding vocabulary, failing to retrieve the target sound from an erroneous form. This polarized distribution suggests that the intermediate stage constitutes a critical transitional period in the development of sound-form mapping in Chinese character learning.

In the advanced group, the overall correct-reading ratio drops below 80%, and the distribution becomes more concentrated. Most advanced learners, after experiencing mid-level fluctuations, have established a relatively stable orthographic recognition system. However, when character misuse occurs, they are less able to compensate through the phonological channel, resulting in a lower proportion of “incorrectly written but correctly read” cases. This shift indicates that as proficiency increases, learners' attentional focus gradually moves from the phonological level toward orthographic and semantic processing. Consequently, the flexibility of phonological activation declines, forming a form-dominant processing pattern, in which visual and semantic cues take precedence over sound-based retrieval.

Across the three proficiency levels, learners show a clear developmental progression in their Chinese character use, moving from phonology-dominant patterns to increased interaction between phonological and orthographic features, and finally toward more stable character-based writing and reading. At the beginner stage, phonological cues play a central role in supporting written and oral output. At the intermediate stage, fluctuations emerge as learners increasingly rely on characters while still exhibiting interference among similar forms. At the advanced stage, character forms and meanings become more consistently controlled, and pronunciation plays a less prominent role in compensating for orthographic uncertainty.

7.4.3 Summary and discussion of cross-modal comparison

By integrating the reading results of incorrectly written characters (IWC) and incorrectly used characters (IUC) (see Table 7.22 and Figures 7.25–7.26), clear stage-specific patterns emerge in learners' cross-modal performance.

Figure 7.25

Ratio of IWC and IUC correctly read across proficiency levels

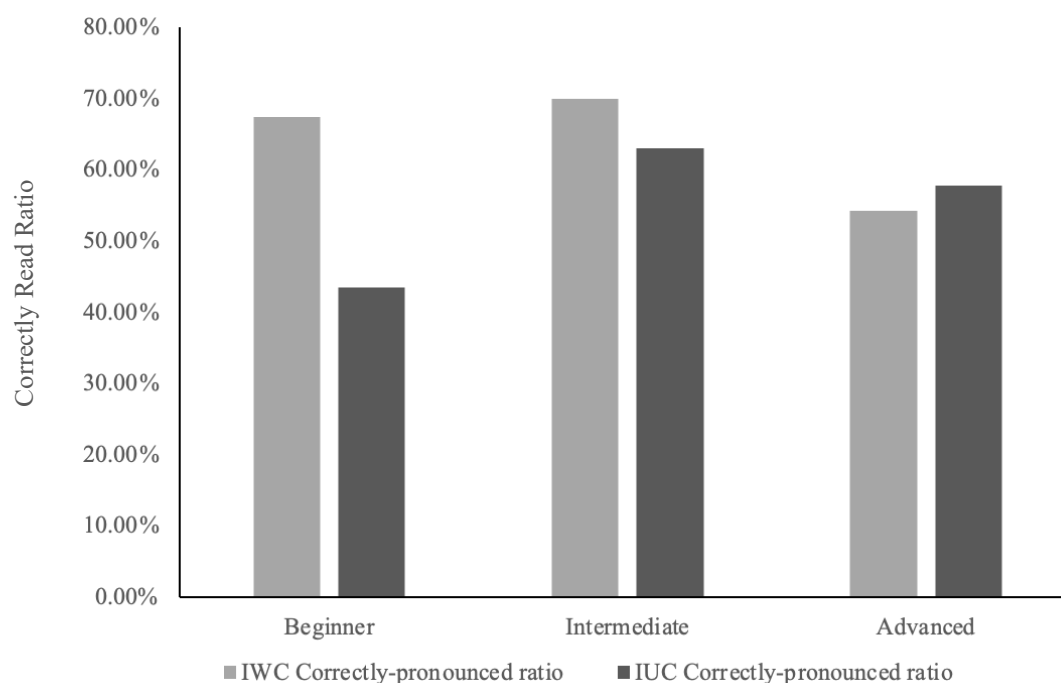
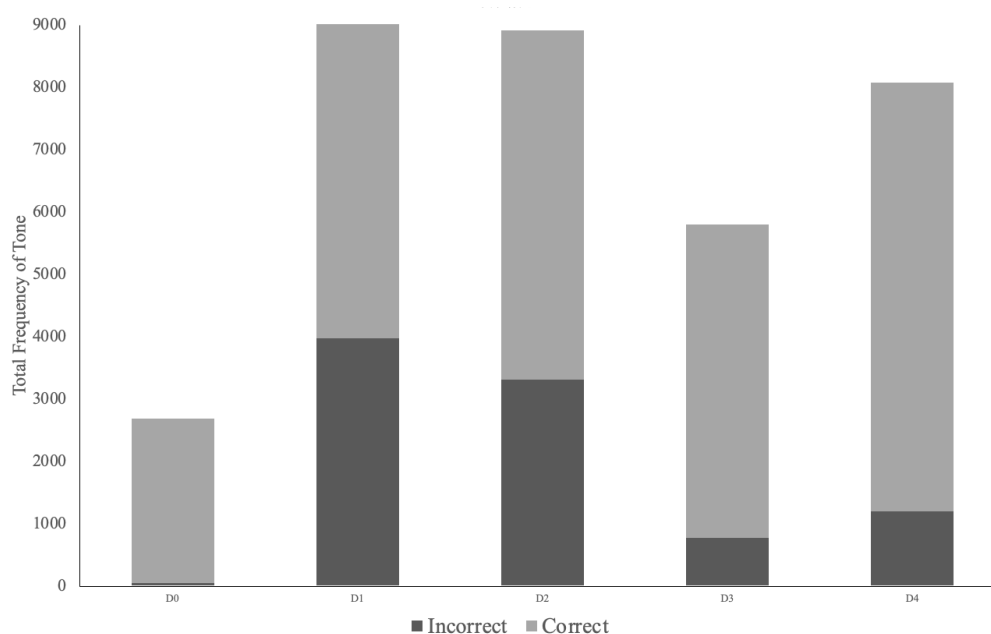


Figure 7.26

Ratio of IWC and IUC reading accuracy across proficiency levels



Across all proficiency levels, the correct-reading ratio of IWCs consistently exceeds that of IUCs, indicating that phonological access is more resilient when errors are confined to orthographic execution than when they involve lexical substitution. However, the magnitude of this difference varies across proficiency levels, revealing a non-linear developmental trajectory in learners' mastery of orthography–phonology mapping and cross-modal compensation.

At the beginner level, the correct-reading ratio for IWCs reaches 67.42% (60/89), substantially higher than that for IUCs (43.48%, 10/23). This pattern suggests that early-stage learners are often able to retrieve target pronunciations even when written forms deviate from the standard. Miswriting at this stage mainly reflects instability in character form, while pronunciation accuracy remains comparatively robust. In contrast, incorrectly used characters involve inappropriate character selection, which limits learners' ability to rely on pronunciation as a compensatory resource during reading-aloud.

At the intermediate level, both correct-reading ratios increase markedly, reaching 69.93% for IWCs and 63.01% for IUCs. This stage represents the highest level of cross-modal consistency observed in the data. Learners at this level appear to have established relatively

stable links between written forms and their corresponding pronunciations, allowing them to maintain reading accuracy even in the presence of writing errors. The intermediate stage therefore constitutes a critical period in which sound–form correspondence is most effectively coordinated.

At the advanced level, the correct-reading ratios decrease to 54.25% for IWCs and 57.69% for IUCs, and the gap between the two error types narrows further. This decline indicates that the stability of sound–form correspondence does not increase linearly with proficiency. Instead, as learners' written output becomes more accurate and complex, correct reading is less frequently maintained when writing errors occur. This suggests a shift in processing preference, in which reading-aloud performance relies increasingly on visually and semantically well-formed input rather than on compensatory pronunciation strategies.

Taken together, the cross-modal results reveal a three-stage developmental pattern in learners' acquisition of sound–form correspondence in Chinese characters. At the beginner stage, learners display a phonology-supported pattern, in which pronunciation accuracy often exceeds orthographic accuracy. At the intermediate stage, sound–form coordination reaches its highest level, resulting in the strongest cross-modal consistency. At the advanced stage, orthographic precision improves further, while cross-modal compensation becomes less prominent. This trajectory demonstrates that the development of sound–form correspondence in L2 Chinese is non-linear, involving phases of reliance, integration, and subsequent stabilization rather than continuous incremental improvement.

From a pedagogical perspective, these findings highlight the importance of maintaining sound–form integration throughout all stages of learning. While intermediate learners benefit most clearly from activities that reinforce the coordination between writing and reading, advanced learners also require continued cross-modal practice to prevent the weakening of this connection. Instruction should therefore incorporate tasks that explicitly link writing and reading, such as reading aloud self-produced texts, correcting and rereading incorrectly written characters, and pairing character forms with pronunciation in integrated practice sequences. Such approaches can support balanced development of reading and writing skills and promote

long-term stability in Chinese character acquisition.

7.5 Summary of cross-proficiency analysis

Drawing on systematic comparisons of writing, reading, and cross-modal tasks across proficiency levels, this chapter delineates the staged developmental trajectory of L1 Italian learners' Chinese character acquisition. Taken together, the three data types reveal a non-linear developmental path: from an early reliance on phonological resources at the beginner stage, through a period of intensive sound–form coordination at the intermediate stage, to a more stable and form-oriented performance profile at the advanced stage.

From the perspective of writing data, orthographic performance develops from instability toward increasing stability. The proportion of incorrectly written characters decreases markedly with proficiency, while within-group dispersion narrows from the highly heterogeneous beginner group to more concentrated patterns at intermediate and advanced levels. This shift reflects learners' growing control over strokes, components, and structural configurations. In contrast, incorrectly used characters peak at the intermediate stage, indicating a temporary fluctuation during the transition toward fuller character-based writing. As learners expand their character inventories and reduce reliance on Pinyin, they encounter increased difficulty differentiating visually or phonologically similar characters. Pinyin use follows a clear “high → low → slight rebound” trajectory, evolving from a dependence-driven support at the beginner stage into a selective compensatory resource used strategically at higher levels.

Reading-aloud data display a hierarchical pattern of development. Tonal errors remain the most persistent difficulty across all proficiency levels, with relatively slow reduction over time. Errors involving initials and finals are rare throughout, whereas compound errors combining segmental and tonal deviations show a temporary rise at the intermediate level before declining. This distribution indicates that intermediate learners experience a transitional phase in which coordination among different phonological elements is less stable, while more advanced learners achieve greater overall consistency despite occasional local fluctuations.

Cross-modal analysis further demonstrates that writing accuracy and reading accuracy do not develop in parallel. Across all proficiency levels, learners are often able to read correctly

characters that they have written incorrectly or used inappropriately, indicating that pronunciation accuracy can be maintained independently of orthographic precision. Cross-modal correctness reaches its highest point at the intermediate stage, where the relationship between written form and pronunciation appears most flexible. At the advanced level, this flexibility decreases slightly, suggesting that accurate reading becomes more closely tied to visually and semantically well-formed input. These results confirm that sound–form correspondence does not strengthen in a linear fashion but instead follows a trajectory of early reliance, mid-stage integration, and later stabilization.

Across writing, reading, and cross-modal perspectives, a consistent stage-based pattern emerges. Beginner learners rely heavily on phonological resources and Pinyin support, producing unstable written forms while maintaining relatively strong reading-aloud performance. Intermediate learners represent a critical transition stage characterized by temporary fluctuations in both writing and reading, alongside the strongest coordination between sound and form. Advanced learners exhibit the highest level of orthographic precision and overall stability, though with reduced reliance on phonological compensation.

Overall, learners progress from phonological reliance to integrated sound–form coordination and finally toward stable character-based performance. Importantly, these findings indicate that development across modalities does not proceed in a uniform or linear manner. Instead, stability emerges through cycles of adjustment and redistribution across writing, reading, and cross-modal tasks.

From a pedagogical perspective, these patterns support the need for continuous, cross-modal instruction rather than strictly stage-segmented teaching. At the beginner stage, instruction should emphasize sound–character conversion and basic stroke and component awareness. At the intermediate stage, teaching should focus on reinforcing coordination between writing and reading through multimodal tasks such as reading-based error correction, sound–form pairing, and contrastive character practice. At the advanced stage, instruction should aim to maintain balance between accuracy and fluency by integrating reading, writing, and extended discourse tasks, thereby preventing the emergence of imbalanced performance

profiles.

In sum, the cross-proficiency analysis clarifies stage-specific regularities in the development of Chinese character skills among L1 Italian learners and underscores the dynamic nature of interlanguage development. Writing, reading, and cross-modal abilities do not advance in lockstep; rather, they evolve through successive phases of adjustment and stabilization. The MICICL corpus provides robust empirical support for this account and yields concrete, stage-sensitive implications for Chinese character instruction.

Chapter 8. Conclusions and Future Work

8.1 Synthesis of findings from MICICL

8.1.1 MICICL as a pedagogically oriented interlanguage corpus

As a methodological endeavor, the value of MICICL does not lie in proposing a single or definitive model of corpus construction, but rather in offering a realistic and operational reference path for future design and application of pedagogically oriented Chinese interlanguage corpora. As discussed in Chapter 2, despite the continuous growth of Chinese interlanguage corpora in both scale and number, their pedagogical orientation remains relatively underdeveloped. Existing corpora have often been driven primarily by researchers' analytical needs, while the process-related information embedded in learners' authentic production and its pedagogical interpretability have received comparatively less attention. To some extent, this orientation has limited the transfer of corpus-based findings into teaching practice, leaving a gap between interlanguage corpora and classroom instruction that requires additional interpretation and reconstruction. MICICL was constructed as a direct response to this reflection, pedagogical considerations were integrated into the corpus design and construction process itself, instead of being limited to later applications.

First, MICICL places writing and reading aloud at the center of its design and aligns them within a single corpus framework. This choice is motivated not by a technical pursuit of multimodality, but by a long-standing issue in Chinese teaching: learners often show uneven development in Chinese writing, character recognition, and reading aloud, even though these abilities are frequently assessed together in instructional contexts. By comparing writing and reading-aloud data, MICICL makes it possible to distinguish different aspects of learners' character knowledge from a pedagogical perspective and to reinterpret error patterns that cannot be explained by a single modality alone. At the same time, MICICL shifts the focus from describing final error outcomes to understanding the learning process itself, it enables simultaneous observation of multiple facets of learner performance on the same task, thereby supporting the analysis of why particular errors emerge at specific developmental stages—

evidence that is essential for pedagogical diagnosis and instructional decision-making.

Second, MICICL adopts a multimedia structure that systematically aligns images of handwritten production, audio of reading-aloud activities, textual transcriptions, and annotations in order to preserve learners' authentic production rather than only normalized textual outcomes. Handwritten images capture fine-grained visual details such as strokes and character shapes, while reading-aloud audio retains dynamic phonetic features such as tone realization and pauses, allowing learner performance to be examined in a more nuanced way than a simple correct–incorrect distinction. This design is closely related to the characteristics of Chinese language learning, as many difficulties in Chinese character acquisition cannot be adequately represented through text alone, and reading-aloud data provides crucial phonological evidence for examining the sources of writing errors. By enabling systematic comparison between writing and reading-aloud within a single corpus framework, MICICL addresses the limitation of many existing Chinese interlanguage corpora that treat different modalities in isolation, and it also helps connect corpus-based findings more directly with teachers' classroom experience, thereby enhancing their pedagogical interpretability.

Third, with regard to data collection, MICICL adopts a task-based, one-on-one collection mode and deliberately avoids placing data collection in examination or high-pressure assessment contexts. This choice reflects an important conceptual shift in pedagogically oriented corpus construction: prioritizing learners' language behavior in relatively natural, low-anxiety environments over test-oriented performance under evaluative pressure. From a pedagogical perspective, such data are better suited for analyzing learning bottlenecks and strategy use, rather than for simple proficiency classification. The MICICL data collection design thus offers a practical reference for how pedagogically oriented interlanguage corpora may strike a balance between authenticity and controllability.

Fourth, in its temporal structure, MICICL introduces both longitudinal and quasi-longitudinal dimensions by repeatedly collecting data from learners at different points in time. Although this design does not aim for strict experimental control in the sense of fully longitudinal studies, its pedagogical significance lies in enabling a developmental perspective

on learners' interlanguage states, rather than restricting analysis to static descriptions. For pedagogically oriented corpora, this distinction is particularly important: teaching practice is concerned not only with where learners currently make errors, but also with which problems tend to resolve naturally over time, and which are likely to persist and require targeted intervention. As shown in the analyses presented in Chapters 5–7, this temporal dimension plays a crucial role in identifying developmental trajectories in Chinese character learning and shifts in pedagogical priorities, highlighting the structural advantages of pedagogically oriented corpus design.

Fifth, with respect to transcription principles, MICICL places particular emphasis on the pedagogical significance of Pinyin-based transcription in reading-aloud data. Rather than directly normalizing learners' spoken output to target-like forms, reading-aloud data are first transcribed faithfully in Pinyin to reflect learners' actual pronunciation, with deviations from standard forms indicated at subsequent annotation levels. This “record first, analyze later” approach reflects a learner-centered stance toward interlanguage: the corpus prioritizes learners' real production over presupposed target norms. From a pedagogical perspective, this transcription closely mirrors teachers' common classroom practices of listening, noting, and correcting, thereby making the corpus data more compatible with teaching realities and lowering the expertise threshold required for phonological analysis.

Sixth, in annotation design, MICICL adopts pedagogically identifiable, interpretable, and actionable error types as its core organizing principle. Error categories in the written modality—such as incorrectly written characters, incorrectly used characters, and pinyin use—as well as tone and segmental errors in the reading-aloud modality, directly correspond to central difficulties in Chinese language teaching. Rather than pursuing maximal formal granularity, the annotation scheme emphasizes the correspondence between annotation categories and pedagogical concerns. By incorporating a pedagogical perspective at the corpus construction stage, MICICL reduces the cost of translating research analysis into teaching applications, making the corpus potentially usable not only for academic research, but also for teacher training, instructional material development, and diagnostic analysis.

Overall, from a methodological perspective, the pedagogical orientation of MICICL does not reside in any single design decision, but emerges from a constellation of interconnected choices. Multimodal data render learning processes visible; the task-based data collection enhances ecological validity; the quasi-longitudinal structure supports developmental analysis; and interpretable transcription and annotation provide a bridge to pedagogical applications. This integrated design enables MICICL to support systematic discussion of character-level interlanguage patterns, developmental trajectories, and teaching implications in subsequent chapters, rather than serving merely as a data repository for academic research. In this sense, MICICL is intended not as a closed or exclusive solution, but as a practical example that may inform the development of pedagogically oriented Chinese interlanguage corpora, and it provides a solid foundation for the subsequent discussion of teaching implications and assessment applications.

8.1.2 Character-level interlanguage patterns

Chapters 5–7 examine character-level Chinese interlanguage from four complementary perspectives: writing, reading aloud, cross-modal interaction, and cross-proficiency comparison. Taken together, the results show that the errors produced by L1 Italian learners are not random slips, but systematic patterns shaped by the visual-structural complexity of Chinese characters, the opacity of form–sound correspondences, and processing strategies rooted in learners’ experience with alphabetic writing systems. These constraints surface differently across modalities and proficiency stages, leading to a development that is coordinated but not synchronized: while some representational dimensions stabilize relatively early, others remain persistently challenging. Importantly, difficulties in writing do not necessarily entail breakdowns in reading aloud, pointing to partial independence and compensatory relations between processing channels.

In the handwriting modality, character-level interlanguage is characterized not only by incorrectly written characters, but also by systematic reliance on Pinyin as a supportive resource. In MICICL, 607 instances of incorrectly written characters were annotated, showing clear clustering across learners and character types. Structural analysis reveals a strongly

hierarchical pattern: component-level errors account for 67.71% of all incorrectly written characters, far exceeding stroke-level (17.63%) and structural errors (7.08%). This distribution indicates that learners rely on components as intermediate units for character construction, rather than copying strokes mechanically or relying solely on holistic visual memory. At the same time, pinyin use constitutes a central feature of writing behavior. Pinyin appears consistently across proficiency levels, often in cases where learners are unable to retrieve orthographic forms but retain access to pronunciation. This pattern reflects learners' strong reliance on phonological representations and shows that writing difficulty does not necessarily stem from loss of character knowledge, but from bottlenecks in orthographic retrieval. Taken together, incorrectly written characters and pinyin use point to a writing system in which phonological representations are relatively stable, while orthographic representations—especially at the component-combination level—remain under consolidation.

Reading-aloud data provide complementary evidence for the phonological dimension of character-level interlanguage. The MICICL reading-aloud subcorpus contains 35,053 Chinese syllables, among which 9,309 tonal errors were identified, accounting for 25.27% of all syllables. Tonal errors occur in almost all recordings and are produced by all learners included in the analysis, indicating a highly systematic pattern. In contrast, errors involving initials and finals are comparatively rare, showing a clear internal differentiation within the phonological system. This asymmetry reflects the influence of Italian as a language with highly transparent orthography: segmental structure aligns closely with pinyin and learners' native phonological experience and therefore stabilizes relatively early, whereas tone, as a suprasegmental feature absent from Italian, remains a long-term bottleneck. As a result, learners' phonological system is globally functional but locally fragile, with tone representing the most persistent source of instability across proficiency levels.

Cross-modal analyses further clarify the relationship between writing and reading aloud. The results show that many characters that are written incorrectly are nevertheless pronounced correctly in reading aloud. Correct pronunciation rates reach 67.29% for stroke-related writing errors, 64.48% for component-related errors, 65.12% for structural errors, and remain as high

as 58.70% even for cases involving multiple writing errors. This pattern indicates that phonological representations at the character level are often already established and accessible, while orthographic representations lag behind. When reading-aloud errors do occur in these cases, they are overwhelmingly tonal (87.44%), whereas segmental errors are extremely rare. This suggests that during cross-modal access, segmental structure remains stable, while tone is more difficult to bind firmly to character–syllable representations. Overall, writing and reading aloud do not stand in a relation of simple correspondence, but rather stabilize at different rates, with the more stable phonological channel partially compensating for orthographic instability.

Cross-proficiency comparisons show that these patterns are developmentally staged rather than static. At lower proficiency levels, learners show high orthographic instability in writing and strong reliance on supportive strategies, including pinyin. As proficiency increases, the overall rate of writing errors decreases sharply and then plateaus, while pinyin use drops in frequency but does not disappear. At the same time, writing errors become increasingly concentrated at the component level, indicating ongoing restructuring rather than uniform improvement. Cross-modal consistency peaks at the intermediate stage and then declines at the advanced stage, suggesting partial decoupling between writing and reading as tasks become more automated and cognitively demanding.

At the system level, these findings indicate that character-level interlanguage is best understood as a composite system consisting of partially independent subsystems—orthographic, phonological, and semantic—each following its own developmental trajectory. When one subsystem is unstable, others may temporarily compensate to maintain functional performance. Overall, the character-level interlanguage patterns observed in MICICL are systematic, developmentally structured, and strongly shaped by typological differences between Italian and Chinese, rather than by random errors or individual deficiency.

8.2 Pedagogical implications of MICICL findings

8.2.1 Teaching implications for Chinese characters

The findings from MICICL demonstrate that the character-level interlanguage of NL Italian

learners displays a high degree of systematicity. Error types do not occur as random collections, but consistently correspond to different processing stages, representational channels, and strategic choices. This observation poses a direct challenge to the result-oriented perspective that is still prevalent in Chinese character instruction. Under a result-oriented approach, learner performance is often reduced to surface outcomes such as “correct vs. incorrect writing” or “correct vs. incorrect reading aloud,” which then serve as the sole basis for feedback and correction. However, the evidence from MICICL suggests that judging learner ability solely on the basis of outcomes easily conflates problems of different nature, leading to feedback that is overly coarse-grained and, in some cases, to misdiagnoses of learners’ actual difficulties. More importantly, such misdiagnoses may shape classroom practice in unproductive ways: instruction may continue to expand character inventories within a whole-character memorization through texts framework, while insufficient attention is paid to the specific levels at which learners are systematically constrained—namely component-level rules, form–sound integration, and tonal automatization.

First, the handwriting data show clearly that incorrectly written characters, incorrectly used characters, and pinyin use do not constitute the same type of “error”. Rather, they index different vulnerable points within the character-processing system. Incorrectly written characters primarily reflect instability in visual–orthographic processing, involving strokes, components, and structural organization. Incorrectly used characters are more closely associated with lexical retrieval and selection, often triggered by competition among homophonous, near-homophonous, or visually similar candidates. Pinyin use, by contrast, typically functions as a compensatory strategy under conditions of orthographic retrieval difficulty, allowing learners to maintain production continuity and reduce immediate cognitive load. From a pedagogical perspective, the key implication is therefore not to correct more strictly, but to differentiate correction. In practical terms, incorrect writing calls for instructional support targeting structural awareness and constructional rules (e.g. component functions, positional constraints, and combinatory patterns); incorrect character choice calls for increased discriminability between target characters and competitors through contrastive work

on form, sound, and context; and pinyin use should be treated not as a pedagogical failure, but as a transitional placeholder whose pedagogical relevance lies in designing a clear exit path rather than enforcing outright prohibition.

At this point, the observations made in Chapter 1 regarding Chinese textbooks used in the Italian context can be meaningfully aligned with the MICICL evidence. Most widely used materials adopt a 随文识字(*suíwén shízi*, ‘characters learned through texts’) approach, in which new characters are introduced incrementally along with vocabulary in reading-aloud passages. While this approach is efficient for lexical progression, it often provides limited systematic support for character structure. The persistent centrality of component-level errors in MICICL suggests that for L1 Italian learners, instruction that focuses primarily on whole characters and lexical accumulation—without establishing an early and transferable framework of component-based knowledge—risks leaving learners repeatedly blocked by structurally complex or rule-opaque characters. In response, a better pedagogically aligned adjustment would be to retain the advantages of text-based character introduction while foregrounding component awareness more explicitly and systematically. Components should not be treated as isolated mnemonic tricks, but as the intermediate processing units learners actually rely on. Grouping characters by shared components, training sensitivity to spatial constraints (e.g. left–right, top–bottom, enclosure structures), and highlighting functional regularities can help learners develop predictive expectations about character structure, thereby reducing cognitive load in subsequent learning.

Second, the cross-modal results indicate that character competence cannot be equated with a single, unified ability. Rather, it emerges from several relatively independent yet interacting subsystems. MICICL shows that errors in writing do not necessarily imply an inability to pronounce characters correctly, and that tonal errors in reading aloud do not necessarily co-occur with orthographic instability. This finding calls into question an often implicit assumption in language teaching that writing and reading abilities develop in parallel, such that learners who can write are expected to read, and vice versa. From a pedagogical standpoint, a more realistic implication is that instruction and assessment should explicitly

differentiate at least three dimensions: orthographic competence (visual structure and constructional rules), phonological competence (especially tonal categorization and production control), and form-sound integration (the ability to map characters to pronunciation rapidly and reliably in task contexts). When these dimensions are conflated, teachers may misinterpret learner performance—for example, attributing writing errors with correct reading-aloud to carelessness, or correct writing with incorrect reading-aloud to weak foundations—while overlooking the actual locus of difficulty. The cross-modal evidence from MICICL thus supports a more process-oriented and diagnostically sensitive approach to classroom evaluation.

Third, MICICL highlights a strong tendency among L1 Italian learners to rely on phonological cues as anchoring points in character learning. This tendency has direct implications for classroom organization. If learners naturally gravitate toward phonological scaffolding (pinyin, segmental skeletons), instruction must actively counterbalance this inclination by reinforcing attention to visual form. In practical terms, this means placing greater emphasis on board writing, structural decomposition, and visual comparison of characters, rather than relying heavily on pinyin-based input methods or screen-based presentations. Especially at beginner and lower-intermediate levels, frequent use of pinyin input as the primary mode of presentation risks reinforcing pinyin as the default processing channel, with characters relegated to a secondary role. Correspondingly, encouraging handwritten production over pinyin typing in classroom tasks is not a return to rote copying, but a way of ensuring that orthographic retrieval and structural organization become obligatory stages of processing. For L1 Italian learners, such “forced activation” of the visual–orthographic channel serves as a necessary counterweight to phoneme-driven processing habits rooted in their L1 experience.

Furthermore, the persistence of tonal difficulty revealed by the reading-aloud data points to the need for more task-oriented and contrastive approaches to tone instruction. Since tonal errors remain salient across proficiency levels, tone training cannot be confined to isolated syllable imitation or short-term corrective drills. Instead, it should involve contrastive perception and production tasks (e.g. minimal and near-minimal tone pairs) and be embedded within meaningful activities such as reading aloud, short recitations, or rhythmically structured

phrases and sentences. Crucially, tonal instability tends to surface as task complexity increases, particularly when learners are required to coordinate multiple processing demands simultaneously; instructional design should therefore deliberately simulate such conditions, requiring learners to manage character recognition, segmental retrieval, tonal control, and fluency at the same time, rather than aiming for correctness under artificially low processing load.

Finally, the cross-modal design of MICICL suggests a highly practical classroom exercise framework: cross-modal integration tasks centered on “read–pause–write” sequences. In such tasks, learners first access characters through the reading-aloud channel (teacher modeling, choral reading-aloud, or silent reading), then briefly suspend phonological support and produce the target characters in handwriting. This format capitalizes on the relative availability of phonological representations among L1 Italian learners while compelling genuine activation of orthographic structure and component rules at the output stage. Compared with traditional dictation, read–pause–write tasks provide finer-grained diagnostic value: correct reading aloud but failed writing points to orthographic retrieval problems; correct writing but incorrect reading-aloud suggests tonal or form–sound binding difficulties; instability in both indicates the need to return to more fundamental component and tone contrasts. In this sense, such tasks function simultaneously as training and classroom-level diagnostic tools.

In sum, the findings from MICICL call for a shift in Chinese character instruction from a narrow focus on outcome correctness toward a process-oriented pedagogical framework. This entails introducing component-based instruction more systematically alongside text-based character learning, designing explicit exit paths for pinyin rather than suppressing it indiscriminately, differentiating ability dimensions in teaching and assessment to avoid single-standard judgments, advancing tonal automatization through contrastive and task-based practice, and exploiting cross-modal exercises to promote form–sound integration. Such adjustments align more closely with the actual dynamics of learners’ interlanguage systems and provide concrete, classroom-implementable pathways for diagnostic and formative assessment grounded in MICICL.

8.2.2 Teaching priorities across proficiency levels

As argued in Section 8.2.1, evidence from MICICL indicates that Chinese character instruction needs to move away from a single outcome-oriented perspective toward a process-oriented teaching framework that distinguishes different ability dimensions in classroom training and assessment. Within this overarching framework, a more specific and pedagogically grounded question arises: as learners' proficiency gradually increases, which aspects of instruction should be prioritized at each stage in order to alleviate the most critical processing bottlenecks at that point in development?

Cross-proficiency comparisons and quasi-longitudinal evidence from MICICL clearly show that difficulties in Chinese character learning do not disappear linearly as proficiency increases. Instead, they follow a developmental trajectory characterized by the relocation of problems across processing levels and modalities. Consequently, teaching priorities should not simply track textbook progression or lexical coverage. Rather, they should be dynamically adjusted according to those processing components that are both the most unstable and the most developmentally malleable at a given stage.

At the initial stage of learning, the primary instructional goal is not full orthographic normalization, but the establishment of a character processing system that can operate sustainably over time. As discussed in Section 8.2.1, beginner learners rely heavily on phonological cues and commonly resort to pinyin or component-based decomposition to cope with the high cognitive demands of writing. If instruction at this stage places excessive emphasis on whether whole characters are written “perfectly correctly”, it is unlikely to improve orthographic mastery and may instead suppress learners' willingness to produce written output. Accordingly, the priority in the writing dimension should be to help learners develop basic structural awareness—understanding component boundaries, spatial positioning, and common structural patterns—rather than enforcing strict whole-character accuracy. In practice, this can be achieved through board-based character decomposition, structural contrasts, and clustered presentation of characters sharing common components, enabling learners to approach character forms “through structure.”

In the reading-aloud dimension, instruction should aim to maintain the usability of the segmental channel so that reading-aloud can function as a stable anchor for later cross-modal training, rather than attempting to eliminate tonal errors prematurely. At the cross-modal level, low-load integration tasks are particularly appropriate, such as reading-aloud first and then writing after a brief pause, or reading-aloud followed by short delayed writing, which use phonology to initiate orthographic retrieval. Pinyin use at this stage should be handled developmentally: it can be allowed as a temporary placeholder in specific contexts, while its functional space is gradually reduced through structural training and consolidation of high-frequency characters. The key indicator of success at this stage is not a rapid reduction in errors, but the emergence of a set of sustainable character learning strategies.

As proficiency increases, instructional priorities need to shift from basic system usability toward integration across processing channels. At this stage, learners' vocabulary typically expands substantially, and both writing and reading-aloud appear more fluent on the surface. However, MICICL shows that this is precisely the stage at which incorrect character choice, confusion among visually similar characters, and tonal instability become most salient. If instruction continues to rely primarily on repetition and single-item correction strategies typical of the beginner stage, it is unlikely to address the core difficulties.

In the writing dimension, the instructional focus should move from "can the learner write this character" to "can the learner write the correct character under competitive conditions." This can be operationalized through systematic comparison of visually similar or phonologically similar characters, strengthening the joint use of structural, phonological, and semantic cues for discrimination. In the reading-aloud dimension, tonal training should shift away from isolated syllable practice toward task-based activities that require tone control in words, short phrases, or continuous speech. At the cross-modal level, instruction should deliberately design tasks that require coordinated activation of multiple channels, such as "read-pause-write-read again" or timed writing following oral reading-aloud, forcing learners to integrate form and sound under realistic processing conditions. Instructional effectiveness at this stage is better reflected in changes in error types rather than in a rapid decline in overall

error counts.

At higher proficiency levels, teaching priorities should further shift from integration toward stabilization and automatization. Learners at this stage typically demonstrate high overall functional adequacy and can handle relatively complex writing and reading-aloud tasks. Nevertheless, under conditions of increased speed, higher cognitive load, or low-frequency character use, localized bottlenecks may still surface. According to the findings from MICICL, such phenomena do not indicate regression, but rather reflect normal system behavior under pressure. Consequently, advanced instruction should move away from broad coverage and instead adopt diagnostic, problem-oriented prioritization.

In writing, this may involve targeting recurring problems associated with specific character groups, structural patterns, or task conditions. In reading-aloud, the instructional focus for tone should no longer be whether learners know tones, but whether they can maintain stable tonal control in authentic tasks. At the cross-modal level, integration activities that closely approximate real-world use—such as immediate writing following continuous reading-aloud—can be used to test and reinforce system stability. The central goal at this stage is to help learners sustain coordinated multi-channel processing while gradually reducing reliance on strategic compensatory mechanisms.

Overall, within the process-oriented teaching framework proposed in Section 8.2.1, evidence from MICICL supports a principle of dynamically adjusted instructional priorities across proficiency development: an early focus on structural awareness and system usability, a mid-stage focus on form–sound integration and competition management, and an advanced focus on diagnostic fine-tuning and automatization. These priorities are not intended to replace textbook content, but to provide instructors with a higher-level reference for deciding where instructional effort is most productively invested at a given moment. By continuously aligning teaching focus with the most fragile yet developmentally promising processing levels, Chinese character instruction can better match the actual operating logic of the interlanguage system, rather than remaining confined to the pursuit of surface-level correctness.

8.2.3 Using MICICL for diagnostic and formative assessment

Building on the process-oriented teaching perspective and the proficiency-sensitive priority framework outlined in the previous two sections, the value of MICICL lies not only in its theoretical contribution to understanding character-level interlanguage, but also in its direct applicability to classroom assessment and feedback. Unlike traditional assessment practices that rely primarily on summative testing and aim to produce a final judgment of learners' performance, MICICL is particularly well-suited to diagnostic and formative assessment, whose purpose is to help teachers identify learners' current processing bottlenecks and continuously adjust instruction during the learning process.

Drawing on the empirical findings in Chapters 5–7, the pedagogical use of MICICL for assessment can be concretely articulated along three closely related dimensions: (a) the use of frequency-based statistical results, (b) the use of multimodal data, and (c) the use of corpus annotation information. Together, these dimensions support an assessment framework that is both data-driven and directly actionable in classroom contexts.

First, at the level of statistical results, the high-frequency lists derived from MICICL—such as frequently incorrectly written characters, frequently misused characters, characters with frequent pinyin use, and characters showing recurrent reading-aloud errors—provide an efficient diagnostic tool for teachers of Chinese, particularly those working with L1 Italian learners. While the specific items in these lists are corpus-dependent, the underlying procedure—using frequency-based error profiles to prioritise instruction—is transferable to other learner populations, provided that comparable learner corpora are available. The analyses in Chapters 5–7 show that learners' difficulties are not evenly distributed across all characters, but are instead highly concentrated in a limited set of structurally complex, visually similar, or phonologically similar characters. From a pedagogical perspective, this means that teachers do not need to rely on large-scale tests to assess learners' overall proficiency. Instead, by examining recurring errors in everyday classroom output (e.g. homework, short writing tasks, or reading-aloud recordings), teachers can quickly identify the “high-risk character sets” that currently pose the greatest challenge. In practice, these frequency lists can be used to guide the

selection of instructional targets. Rather than mechanically following textbook progression, teachers can prioritize a small number of characters that repeatedly cause difficulty and design focused practice around them. Similarly, in reading-aloud instruction, characters that repeatedly trigger tonal errors across learners can be isolated and addressed through targeted listening and production activities. In this way, assessment becomes embedded in teaching, directly informing instructional decisions rather than functioning as a separate evaluative endpoint.

Second, at the level of multimodal data, MICICL's handwritten image data and reading-aloud audio recordings provide forms of diagnostic evidence that are difficult to obtain through traditional paper-based tests. These data are particularly valuable for understanding why an error occurs, not just whether it occurs.

In the handwriting modality, handwritten images allow both teachers and learners to examine the construction process of characters. When a character is judged as incorrectly written, teachers can guide learners to revisit their own handwritten output and identify whether the problem lies in component boundaries, component placement, or overall structural balance. Even when the final form is incorrect, learners can often recognize whether they have correctly identified key components or whether their production relies on vague visual memory. Such reflective use of handwritten images shifts attention away from binary correctness toward structural awareness and character construction strategies.

In the reading-aloud modality, audio data are particularly effective for diagnosing tonal problems. Teachers can select short excerpts from learners' own recordings and compare them with target pronunciations, asking learners to judge where the problem lies—whether in tonal category selection or in tonal weakening during continuous speech. In the reading-aloud modality, audio data are particularly effective for diagnosing tonal problems. Teachers can select short excerpts from learners' own recordings and compare them with target pronunciations, asking learners to judge where the problem lies—whether in tonal category selection or in tonal weakening during continuous speech. Within the MICICL framework, this can be operationalised by using the aligned audio data and transcription layers to identify

representative learner errors and pairing them with corresponding target forms. Such corpus-based examples can be used either as model cases in class or, where available, in combination with students' own recordings. While teachers may optionally collect audio data from their own students, the MICICL corpus itself already provides a structured resource for this type of diagnostic and comparative activity. This approach also helps teachers distinguish between learners who have not yet established tonal categories and those whose tonal representations exist but are not yet automatized.

More importantly, at the cross-modal level, MICICL's multimodal data can be directly used to design tasks that integrate writing and reading-aloud for both assessment and training purposes. Given the documented developmental asymmetry between orthographic and phonological channels, teachers can deliberately design activities that require learners to switch between modalities. For example:

- learners are shown a character and listen to an audio recording, then judge whether the pronunciation (especially the tone) is correct;
- learners first listen to a spoken form and are then asked to write the corresponding character without pinyin support;
- learners read a character or word aloud, pause briefly with auditory support removed, and then write it by hand.

Such tasks function simultaneously as practice and diagnosis. If a learner can read correctly but cannot write, the bottleneck is likely orthographic retrieval or structural organization; if a learner writes correctly but reads inaccurately, tonal control or phonological automatization is likely the issue. Cross-modal tasks therefore allow teachers to locate processing difficulties more precisely than single-modality assessments.

Third, at the level of corpus annotation information, MICICL's fine-grained annotation system enables assessment to move from simply pointing out errors to explaining their underlying mechanisms. Different annotation categories—such as component errors, structural errors, character misuse, pinyin placeholder use, or tonal errors—correspond to different processing levels within the interlanguage system. In classroom assessment, these categories

can be translated into explanatory feedback: teachers can not only indicate what is wrong, but also clarify what type of processing difficulty the error likely reflects.

For instance, when a learner repeatedly produces component-level errors, teachers can explicitly explain that the issue lies in structural integration rather than memory failure; when misused characters cluster around homophones, learners can be guided to understand the role of lexical competition rather than lack of vocabulary knowledge. Such feedback helps learners recognize the systematic nature of their errors and reduces the tendency to attribute difficulties to carelessness or insufficient effort.

Annotation information also supports longitudinal formative assessment. By tracking whether certain error types decrease after targeted instruction, teachers can evaluate whether an intervention has addressed the underlying bottleneck rather than merely improving short-term accuracy. This makes assessment an ongoing process of observation, adjustment, and refinement.

Taken together, MICICL enables a practical and coherent framework for diagnostic and formative assessment in Chinese character instruction: frequency-based statistics help identify focal problem areas, multimodal data make processing difficulties visible, and annotation information provides an explanatory link between errors and underlying mechanisms. By integrating these resources into writing, reading aloud, and cross-modal classroom activities, assessment becomes an integral part of teaching rather than an external evaluative add-on. In this way, MICICL supports a form of assessment that aligns closely with the actual operation of learners' interlanguage systems and the realities of classroom practice.

8.3 Future research directions and extended uses of MICICL

8.3.1 Lexical and syntactic levels of interlanguage

Rather than introducing new empirical analyses, this section is intended as a methodological note on how MICICL's existing annotation system can be operationalized in future research. By making explicit the analytical entry points associated with different annotation levels, this section aims to guide researchers in the principled reuse of MICICL for investigating

interlanguage phenomena beyond the character level. In this sense, the focus here is not on findings, but on analytical affordances.

The present dissertation deliberately focuses its analysis on the character level, a choice motivated by the research objectives and the specific characteristics of the data. This focus, however, should not be interpreted as a limitation of MICICL's research potential. On the contrary, as demonstrated in Chapter 4, MICICL preserves learners' complete written texts and aligned reading-aloud transcriptions, and it establishes a clear, extensible, and internally coherent annotation system across multiple linguistic levels, including characters, words, phrases, and sentences. This design not only supports the analyses conducted in this study but also provides explicit operational pathways for future investigations of interlanguage phenomena at higher linguistic units.

At the lexical level, MICICL's word-level annotation scheme—most notably the [/CC] (incorrect word) label—can be treated as a primary analytical entry point for examining learners' lexical choice strategies and patterns of lexical error. Future studies may take word-level errors as the basic unit of analysis, first identifying the distribution of error types, high-frequency incorrect words, and indices of lexical diversity across proficiency levels. These results can then be systematically aligned with character-level annotations (such as </C> incorrectly written characters and incorrectly used characters) to examine whether lexical errors tend to occur in words containing characters with higher orthographic difficulty. In this way, word-level error annotations serve not merely as descriptive labels, but as indicators of the interaction between lexical selection and character-level processing constraints in the interlanguage system.

At the phrasal level, MICICL introduces dedicated phrase-level annotations (e.g. {/DPC} phrase collocation error and {/DXC} phrase order error), which function as operational tools for isolating learners' difficulties in lexical combination and structural organization. These annotations allow researchers to extract recurrent collocational patterns—such as classifier–noun mismatches or modifier–head ordering issues—without conflating them with purely lexical or sentential errors. In analytical practice, phrasal annotations can be used to compare

the range, diversity, and target-likeness of phrase-level constructions across proficiency groups, and to explore whether collocational simplification strategies co-occur with instability at the lexical or character level.

At the sentential level, MICICL adopts an annotation framework oriented toward common ill-formed sentence types encountered in Chinese language teaching, including labels such as “/JXC” (intra-sentential word order error), “/JPC” (intra-sentential collocation error), and “/JQ” (incomplete sentence component). These sentence-level labels provide a ready-made unit of analysis for examining learners’ syntactic choices and error distributions. They allow researchers to systematically compare sentence length, structural complexity, and error types across proficiency levels, and to assess whether learners rely on structurally simpler or more formulaic sentence patterns as a strategy for managing the combined demands of character writing and lexical retrieval.

In addition, the multimodal design of MICICL enables cross-modal investigation of lexical and syntactic phenomena. Because written and read-aloud texts are aligned at the content level, lexical and sentential structures that are problematic in writing can be directly compared with their realization in reading-aloud. Such cross-modal alignment allows researchers to distinguish between knowledge-related limitations and performance effects associated with writing-specific processing load, thereby offering a more nuanced methodological basis for interpreting interlanguage development.

8.3.2 Phonological development beyond citation forms

This dissertation uses the reading-aloud data in MICICL to systematically analyze tonal and segmental errors produced by Italian L1 learners at the character level. The analysis mainly focuses on reading-aloud tasks under relatively controlled conditions. Such tasks are useful for clearly identifying pronunciation problems at the level of individual syllables and therefore play an important role in phonological acquisition research. However, from the perspective of language development and teaching practice, phonological ability is not limited to the accurate pronunciation of isolated characters. It also involves the ability to apply tone sandhi rules in connected speech, maintain fluency, and carry out self-correction when problems arise. With

this in mind, this section explains how MICICL's existing reading-aloud annotations can be used to study phonological development beyond citation-form pronunciation.

First, with regard to tones, MICICL includes the `</BDC>` (tone sandhi error) label in its reading-aloud data to mark cases where learners fail to apply tone sandhi correctly in connected speech. Unlike single-character tone errors, tone sandhi errors do not necessarily indicate that learners lack knowledge of basic tone categories. Instead, they are more likely related to difficulties in applying tonal rules consistently during reading-aloud-aloud. Future studies can take `</BDC>` as a starting point to examine whether tone sandhi errors occur more frequently in tasks with faster speech rate, longer sentences, or higher information load. By comparing the performance of the same learners across different task conditions, researchers can better determine whether tone sandhi problems reflect weak rule knowledge or insufficient automaticity during production.

Second, MICICL's reading-aloud transcription system includes clear annotations for pauses, repetitions, and phonological self-correction, such as natural pauses (`/ZT`), non-natural pauses (`/FZT`), repetitions (`{/CF}`, `{/CFH}`), and self-corrections and corrected forms (`{/XZ}`, `{/XZH}`). These annotations provide direct evidence for how learners plan speech, notice problems, and adjust their pronunciation during reading-aloud-aloud. Future research can use these labels to examine where and how often pauses, repetitions, and self-corrections occur, for example whether learners tend to pause at difficult pronunciation points or correct themselves immediately after an error. In particular, the `{/XZ}` and `{/XZH}` labels make it possible to observe the process from an incorrect pronunciation to a corrected one. This allows researchers to examine whether learners are able to notice their own phonological errors in real time and successfully repair them. Such process-based information is difficult to capture through simple accuracy measures alone, but it is important for understanding different stages of phonological development.

Third, the multimodal structure of MICICL makes it possible to compare the development of phonological ability and writing ability within the interlanguage system. As shown in the analysis in Chapter 6, learners often produce correct pronunciations during reading-aloud-

aloud even when their written output contains character errors. This suggests that phonological ability can partly compensate for instability in writing. Future research can further test this pattern in more demanding reading-aloud or semi-controlled speaking tasks. For example, when learners are required to recognize characters quickly, retrieve words, and maintain fluent speech at the same time, their pronunciation may become less stable and start to show difficulties similar to those observed in writing. By comparing tonal, segmental, tone sandhi, pause, repetition, and self-correction patterns in reading-aloud with character errors in writing, researchers can gain a clearer picture of how phonological and orthographic abilities interact at different stages of development.

Overall, MICICL's reading-aloud annotation system supports not only the analysis of citation-form pronunciation but also the study of phonological performance in connected speech. By making systematic use of tone sandhi, pause, repetition, and self-correction annotations, and by combining these with cross-modal comparisons, future research can examine phonological development in a more comprehensive and development-oriented way without changing the basic structure of the corpus. This approach contributes to a better understanding of Chinese phonological acquisition by learners from non-tonal language backgrounds and further highlights MICICL's value as a pedagogically oriented interlanguage corpus.

8.3.3 Cross-linguistic extensions and reuse of MICICL

This dissertation focuses on learners with Italian as their first language in order to identify interlanguage patterns in Chinese character learning under relatively controlled L1 conditions. This design helps reduce variability and improves the internal consistency of the analysis. The systematic design of MICICL in terms of task construction, transcription rules, and annotation schemes means that it is not limited to a single native language background. Instead, MICICL can be viewed not only as a specific corpus project, but also as a reusable and transferable framework for research and teaching.

At the research level, cross-linguistics extension of MICICL is primarily supported by methodological consistency and comparability. The analyses in this dissertation show that

Italian L1 learners display relatively stable patterns of incorrectly written characters, incorrectly used characters, and pinyin use in both writing and reading-aloud. Future studies can extend this design by introducing learners with other L1 backgrounds—such as English or Spanish, which also use alphabetic writing systems but differ in orthographic transparency—while keeping task types, data collection conditions, transcription rules, and annotation schemes unchanged. By comparing the overlap and divergence in high-frequency error types and their distributions across L1 groups, researchers can better distinguish errors that are more likely due to the inherent complexity of the Chinese writing system from those that are more strongly influenced by L1 orthographic or phonological properties.

Building on this approach, MICICL's existing annotation system also allows for a more detailed examination of compensatory strategies used by learners from different L1 backgrounds. The analyses in this dissertation indicate that Italian L1 learners systematically use pinyin as a transitional strategy in writing, a pattern that is closely related to their experience with a highly transparent orthography and long-term reliance on sound–spelling consistency. In cross-linguistic extensions, researchers can directly compare the frequency, developmental stages, and exit points of pinyin use among learners with different alphabetic L1s, and examine whether differences in orthographic transparency lead to systematic differences in reliance on pinyin. In addition, comparisons between alphabetic and non-alphabetic L1 backgrounds can shed light on whether learners adopt different strategies when dealing with uncertainty in Chinese character writing. Because pinyin use has already been consistently annotated in MICICL, such comparisons are methodologically straightforward and require no changes to the existing annotation scheme.

The multimodal structure of MICICL also provides a unified framework for cross-linguistics and cross-modal comparisons. As shown in Chapter 6, Italian L1 learners often produce accurate pronunciations in reading-aloud even when their written output contains character errors, suggesting that the phonological channel can compensate for orthographic instability. Future research can replicate the same analytical procedures with learners from other L1 backgrounds to examine whether this writing–reading-aloud asymmetry is a general

pattern or whether it is influenced by L1 phonological characteristics, such as whether the L1 is tonal or the complexity of its syllable structure. By applying the same tasks and analytical steps across L1 subcorpora, researchers can ensure comparability while extending interlanguage research across language types.

From a research design perspective, cross-linguistic extensions of MICICL should prioritize comparability over corpus size. In practical terms, this means following a principle of minimal comparable design: even with relatively small datasets, reliable comparisons can be made as long as task design, data collection conditions, transcription conventions, and annotation schemes are kept consistent. For pedagogically oriented interlanguage research, this approach often provides more meaningful insights than large but methodologically inconsistent datasets.

Beyond cross-linguistic research, MICICL also has clear potential as a reusable resource for both research and teaching. Its reusability does not depend primarily on corpus size or technical complexity, but on its clear data structure, multi-level annotation design, and close alignment with real learning contexts. By distinguishing between original learner output, transcribed text, and corrected text, and by preserving detailed error information, MICICL allows future researchers to reanalyze the same data from new perspectives without modifying the original dataset. This reanalyzability supports methodological replication, theoretical testing, and dialogue between different research approaches.

In teaching contexts, MICICL's reusability lies in its direct reflection of real learner problems. The data capture learners' language production in non-test situations, which closely resembles what teachers observe in actual classrooms. Based on the lists of high-frequency writing and reading-aloud errors derived from MICICL, teachers can better judge which difficulties are likely related to the Chinese writing system itself and which may be associated with specific L1 backgrounds, such as differences in orthographic transparency. In addition, the multimodal data—including handwritten images and reading-aloud audio—can be used as case material in teacher training or reflective practice, helping teachers understand not only where learners make errors, but also why those errors occur from a processing perspective.

It should be emphasized that the reusability of MICICL also has clear limits. The data sources, task types, and collection contexts are tied to specific instructional settings. Therefore, when MICICL is used for cross-linguistic comparison or pedagogical reference, researchers and teachers must carefully consider its scope and constraints. Only by making these boundaries explicit can MICICL be used responsibly as a research and teaching resource.

Overall, the extension of MICICL to cross-linguistic research and reusable applications does not represent a departure from the goals of this dissertation, but rather a natural continuation of its research framework. By maintaining methodological consistency, explicitly addressing the role of L1 orthographic type, and making full use of MICICL's existing multimodal and annotation design, future studies can adopt MICICL as a concrete and operational framework for investigating Chinese interlanguage development from a cross-linguistic perspective.

8.4 Conclusion

8.4.1 Overall conclusion and main contributions

The primary aim of this dissertation is not simply to describe linguistic error patterns in a specific learner group, but to systematically examine how a pedagogically oriented Chinese interlanguage corpus can be constructed and effectively applied. Through the design, construction, and analytical workflow of MICICL, this study addresses two closely related questions: first, what key components are required in the construction of a pedagogically oriented Chinese interlanguage corpus; and second, how such a corpus can be meaningfully used to support interlanguage research and pedagogical understanding.

To this end, the dissertation first presented a comprehensive construction framework for a pedagogically oriented Chinese interlanguage corpus, as exemplified by MICICL. This framework is grounded in authentic instructional contexts and focuses on two core learner output modalities: writing and reading aloud. By integrating handwritten images, reading-aloud audio recordings, layered transcriptions, and pedagogically interpretable annotation schemes, MICICL preserves rich traces of learners' production processes. In contrast to traditional

corpora that prioritize normalized textual output, MICICL emphasizes the observability of learning processes, enabling the corpus to support not only formal description but also explanations of learner strategies and processing mechanisms.

Building on this foundation, the dissertation uses character-level analysis as an illustrative case to demonstrate how such a corpus can be applied in interlanguage research. Through systematic analyses of writing and reading-aloud data, the study shows how a multimodal, pedagogically oriented corpus can support cross-modal comparison, cross-proficiency analysis, and structured classification of error types. The choice of Chinese characters as the primary unit of analysis does not imply that this level is uniquely important, but rather reflects the fact that characters represent a high-processing-load unit in Chinese learning and are particularly sensitive to systematic difficulties in orthography, form–sound mapping, and strategy use. As such, character-level analysis serves a demonstrative function in this dissertation, illustrating the analytical feasibility and explanatory potential of the MICICL framework.

Through this illustrative application, the study also identifies several interlanguage characteristics of Italian L1 learners in Chinese character learning. In the writing modality, incorrectly written characters, incorrectly used characters, and pinyin use reflect different types of processing difficulty and compensatory strategies. In the reading-aloud modality, tonal errors show notable persistence, while segmental errors are relatively limited. Cross-modal comparison further reveals that writing and reading-aloud abilities do not develop synchronously: learners who produce orthographic errors in writing are often still able to pronounce the same characters accurately in reading-aloud, indicating a compensatory role of the phonological channel under conditions of orthographic instability. Although these empirical findings are not the sole focus of the dissertation, they provide grounded evidence for understanding Italian learners' character learning processes and offer relevant reference points for pedagogical practice.

Overall, the main contribution of this dissertation lies in demonstrating, through the construction and application of MICICL, that a pedagogically oriented Chinese interlanguage corpus is not only feasible in principle, but also methodologically viable and pedagogically

meaningful. By integrating corpus construction, methodological illustration, and pedagogical interpretation, the study presents a corpus-based research pathway that uses learner data as a bridge between interlanguage research and teaching practice. This contribution is reflected not only in the empirical observations reported in the dissertation, but also in the research perspective and methodological framework that it proposes and puts into practice.

8.4.2 Limitations and Future Research Directions

While this study has undertaken a systematic exploration of corpus construction and multimodal analysis, its research design and data sources are inevitably shaped by real instructional contexts and practical research conditions. The following discussion aims to clarify the scope within which the findings should be interpreted.

First, with regard to the research participants, the data used in this study were drawn primarily from adult learners of Chinese at the university level. This group is relatively homogeneous in terms of cognitive development, learning context, and strategy use, which helps reduce certain types of variability in the analysis. At the same time, this focus limits the extent to which the findings may be applicable to other learner populations, such as younger learners or early immersion contexts. Future studies replicating a similar design across different age groups and learning backgrounds would be valuable in assessing the broader applicability of the patterns observed here.

Second, concerning the temporal structure of the data, MICICL adopts a quasi-longitudinal design shaped by actual teaching conditions. Because not all learners participated in every data collection session, it was not possible to conduct strict longitudinal tracking at the individual level. As a result, interpretations of developmental change are based primarily on group-level tendencies rather than on detailed individual trajectories. Future research conducted in more stable instructional settings, with continuous tracking of the same learners, would allow for a more fine-grained analysis of developmental pathways in Chinese character acquisition.

Third, in terms of analytical scope, this study focuses on the character level and does not attempt a systematic account of interlanguage phenomena at the lexical, syntactic, or discourse levels. This focus reflects the research objectives and the design of the corpus, given that

Chinese characters represent a high-load unit in both writing and reading-aloud, particularly in relation to orthographic processing and form–sound mapping. Nevertheless, this scope means that interactions between character-level phenomena and higher-level linguistic organisation remain underexplored. Future research could extend the present analysis by incorporating multi-level approaches to examine how character processing relates to broader aspects of language production.

In addition, the transcription and annotation of multimodal data inevitably involve a degree of human judgment. Although efforts were made to enhance consistency through clearly defined guidelines and repeated verification, the analysis presented here prioritises the identification of general patterns over fine-grained quantitative distinctions. Future work may further strengthen reliability by incorporating semi-automatic annotation tools and by introducing inter-annotator agreement measures.

A further limitation concerns the predominantly descriptive nature of the analysis. Given the relatively small sample size and the case-study character of the dataset, no inferential statistical procedures were applied. The findings should therefore be interpreted as indicative of patterns within this specific learner group rather than as generalisable evidence concerning the broader population of Chinese language learners. In this sense, the study adopts an exploratory, data-driven approach, in which corpus-based observations are used to identify tendencies that may be further examined in future research.

With respect to theoretical grounding, while the interpretation of the data is informed by perspectives from interlanguage and second language acquisition research, the integration of a fully explicit acquisitional framework remains limited. As a result, the pedagogical implications proposed in this study should be understood as exploratory and suggestive rather than as fully theory-driven generalisations. Future research could build on the present findings by embedding them within more clearly specified theoretical models and by testing them against larger and more diverse datasets.

References

- Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. (2014). KoKo: An L1 learner corpus for German. In *Proceedings of the Language Resources and Evaluation Conference (LREC)* (pp. 2414–2421).
- Abuczki, Á., & Ghazaleh, E. B. (2013). An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, 9(1), 86–98.
- Adjemian, C. (1976). On the nature of interlanguage systems. *Language Learning*, 26, 297–320.
- Aggazzotti, C., Andrews, N., & Smith, E. A. (2024). Can authorship attribution models distinguish speakers in speech transcripts?. *Transactions of the Association for Computational Linguistics*, 12, 875–891.
- Afiana, D., Fauziati, E., & Nurkamto, J. (2018). Permeability of students' interlanguage: A case study on Indonesian students learning English as a foreign language. *International Journal of Scientific and Research Publications*, 8(11), 391–394.
- Aljumah, F. H. (2020). Second language acquisition: A framework and historical background on its research. *English Language Teaching*, 13(8), 200–207.
- Al-Khresheh, M. H. (2015). A review study of interlanguage theory. *International Journal of Applied Linguistics & English Literature*, 4(3), 123–131.
- Allen, J. R. (2008). Why learning to write Chinese is a waste of time: A modest proposal. *Foreign Language Annals*, 41(2), 237–251.
- Allwood, J. (2008). Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 207–225). Mouton de Gruyter.
- Alshammari, H. (2020). Chinese language in Saudi Arabia: Challenges and recommendations. *English Language Teaching*, 13(2), 75–85.
- Amara, N. (2015). Error correction in foreign language teaching. *The Online Journal of New Horizons in Education*, 5(3), 58–68.
- Andersen, G. (2016). Semi-lexical features in corpus transcription: Consistency, comparability, standardisation. *International Journal of Corpus Linguistics*, 21(3), 323–347. <https://doi.org/10.1075/ijcl.21.3.04>
- Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1–16.
- Bardovi-Harlig, K., & Sprouse, R. A. (2018). Negative versus positive transfer. In *The TESOL encyclopedia of English language teaching* (pp. 1–6). Wiley.

- Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 511–533). Cambridge University Press.
- Barone, S. (2024). The importance of interlanguage in language teaching: An analysis of its development in L2 learners. *Vestnik of Samara State Technical University: Psychological and Pedagogical Sciences*, 21(3), 81–97.
- Baroni, A. (2011). Alphabetic vs. non-alphabetic writing: Linguistic fit and natural tendencies. *Rivista di linguistica*, 23(2), 127–159.
- Beare, S., & Bourdages, J. S. (2007). Skilled writers' generating strategies in L1 and L2: An exploratory study. In G. Rijlaarsdam (Series Ed.), M. Torrance, L. Van Waes, & D. Galbraith (Vol. Eds.), *Studies in writing* (Vol. 20, Writing and cognition: Research and applications, pp. 151–161). Elsevier.
- Benson, C. (2002). Trasfer/ Cross-linguistic influence. *ELT Journal: English Language Teaching Journal*, 56(1).
- Bi, Y. (2001). 美国学生学习汉语的声韵难点分析 [Analysis of phonetic difficulties in learning Chinese for American students]. *辽宁工业大学学报 (社会科学版) [Journal of Liaoning Institute of Technology (Social Science Edition)]*, (2), 39–41.
- Birdsong, D., Gertken, L.M., & Amengual, M. (2012). *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. COERLL, University of Texas at Austin. <<https://blp.coerll.utexas.edu>>.
- Boersma, P., & Weenink, D. (2007). *Praat: Doing phonetics by computer (Version 5.0.02)*. University of Amsterdam, Institute of Phonetic Sciences. <http://www.praat.org>
- Bonsignori, V., & Cappelli, G. (2022). Developing strategies for conceptual accessibility through multimodal literacy in the English for tourism classroom. *Multimodal Communication*, 11(1), 31–47. <https://doi.org/10.1515/mc-2020-0026>
- Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology*, 25(3).
- Bowles, M. A., Yan, X., & Malone, M. (2021). A narrative synthesis of assessment research on languages other than English: Identifying unique assessment issues.[Paper presentation]. Midwestern Association of Language Testers (MwALT) and Technology and Second Language Learning (TSL) Conference 2021, Iowa State University, USA.
- 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., 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>
- Brown, H. D. (1994). *Principles of language learning and teaching (3rd ed.)*. Prentice Hall Regents.
- Brugman, H., & Russel, A. (2004). Annotating multi-media / multi-modal resources with ELAN. In M. Lino et al. (Eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (pp. 2065–2068). Lisbon, Portugal.
- Callies, M. (2023). Current perspectives on learner corpus research. *Arbeiten aus Anglistik und Amerikanistik*, 48(1), 37–51.
- Callies, M. (2024). Challenges in the compilation, annotation, and analysis of learner corpus data. In M. Kaunisto & M. Schilk (Eds.), *Studies in Corpus Linguistics* (Vol. 118, pp. 55–67). John Benjamins Publishing Company.
- Callies, M., & Götz, S. (2015b). Learner corpora in language testing and assessment: Prospects and challenges. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 1–9). John Benjamins.
- Cao, D., & Xu, L. (2017). The study of genuine tenseless languages (in Chinese with English abstract). *Contemporary Linguistics*, 19, 93–121.
- Cao, F., Rickles, B., Vu, M., Zhu, Z., Chan, D. H. L., Harris, L. N., ... & Perfetti, C. A. (2013). Early stage visual-orthographic processes predict long-term retention of word form and meaning: a visual encoding training study. *Journal of Neurolinguistics*, 26(4), 440–461.
- Cao, J. (1995). 连读变调与轻重对立[Tone sandhi in connected speech and the opposition between stress and lightness]. *中国语文[Studies of the Chinese Language]*, (4), 312–320.
- Cao, W., & Zhang, J. (2009). 面向计算机辅助正音的汉语中介语语音语料库的创制与标注 [The creation and annotation of a Chinese interlanguage speech corpus for computer-assisted pronunciation correction]. *语言文字应用 [Applied Linguistics]*, (4), 122–131. <https://doi.org/10.16499/j.cnki.1003-5397.2009.04.010>
- Cao, X. (2013). 留学生汉语中介语纵向语料库建设的若干问题 [Issues in constructing a longitudinal Chinese interlanguage learner corpus]. *语言文字应用 [Applied Linguistics]*, (2), 127–134.
- 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.
- Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap.

- Language Teaching*, 52(4), 460–475.
- Chan, J., Woore, R., Molway, L., & Mutton, T. (2022). Learning and teaching Chinese as a foreign language: A scoping review. *Review of Education*, 10(3), e3370.
- Chao, Y. R. (1930). A system of “tone letters”. *Le Maître Phonétique*, 8(45), 24–27.
- Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.
- Chen, B., & Ning, A. (2005). 汉字识别中的同音字效应：语音影响字形加工的证据 [Homophone effects in Chinese character recognition: Evidence for phonological influence on orthographic processing]. *心理学探新* [Psychological Exploration], (4), 35–39, 55.
- Chen, G., & Liang, M. C. (2010). Summary of the first national seminar on learner corpus research. *Foreign Language Education in China*, 2010(4), 77–80.
- Chen, J. D. (2014). 泰国学生汉字偏误研究综述 [A review of research on Chinese character writing errors of Thai students]. *语文学刊* [Journal of Philology], (11), 31–32, 68.
- Chen, M. (2011). 无声调语言母语者汉语声调范畴习得的实验研究 [An experimental study on the acquisition of Chinese tone categories by learners whose native languages are toneless]. *华文教学与研究* [TCSOL Studies], (4), 9–15.
<https://doi.org/10.16131/j.cnki.cn44-1669/g4.2011.04.002>
- Chen, Q., Liu, J., & Zhu, L. (2009). 泰国学生汉字书写偏误分析 [An analysis of Chinese character writing errors of Thai students]. *云南师范大学学报(对外汉语教学与研究版)* [Journal of Yunnan Normal University (Teaching and Researching Chinese as a Foreign Language Edition)], 7(2), 76–82.
<https://doi.org/10.16802/j.cnki.ynsddw.2009.02.014>
- Chen, S., He, Y., Wayland, R., Yang, Y., Li, B., & Yuen, C. W. (2019). Mechanisms of tone sandhi rule application by tonal and non-tonal non-native speakers. *Speech Communication*, 115, 67–77.
- Chen, Y. (2003). 病句类型及修改方法 [Types of ill-formed sentences and methods of correction]. *甘肃教育* [Gansu Education], (Z2), 51–52.
<https://doi.org/CNKI:SUN:GSJY.0.2003-Z2-040>
- Chuang, H. K., Joshi, R. M., & Dixon, L. Q. (2012). Cross-language transfer of reading ability: Evidence from Taiwanese ninth-grade adolescents. *Journal of Literacy Research*, 44(1), 97–119.
- Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching*, 5(4), 161–170.
- Corder, S. P. (1975). Error analysis, interlanguage, and second language acquisition. *Language Teaching*, 8(4), 201–218.

- Corder, S. P. (1982). *Error analysis and interlanguage*. Oxford University Press.
- Croft, W. (1994). Semantic universals in classifier systems. *Word*, 45(2), 145–171.
- Crosthwaite, P., & Boulton, A. (2022). DDL is dead? Long live DDL! Expanding the boundaries of data-driven learning. *Discovering language: Learning and affordance*.
- Cui, X. L., & Zhang, B. L. (2011). The global Chinese learner corpus construction plan. *Applied Linguistics*, 2, 100–108.
- Cui, Y. (1997). 汉字部件和对外汉字教学 [Chinese character components and teaching Chinese characters as a foreign language]. *语言文字应用*[*Applied Linguistics*], (3), 51–56.
- Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, 26(2), 163–174.
- Deng, D. (2020). 汉语语音研究与对外汉语语音教学[Research on Chinese phonetics and teaching Chinese pronunciation as a foreign language]. *国际汉语教育(中英文)* [*International Chinese Education (Chinese and English)*], 5(2), 51–59.
- Deng, S., & Hu, W. (2022). An examination of Chinese character writing errors: Developmental differences among Chinese as a foreign language learners. *Journal of Chinese Writing Systems*, 6(1), 39–51.
- Deng, S., & Hu, W. (2022). An examination of Chinese character writing errors: Developmental differences among Chinese as a foreign language learners. *Journal of Chinese Writing Systems*, 6(1), 39–51. <https://doi.org/10.1177/25138502211066611>
- Díaz-Negrillo, A., & Domínguez, J. F. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada*, (19), 83–102.
- Díaz-Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2), 139–154.
- Dong, Y., Tsubota, Y., & Dantsuji, M. (2013, November). Difficulties in perception and pronunciation of Mandarin Chinese disyllabic word tone acquisition: A study of some Japanese university students. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)* (pp. 143–152).
- Du, Y. (2007). 非汉字文化圈留学生汉字课教学策略研究[Teaching strategies for Chinese character courses for learners from non-character-based writing systems]. *上海商学院学报*[*Journal of Shanghai Business School*], (2), 66–68.
- Duan, Y. Y. (2019). 柬埔寨大学生汉字笔画书写偏误分析及教学策略 [An analysis of stroke writing errors of Cambodian university students and teaching strategies]. 辽宁大

- 学硕士学位论文[Master's thesis, Liaoning University].
- Duan, Z. (2015). 常见病句类型辨析 [An analysis of common types of erroneous sentences]. *语文教学通讯·D 刊(学术刊)* [*Language Teaching Communication · Series D (Academic Edition)*], (6), 67–68. <https://doi.org/10.13525/j.cnki.bclt.201506030>
- Du Bois, J. W., Edwards, J. A., & Lampert, M. D. (1993). Outline of discourse transcription. In *Talking Data: Transcription and Coding in Discourse Research* (pp. 45–87). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dumont, A. (2017). The contribution of learner corpora to the substantiation of fluency levels. In P. De Haan, R. D. Vries, & S. Van Vuuren (Eds.), *Language, learners and levels: Progression and variation* (pp. 281–308). Presses universitaires de Louvain.
- Ellis, R. (2008). *The Study of Second Language Acquisition (2nd ed.)*. Oxford University Press.
- Ellis, R. (2015). *Understanding second language acquisition 2nd edition*. Oxford university press.
- Erbaugh, M. S. (2004). Chinese classifiers: Their use and acquisition. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *Handbook of East Asian psycholinguistics* (Vol. 1, Chinese, pp. 39–51). Cambridge University Press.
- Fan, Y. (2015). 论对外汉语教学中的汉字教学 [On Chinese character teaching in teaching Chinese as a foreign language. *赤峰学院学报(哲学社会科学版)* [*Journal of Chifeng University(Philosophy and Social Science Chinese Edition)*], 36(2), 179–180.
- Fauziati, E. (2011). Interlanguage and error fossilization: A study of Indonesian students learning English as a foreign language. *Indonesian Journal of Applied Linguistics*, 1(1), 23–38.
- Fauziati, E. (2017). Native and target language influence on students' interlanguage production: A case of Indonesian EFL compositions. *Indonesian Journal of Applied Linguistics*, 7(1), 54–63.
- Fei, J. (1996). 现代汉字部件探究 [An investigation into modern Chinese character components]. *语言文字应用* [*Applies Linguistics*], (2), 20–26.
- Fei, J. (1997). 现代汉字笔画规范刍议 [On the standardisation of strokes in modern Chinese characters]. *世界汉语教学* [*Chinese Teaching in the World*], (2), 10–16. <https://doi.org/10.13724/j.cnki.ctiw.1997.02.002>
- Fei, J. (1998). 对外汉字教学的特点、难点及其对策 [Characteristics, difficulties and countermeasures in teaching Chinese characters to foreign learners]. *北京大学学报(哲学社会科学版)* [*Journal of Peking University(Philosophy and Social Sciences)*], (3), 118–126.

- Feng, L. (2015). 入门阶段汉字教学条件的分析与教学方法的思考 [An analysis of conditions and teaching methods for introductory Chinese character instruction]. *云南师范大学学报 (对外汉语教学与研究版)* [Journal of Yunnan Normal University(Teaching and Research on Chinese as A Foreign Language Edition)], 13(1), 1–11.
- Feng, T. (2004). “汉字文化圈”刍议 [On the “Chinese character cultural sphere”]. *吉首大学学报 (社会科学版)*, [Journal of Jishou University (Social Sciences)] 25(2), 1.
- Fillmore, L. W. (1976). The second time around: Cognitive and social strategies in second language acquisition. Stanford University.
- Fillmore, L. W. (1979). Individual differences in second language acquisition. In C. Fillmore (Ed.), *Individual Differences in Language Ability and Behavior* (pp. 203–228). Academic Press.
- Foster, M. E., & Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. In *Proceedings of the International Language Resources and Evaluation Conference (LREC)* (pp. 305–323).
- Franceschi, D. (2018). Simulating a lawyer–client interview: The case of Italian EFL university learners. *ESP Across Cultures*, 15, 41–59.
- Friginal, E., & Hardy, J. (2014). *Corpus-Based Sociolinguistics: A Guide for Students*. Routledge.
- Fu, Y. (1992). 汉字的笔画 [Strokes of Chinese characters]. *语文建设*, [Language Planning] (1), 8–12. <https://doi.org/10.16412/j.cnki.1001-8476.1992.01.003>
- Gablasova, D., Brezina, V., & McEnery, T. (2019). 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>
- Gan, J. (2020). The study on Chinese character acquisition errors of foreign students. *International Journal of Learning and Teaching*, 6, 38–42.
- Gan, R. R., & Zhang, P. (2005). A study on the development of the Korean-specific word list for Chinese as a second language teaching. *Applied Linguistics*, 2, 142–150.
- Gao, H. H. (2009, June). Cognitive barriers to the learning of Chinese noun classifiers by native Swedish speakers. In *Proceedings of the Second Conference of the Swedish Association for Language and Cognition*. Stockholm University.
- Gao, L. Q. (2001). 外国留学生规则字偏误分析——基于中介语语料库的研究 [An analysis of regular character errors of foreign students: Based on an interlanguage corpus]. *语言教学与研究* [Language Teaching and Linguistic Studies], (02), 47–52.
- Gao, L. Q., & Meng, L. (2000). 外国留学生汉语阅读中音、形信息对汉字辨认的影响

- [The influence of phonological and graphic information on character recognition by foreign students of Chinese]. *世界汉语教学 [Chinese Teaching in the World]*, (4), 67–76. <https://doi.org/10.13724/j.cnki.ctiw.2000.04.008>
- Gass, S. M. (1988). Second language acquisition and linguistic theory: The role of language transfer. In S. Flynn & W. O’Neil (Eds.), *Linguistic theory in second language acquisition* (pp. 384–403). Springer. https://doi.org/10.1007/978-94-009-2733-9_21
- Gass, S. M., & Selinker, L. (2001). *Second Language Acquisition: An Introductory Course*. Lawrence Erlbaum Associates.
- Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum* (pp. 1–8). Somerville, MA: Cascadia Proceedings Project.
- Gil, J. (2011). A comparison of the global status of English and Chinese: Towards a new global language? *English Today*, 27(1), 52–59.
- Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O’Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 359–370). Routledge.
- Gilquin, G., & Granger, S. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 9–34). Cambridge University Press.
- Gilquin, G., & Granger, S. (2015). From design to collection of learner corpora. *The Cambridge handbook of learner corpus research*, 3(1), 9-34.
- Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *International Journal of Corpus Linguistics*, 14(1), 1–26.
- Gong, C. (2014). 非汉字圈外国学生汉字偏误分析 [An analysis of Chinese character errors by non-Chinese character background students]. *现代语文（学术综合版） [Modern Chinese Language (Academic Edition)]*, (4), 114–115. <https://doi.org/CNKI:SUN:YWCZ.0.2014-04-046>
- Gong, M. (2022). 意大利本土中文教学资源发展现状、影响因素和未来建设 [The development status, influencing factors and future directions of local Chinese teaching resources in Italy]. *国际中文教育（中英文） [International Chinese Language Education]*.
- Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics*, 25(2), 207–231.
- Götz, S., & Granger, S. (2024). Learner corpus research for pedagogical purposes: An overview

- and some research perspectives. *International Journal of Learner Corpus Research*, 10(1), 1–38.
- Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & J. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). John Benjamins.
- Granger, S. (2003). Error-labeled learner corpora and CALL. *CALICO Journal*, 20(3), 465–480.
- Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In A. O'Keeffe & M. McCarthy (Eds.), *Corpora and Language Teaching* (pp. 13–32). Cambridge University Press.
- Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 485–510). Cambridge University Press.
- Granger, S. (2021). Commentary: Have learner corpus research and second language acquisition finally met? In S. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 258–273).
- Granger, S., & Paquot, M. (2017). Towards standardization of metadata for L2 corpora. Invited talk at the CLARIN workshop on Interoperability of Second Language Resources and Tools, 6–8 December 2017, University of Gothenburg, Sweden.
- Gu, Y. (2013). 论言思情貌整一原则与鲜活话语研究——多模态语料库语言学方法[On the principle of integrating speech, thought, emotion and manner and the study of vivid discourse: A multimodal corpus linguistic approach]. *当代修辞学*[*Contemporary Rhetoric*], (6), 1–19.
- Guan, C. Q., Liu, Y., Chan, D. H. L., Ye, F., & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology*, 103(3), 509.
- Guo, L., & Tao, L. (2008). Tone production in Mandarin Chinese by American students: A case study. In *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)* (Vol. 1, pp. 123–138).
- Guo, L., & Tao, L. (2008, April). Tone production in Mandarin Chinese by American students: A case study. In *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)* (Vol. 1, No. 1, pp. 123–138).
- Guo, Q. (2022). Interlanguage and its implications to second language teaching and learning. *Pacific International Journal*, 5(4), 8–14.
- Guo, S. L. (2008). 汉字的笔画特点与外国学生汉字笔画偏误 [Features of Chinese strokes

- and stroke errors of foreign students]. *语言教学与研究 [Language Teaching and Linguistic Studies]*, (03), 72–79.
- Guo, X., Tang, J., & Li, J. (2007). 汉字错误记忆的形、音编码差异研究[A study on differences in visual and phonetic encoding in erroneous memory of Chinese characters]. *心理学与行为研究[Research on Psychology and Behavior]*, (2), 86–92.
- Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, 32(3), 395–421.
- Han, L. (2014). 对日汉字教学中的偏误分析及教学建议 [An analysis of errors in teaching Chinese characters to Japanese students and pedagogical suggestions]. *语文学刊 [Journal of Language and Literature Studies]*, (11), 31–32, 68.
- Han, Z. (2005). *Fossilization in adult second language acquisition (Vol.5)*. Multilingual Matters.
- Hana, J., Rosen, A., Štindlová, B., & Štěpánek, J. (2014). Building a learner corpus. *Language Resources and Evaluation*, 48(4), 741–752.
- Hao, D. D. (2019). 柬埔寨吴哥高中孔子课堂学生汉字书写偏误研究及教学建议 [A study on Chinese character writing errors of Cambodian Confucius Classroom students at Angkor High School and pedagogical suggestions] 苏州大学硕士论文[Master's thesis, Suzhou University]. <https://doi.org/10.27351/d.cnki.gszhu.2019.001946>
- Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40(2), 269–279.
- Hepburn, A., & Bolden, G. B. (2017). *Transcribing for social research*. London: Sage.
- Hu, F., & Wang, X. (2011). HSK 动态口语语料库的语料转写研究 [Transcription of the HSK dynamic spoken corpus]. In *语言文字法制化、规范化、标准化、信息化建设——第七届全国语言文字应用学术研讨会论文集 [Proceedings of the 7th National Symposium on Language Application]* (pp. 321–327). Chinese Proficiency Test Center, Beijing Language and Culture University.
- Hu, X. Q. (2018). The construction of dynamic Chinese learner interlanguage corpora. *Journal of Shandong Normal University (Social Sciences)*, 5, 1–18.
- Hu, X., & Xu, X. (2020). 韩国汉语学习者中介语口语语料库的建设及意义 [Construction and significance of a spoken interlanguage corpus of Korean learners of Chinese]. *华文教学与研究 [Research on Chinese Language Teaching]*, (1), 53–59. <https://doi.org/10.16131/j.cnki.cn44-1669/g4.2020.01.009>
- Huang, S. (2022). 浅谈对外汉语课堂语音偏误分析及教学对策——以英语母语背景及法

- 语母语背景者的汉语学习者为例[Analysis and teaching strategies of phonetic errors in CFL classrooms: Taking English and French learners as examples]. *汉字文化 [Sinogram Culture]*, 10, 82–84.
- Huang, W. (2015). Preliminary discussion on the construction of multimodal Chinese interlanguage corpora. *International Chinese Language Teaching Research*, 3, 60–66.
- Huang, W. (2015). 多模态汉语中介语语料库建设刍议 [On the construction of a multimodal Chinese interlanguage corpus]. *国际汉语教学研究 [Journal of International Chinese Teaching]*, (3), 60–66.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- Ide, N., & Romary, L. (2004, May). A Registry of Standard Data Categories for Linguistic Annotation. In *LREC*.
- Ishikawa, S. (2022). L2 English learners' gesture use in a picture description task: A study based on the ICNALE Dialogue Module. In R. Ruiz Cecilia & J. R. Guijarro Ojeda (Eds.), *Investigación e innovación en lengua extranjera: Una perspectiva global* (pp. 721–750).
- Iurato, A. (2022a). Learner corpus research meets Chinese as a second language acquisition: Achievements and challenges. *Annali di Ca'foscari. Serie orientale*, 58, 709–742.
- Iurato, A. (2022b). Investigating Chinese learner corpus research and learner corpora. *Language Learning*, 45(3), 11–25.
- Iurato, A. (2023). Designing and compiling the written sub-corpus of the Bimodal Italian Learner Corpus of Chinese (BILCC): Methodological issues. In *Studies on Chinese Language and Linguistics in Italy* (pp. 197–228). Bologna University Press.
- Iurato, A. (2023). Designing and compiling the written sub-corpus of the Bimodal Italian Learner Corpus of Chinese (BILCC): Methodological issues. In *Studies on Chinese Language and Linguistics in Italy* (pp. 197–228). Bologna University Press.
- Iurato, A. (2024). Exploring the Pragmalinguistic Knowledge of the 是 shì... 的 de Cleft Construction in L1 Italian Learners' L2 Chinese: Triangulation of Corpus and Experimental Data. *Chinese as a Second Language Research from Different Angles*, 76–10.
- Jarvis, S. (2015). The scope of transfer research. *New perspectives on transfer in second language learning*, 17–47.
- Javadi-Safa, A. (2018). A brief overview of key issues in second language writing teaching and research. *International Journal of Education and Literacy Studies*, 6(2), 12–25.
- Jia, Y., & Zhu, X. (2012). 病句类型归纳 [A classification of erroneous sentence types]. *语文教学与研究 [Language Teaching and Research]*, (4), 54–55.
<https://doi.org/CNKI:SUN:YWJY.0.2012-04-033>
- Jiang, X., & Liu, Y. (2004). 拼音文字背景的外国学生汉字书写错误研究 [A study on

- Chinese character writing errors of foreign students with alphabetic writing backgrounds]. *世界汉语教学 [Chinese Teaching in the World]*, (1), 60–70+4. <https://doi.org/CNKI:SUN:SJHY.0.2004-01-007>
- Jin, L., & Hendriks, H. (2005). The development of aspect marking in L1 and L2 Chinese. *Working Papers in English and Applied Linguistics*, 9, 69–99.
- Jin, M. (2015a). 略论对外汉语教学初级阶段“语”与“文”的关系——兼及对外汉语教学初级阶段课程的设置[On the relationship between “language” and “script” at the elementary stage of TCFL]. *赤峰学院学报（哲学社会科学版） [Journal of Chifeng University(Philosophy and Social Science Chinese Edition)]*, 36(3), 253–254.
- Jin, M. (2015b). 汉字是汉语作为第二语言学习的最大难点简论[Chinese characters as the greatest difficulty in learning Chinese as a second language]. *赤峰学院学报（哲学社会科学版） [Journal of Chifeng University(Philosophy and Social Science Chinese Edition)]*, 36(2), 177–178.
- Jin, W. (2019). Acquisition of tone sandhis by English-speaking learners of Chinese. *Journal of the National Council of Less Commonly Taught Languages*, 25(1), 67–107.
- Jin, Z., & Shi, G. (2019). 意大利孔子学院发展现状、问题与策略研究 [The development, problems and strategies of Confucius Institutes in Italy]. *收藏[Collections]*, (3).
- Jin, Z., & Zhuang, Y. (2019). 意大利高校汉语课程设置模式及其启示[Models of Chinese curriculum design in Italian universities and their implications]. *辽宁师范大学学报（社会科学版） [Journal of Liaoning Normal University (Social Sciences Edition)]*, 42(1), 110–116.
- Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Teaching Journal*, 45(4), 1–16.
- Jones, R. H. (2021). Data collection and transcription in discourse analysis: A technological history. *The Bloomsbury handbook of discourse analysis*, 9-20.
- Kang, H. (2011). *Computer-based writing and paper-based writing: A study of beginning-level and intermediate-level Chinese learners' writing* (Doctoral dissertation, The Ohio State University).
- Kang, X. M. (2023). 从书写偏误看不同类型汉语学习者的汉字习得 [An analysis of Chinese character acquisition of different types of Chinese learners through writing errors]. *华中学术 [Huazhong Academic Journal]*, 2023(2), 209–221.
- Karim, K., & Nassaji, H. (2013). First language transfer in second language writing: An examination of current research. *Iranian Journal of Language Teaching Research*, 1(1), 117–134.
- Kasper, G. (2001). Classroom research on interlanguage pragmatics. In K. R. Rose & G. Kasper

- (Eds.), *Pragmatics in language teaching* (pp. 33–60). Cambridge University Press.
- Kasper, G., & Dahl, M. (1991). *Research methods in interlanguage pragmatics*. Second Language Teaching & Curriculum Center, University of Hawai‘i at Mānoa.
- Ke, B. (2003). 汉语拼音在国际汉语教学中的地位 and 运用 [The status and application of Hanyu Pinyin in international Chinese teaching]. *世界汉语教学* [Chinese Teaching in the World], (3), 67–72, 4. <https://doi.org/10.13724/j.cnki.ctiw.2003.03.010>
- Klein, W., Li, P., & Hendriks, H. (2000). Aspect and assertion in Mandarin Chinese. *Natural Language & Linguistic Theory*, 18(4), 723–770.
- Knight, D. (2009). *A multimodal corpus approach to the analysis of backchannelling behaviour* (Doctoral dissertation). The University of Nottingham.
- Knight, D. (2011). The future of multimodal corpora. *Revista Brasileira de Linguística Aplicada*, 11, 391–415.
- Knight, D., Evans, D., Carter, R., & Adolphs, S. (2009). HeadTalk, HandTalk and the corpus: Towards a framework for multimodal, multimedia corpus development. *Corpora*, 4(1), 1–32.
- Kotani, K., & Yoshimi, T. (2023). A design of multimodal learner corpus for practical foreign-language communication. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers.
- Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, 25(1), 161–168.
- Kuo, W. J., Yeh, T. C., Lee, J. R., Chen, L. F., Lee, P. L., Chen, S. S., Ho, L. T., Hung, D. L., Tzeng, O. J. L., & Hsieh, J. C. (2004). Orthographic and phonological processing of Chinese characters: An fMRI study. *NeuroImage*, 21, 1721–1731. <https://doi.org/10.1016/j.neuroimage.2003.12.007>
- Lado, R. (1957). *Linguistics across cultures*. University of Michigan Press.
- Lai, Y.-H. (2012). Perceptual processing of Mandarin nasals by L1 and L2 Mandarin speakers. *Journal of Psycholinguistic Research*, 41, 237–252.
- Lau, E., & Grüter, T. (2015). Real-time processing of classifier information by L2 speakers of Chinese. In *Proceedings of the 39th Annual Boston University Conference on Language Development* (pp. 311–323). Cascadilla Press.
- Lee, C., & Wang, A. (2006). 汉语声调特征教学探讨 [On the teaching of Chinese tone features]. *语言教学与研究* [Language Teaching and Research], (3), 70–75.
- Lee, C.-Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification

- of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics*, 37(1), 1–15.
- Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*, 8(4).
- Leech, G. (1997). Introducing corpus annotation. In G. Garside et al. (Eds.), *Corpus Annotation* (pp. 1–18). London: Longman.
- Leńko-Szymańska, A. (2017). Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology*, 21(3), 217–241.
- Li, B., & Li, Y. (2021). 中文纳入意大利国民教育体系的现状、特点及启示[Chinese as part of the Italian national education system: Status, characteristics and implications]. *国际中文教育（中英文）* [*International Chinese Language Education*], (3).
- Li, C. (2011). 日本学生学习汉语普通话声母的偏误分析及教学对策[Analysis and pedagogical suggestions for Chinese initial consonant errors among Japanese students]. *吉林省教育学院学报* [*Journal of Jilin Education Institute*], 27(2), 122–123. <https://doi.org/10.16083/j.cnki.1671-1580.2011.02.014>
- Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. University of California Press.
- Li, J. (2023). Study on Chinese tone acquisition of learners from different language backgrounds. In *Proceedings of TAI 2023* (pp. 101–105).
- Li, L. (2014). 对外汉字教学之留学生汉字书写偏误研究 [A study on writing errors of Chinese characters by international students in teaching Chinese as a foreign language]. *内蒙古师范大学学报（教育科学版）* [*Journal of Inner Mongolia Normal University (Educational Science Edition)*], 27(3), 102–104. <https://doi.org/CNKI:SUN:NMJX.0.2014-03-032>
- Li, L. (2021). *A spoken Chinese corpus: Development, description, and application in L2 studies* [Doctoral dissertation, Massey University]. Massey University Research Archive.
- Li, M., Wang, W., Tao, S., Dong, Q., Guan, J., & Liu, C. (2016). Mandarin Chinese vowel-plus-tone identification in noise: Effects of language experience. *Hearing Research*, 331, 109–118.
- Li, R., & Ye, B. (2013). 语文分进的教学模式对汉字能力的影响——针对非汉字文化圈学习者的实验研究 [Effects of a differentiated language-script teaching model on Chinese character ability: An experimental study of learners from non-character-based cultures]. *语言文字应用* [*Applied Linguistics*], (4), 98–106.
- Li, S., & Zhao, X. (2012). 欧美学习者的汉语声调偏误分析[Analysis of Chinese tone errors among European and American learners]. *北方文学* [*Northern Literature*], (9), 117–118.

- Li, Y., & Kang, J. (1995). 现代汉语形声字声符研究 [A study of phonetic components in modern Chinese phono-semantic characters]. In 现代汉语用字信息分析 [An analysis of character use in modern Chinese]. 上海: 上海教育出版社[Shanghai: Shanghai Education Press.].
- Li, Y., & Zhang, S. (2006). 汉字性质综论 [A comprehensive study of the nature of Chinese characters]. 北京师范大学学报 (社会科学版) [Journal of Beijing Normal University (Social Sciences)], (1), 68-76.
- Li, Y., Kang, J., Wei, L., & Zhang, S. (1992). 现代汉语形声字研究 [A study of phonetic-semantic compound characters in Modern Chinese]. 语言文字应用 [Applied Linguistics], (1), 74-83.
- Li, Z. Y. (2016). 初级水平汉语学习者汉字偏误调查 [A survey of character writing errors of beginner CFL learners]. 第六届东亚汉语教学研究生论坛暨第九届北京汉语教学研究生论坛论文集 (第 440-444 页) [In Proceedings of the Sixth East Asian Chinese Teaching Postgraduate Forum and the Ninth Beijing Chinese Teaching Postgraduate Forum (pp. 440-444). 北京: 北京外国语大学[Beijing: Beijing Foreign Studies University].
- Liang, M. (2023, July). Design and implementation of English multimedia corpus. In 2023 International Conference on Data Science and Network Security (ICDSNS) (pp. 1-9). IEEE.
- Liang, N. S.-Y. (2008). The acquisition of Chinese shape classifiers by L2 adult learners. In M. K. M. Chan & H. Kang (Eds.), *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)* (pp. 309-326). Ohio State University.
- Liang, Y. (2019). 从书写偏误看汉语二语学习者的汉字习得 [Chinese character acquisition of CFL learners from the perspective of writing errors]. 语言教学与研究 [Language Teaching and Linguistic Studies], (04), 33-44.
- Lin, J. K., Lu, H. C., Wang, C. C., Lin, H. Y., & Chen, B. (2025). Acoustically Precise Hesitation Tagging Is Essential for End-to-End Verbatim Transcription Systems. *arXiv preprint arXiv:2506.04076*.
- Lin, L. (2006). 东亚汉字文化圈及其形成论析 [The East Asian Chinese character cultural sphere and its formation]. 东南亚纵横 [Around Southeast Asia], (8), 58-62.
- Lin, M. (2004). 汉语语调与声调 [Chinese intonation and tone]. 语言文字应用 [Applied Linguistics], (3), 57-67. <https://doi.org/10.16499/j.cnki.1003-5397.2004.03.014>
- Lin, T. (1996). 语音研究和对外汉语教学 [Phonetic studies and teaching Chinese as a foreign language]. 世界汉语教学 [Chinese Teaching in the World], (3), 20-23.

- Lin, T., & Wang, L. (2013). *语音学教程（增订版）* [A course in phonetics (Revised ed.)]. 北京: 北京大学出版社 Beijing: Peking University Press.
- Liu, J. (2017). 面向日本大学生的汉语学习情况调查 [A survey of Chinese learning among Japanese university students]. *国际汉语教育（中英文）* [International Chinese Language Education], 2(1), 73–80.
- Liu, J. (2019). 四款可用于多模态口译语料库建设的软件功能比较研究 [A comparative study of four software programs applicable to multimodal interpreting corpus construction]. *文化创新比较研究* [Comparative Study on Cultural Innovation], 3(31), 63–64. <https://doi.org/CNKI:SUN:WCBJ.0.2019-31-031>
- Liu, J. H. (2008). 对外国学生汉字书写偏误的分析——兼谈汉字教学 [An analysis of Chinese character writing errors of foreign students—With reference to Chinese character teaching]. *语言教学与研究* [Language Teaching and Linguistic Studies], (02), 66–72.
- Liu, J. P. (2016). 对外汉字书写教学法研究——基于外国留学生汉字书写偏误分析 [A study on teaching methods of Chinese character writing: Based on error analysis of foreign students' writing]. *现代语文（语言研究版）* [Modern Chinese Language (Linguistic Research Edition)], (2), 115–118. <https://doi.org/CNKI:SUN:YWLY.0.2016-02-034>
- Liu, L., & Ma, Y. (1986). 普通话声调分布和声调结构频度 [Distribution and frequency of Mandarin tones]. *语文建设* [Language Planning], (3), 21–23. <https://doi.org/10.16412/j.cnki.1001-8476.1986.03.007>
- Liu, Q. H. (2017). 意大利威尼斯大学中文系学生汉字学习情况调查与分析 [A study on the Chinese character learning of students at the Department of Chinese, University of Venice] 辽宁师范大学硕士论文 [Master's thesis, Liaoning Normal University].
- Liu, T. (2009). “字本位”与对外汉字教学 [The character-centered approach and Chinese character teaching in TCFL]. *鲁东大学学报（哲学社会科学版）* [Ludong University Journal (Philosophy and Social Sciences Edition)], (4), 91–94.
- Liu, X. L. (2011). 基于中介语语料库的日本学生汉字书写偏误分析 [An analysis of Japanese students' Chinese character writing errors based on an interlanguage corpus]. *学理论* [Theory Research], (12), 362–364.
- Liu, Y. (2005). 非汉字文化圈留学生汉字认知难点及教学策略 [Cognitive difficulties in Chinese character learning among students from non-character-based cultures and teaching strategies]. *开封大学学报* [Journal of Kaifeng University], 19(2), 45–47.
- Liu, Y. (2022). 汉语口语中介语语料转写若干问题探讨 [On several issues in transcribing

- spoken Chinese interlanguage corpora]. 国际中文教育(中英文) [International Chinese Language Education (Chinese and English Edition)], 7(2), 23–29.
<https://doi.org/CNKI:SUN:GZZY.0.2022-02-002>
- Liu, Z. (2013). 基于对比分析假说的汉语声母习得难易度分析——以英语为基底语的学习者为例 [Difficulty analysis of Chinese initials acquisition based on contrastive analysis hypothesis: A case of English learners]. 云南师范大学学报(对外汉语教学与研究版) [Journal of Yunnan Normal University (Teaching Chinese as a Foreign Language Edition)], 11(6), 43–51. <https://doi.org/10.16802/j.cnki.ynsddw.2013.06.007>
- Lu, J. (2008). 《汉语拼音方案》和汉语教学 [The Scheme for the Chinese Phonetic Alphabet and Chinese teaching]. 语言文字应用 [Applied Linguistics], (3), 16–17.
- Lu, W. (2001). 现代汉语音节的数量与构成分布 [The number and structural distribution of modern Chinese syllables]. 语言教学与研究 [Language Teaching and Linguistic Studies], (6), 28–34.
- Lu, Y. (2007). 俄罗斯人学习汉语语音偏误分析 [Analysis of phonetic errors in Russian learners of Chinese]. 云南师范大学学报(对外汉语教学与研究版) [Journal of Yunnan Normal University (Teaching Chinese as a Foreign Language Edition)], 5(1), 52–55.
- Lu, Y. (2014). 中国汉字之象形字初探 [A preliminary study of pictographic characters in Chinese]. 语文建设 [Chinese Planning], (04X), 66–67.
- Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005*, 1, 14–17.
- Lv, L. N. (2014). 柬埔寨华校学生汉字书写偏误调查及分析 [A study on Chinese character writing errors of Cambodian Chinese school students] 华中师范大学硕士学位论文 [Master's thesis, Central China Normal University].
- Ma, D. Y. (2016). 韩国留学生汉字书写情况调查及偏误分析 [An investigation into Chinese character writing of Korean students and error analysis] 黑龙江大学硕士学位论文 [Master's thesis, Heilongjiang University].
- Ma, H. (2004). 摸清规律有的放矢——韩国学生学习汉语普通话语音的几个问题 [Understanding the patterns and targeted instruction: Some problems of Korean students in learning Mandarin pronunciation]. 天津外国语大学学报 [Journal of Tianjin Foreign Studies University], (2), 75–78.
- Ma, Q. (2003). 关于对外汉语教学的若干建议 [Several suggestions on teaching Chinese as a foreign language]. 世界汉语教学 [Chinese Teaching in the World], 17(3), 13–16.
- Ma, X. (2005). 汉语同音现象分析 [An analysis of homophony in Chinese]. 语文研究

- [*Linguistic Research*], (2), 15–18, 37.
- Ma, X., Gong, Y., Gao, X., & Xiang, Y. (2017). The teaching of Chinese as a second or foreign language: A systematic review of the literature 2005–2015. *Journal of Multilingual and Multicultural Development*, 38(9), 815–830.
- Ma, Y. (1994). 初级汉语水平留学生的普通话声调误区[Tone errors of beginning-level foreign students learning Chinese]. *北京师范大学学报(社会科学版)* [*Journal of Beijing Normal University (Social Sciences)*], (3), 88–93.
- Mackey, A., & Gass, S. M. (Eds.). (2011). *Research methods in second language acquisition: A practical guide*. John Wiley & Sons.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mahmood, A. H., & Murad, I. M. A. (2018). Approaching the language of the second language learner: Interlanguage and the models before. *English Language Teaching*, 11(10), 95–108.
- Mao, L. (2007). 声调偏误与对外汉语声调教学研究综述[A review of tone error research and CFL tone teaching]. *湖南第一师范大学学报*[*Journal of Hunan First Normal University*], (2), 119–122.
- Masini, F. (2009). Teaching Chinese in Italy: Past and present. *CHUN–Chinesischunterricht*, 24, 157–161.
- Matthews, S., & Yip, V. (2011). *Cantonese: A comprehensive grammar (2nd ed.)*. 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.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Mendikoetxea, A., Bielsa, S. M., & Rollinson, P. (2010). Focus on errors: Learner corpora as pedagogical tools. In *Corpus-based approaches to English language teaching* (p. 180).
- Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 379–400). Cambridge University Press.
- Meunier, F. (2021). Introduction to learner corpus research. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 23–36). Routledge.
- Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 537–

- 566). Cambridge University Press.
- Mukherjee, J., & Rohrbach, J.-M. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann & G. Marko (Eds.), *Planning, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop* (pp. 205–232). Peter Lang.
- Myles, F. (2009). Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. In *The Longitudinal Study of Advanced L2 Capacities* (pp. 74–88). Routledge.
- Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Learner corpus research: Current perspectives and applications* (pp. 309–331). Cambridge University Press. <https://doi.org/10.1017/CBO9781139649414.014>
- Myles, F. (2021). Commentary: An SLA perspective on learner corpus research. In S. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 258–273).
- Nai, X., Li, W., & Pu, J. (2007). COLSEC 语料库的设计原则与标注方法[Design principles and annotation scheme of the COLSEC corpus]. *当代语言学* [*Contemporary Linguistics*], 9(3), 235–246.
- Napoles, C., Sakaguchi, K., & Tetreault, J. (2017). JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.
- Nemser, W. (1971). Approximative systems of foreign language learners. *International Review of Applied Linguistics in Language Teaching*, 9(3), 115–134.
- Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 125–152). John Benjamins Publishing Company.
- Odlin, T. (1989). *Language transfer*. Cambridge University Press.
- O’Keeffe, A. (2020). Data-driven learning: A call for a broader research gaze. *Language Teaching*, 54(2), 259–272. <https://doi.org/10.1017/S0261444820000245>
- Osman, O. H., Abuoaf, M., & Al-Rahbi, A. H. (2022). Interlanguage theory revisited: Implications for the classroom. *International Journal of Linguistics, Literature and Translation*, 5(10), 30–40.
- Pan, W. (2006). 汉语的音节描写[On the description of Chinese syllables]. *语言科学* [*Linguistic Sciences*], (2), 39–43.
- Paquot, M., König, A., Stemle, E. W., & Frey, J. C. (2024). The Core Metadata Schema for

- Learner Corpora (LC-meta) Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research*, 10(2), 280-300.
- Park, E. S. (2011). 韩国学生汉字错字成因研究 [A study on the causes of wrongly written Chinese characters of Korean students] 华东师范大学硕士论文[Master's thesis, East China Normal University].
- Pasquarella, A., Chen, X., Gottardo, A., & Geva, E. (2015). Cross-language transfer of word reading accuracy and word reading fluency in Spanish–English and Chinese–English bilinguals: Script-universal and script-specific processes. *Journal of Educational Psychology*, 107(1), 96–110.
- Peng, Y., Yan, W., & Cheng, L. (2021). Hanyu Shuiping Kaoshi (HSK): A multi-level, multi-purpose proficiency test. *Language Testing*, 38(2), 326-337.
- Peng, Z. X. (2001). 常见别字的分类及其更正的理据分析 [The classification of common wrongly used characters and rationale for correction]. 桂林高等师范专科学校学报(综合版) [*Journal of Guilin Teachers' College (Comprehensive Edition)*], (4), 65–69. <https://doi.org/CNKI:SUN:GLJY.0.2001-04-018>
- Pérez-Paredes, P. (2022). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, 35(1–2), 36–61.
- Perfetti, C. A., & Tan, L. H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(1), 101–118.
- Pine, N., Ping'an, H., & Ren Song, H. (2003). Decoding strategies used by Chinese primary school children. *Journal of Literacy Research*, 35(2), 777–812.
- Polio, C. (1994). Non-native speakers' use of nominal classifiers in Mandarin Chinese. *Journal of the Chinese Language Teachers Association*, 29(3), 51–66.
- Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal*, 26(1), 8–14.
- Puspita, D. (2019). Error analysis on learners' interlanguage and intralanguage: A case study of two adolescent students. *Teknosastik*, 17(2), 12–18.
- Qian, X. Y. (2016). 初级意大利学生汉字书写偏误分析及教学策略研究 [An analysis of writing errors of beginner Italian students and pedagogical strategies] 重庆大学硕士论文[Master's thesis, Chongqing University].
- Qiu, X. (1988). 文字学概要 [*An outline of Chinese philology*]. 商务印书馆 [The Commercial Press].

- Ran, Q., & Yu, S. (2019). 汉语语音偏误的特点与模式——基于 25 种母语背景学习者的偏误条目数据的分析 [Characteristics and patterns of Chinese phonological errors: Evidence from learners with 25 L1 backgrounds]. *世界汉语教学*[*Chinese Teaching in the World*], (3), 417–432.
- Razali, N. (2015). Foreign students' common problems in writing Chinese characters. *WILAYAH: The International Journal of East Asian Studies*, 4(1), 63–78.
- Ren, H. (2010). 关于中介语语料库建设的几点思考——以“HSK 动态作文语料库”为例 [Some reflections on the construction of Chinese interlanguage corpora: Taking the “HSK dynamic composition corpus” as an example]. *语言教学与研究*[*Language Teaching and Research*], (6), 8–15. <https://doi.org/10.CNKI:SUN:YYJX.0.2010-06-003>
- Ren, H. B. (2010). Reflections on the construction of interlanguage corpora: Taking the "HSK dynamic composition corpus" as an example. *Language Teaching and Research*, 6, 8–15.
- Ren, Y. (2024). “字本位”理念下的国际中文词汇教学再思考 [Reconsidering international Chinese vocabulary teaching from a character-centered perspective]. *教育进展* [*Advances in Education*], 14(6), 547–551. <https://doi.org/10.12677/ae.2024.146971>
- Richards, J. C., & Schmidt, R. (2003). *Longman dictionary of language teaching and applied linguistics*. Foreign Language Teaching and Research Press.
- Richards, J. C., Platt, J., & Platt, H. (1996). *Dictionary of language teaching and applied linguistics*. Longman.
- Rosen, A., Hana, J., Štindlová, B., & Feldman, A. (2014). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, 48(1), 65–92. <https://doi.org/10.1007/s10579-013-9226-3>
- Saville-Troike, M. (2006). *Introducing second language acquisition*. Cambridge University Press.
- Schachter, J. (1974). An error in error analysis 1. *Language learning*, 24(2), 205–214.
- Schmidt, T., & Wörner, K. (2014). EXMARaLDA. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology* (pp. 402–419). Oxford University Press.
- Schmitt, B., & Zhang, S. (1998). Language structure and categorization: A study of classifiers in consumer cognition, judgment, and choice. *Journal of Consumer Research*, 25(2), 108–122.
- Schnur, E., et al. (2019). Introducing language teachers to learner corpora: The development of online tutorials for pedagogic uses of the MuSSeL corpus. *Learner Corpus Research*, 20(4), 77.

- Schumann, J. H. (1978). *The Pidginization Process: A Model for Second Language Acquisition*. Rowley, MA: Newbury House Publishers.
- Scovel, T. (2001). *Learning new languages: A guide to second language acquisition*. Heinle & Heinle.
- Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 213–234). John Benjamins Publishing Company.
- Selinker, L. (1969). Language transfer. *General Linguistics*, 9(2), 67–92.
- Selinker, L. (1972). Interlanguage. *IRAL: International Review of Applied Linguistics in Language Teaching*, 10, 209–231. <https://doi.org/10.1515/iral.1972.10.1-4.209>
- Selinker, L. (1983). Language transfer. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 33–68). Newbury House.
- Shen, H. H., & Xu, W. (2015). Active learning: Qualitative inquiries into vocabulary instruction in Chinese L2 classrooms. *Foreign Language Annals*, 48(1), 82–99. <https://doi.org/10.1111/flan.12137>
- Shi, L., & Wen, B. (2012). “洋腔洋调”初探——美国学生汉语语调习得 [A preliminary study of foreign-accented intonation: American learners' acquisition of Chinese intonation]. *南开语言学刊[Nankai Linguistics]*, (1), 42–49.
- Shi, W. (2010). 汉语语音特点及对外汉语语音教学[Characteristics of Chinese phonetics and pronunciation teaching in TCFL]. *辽宁教育行政学院学报[Journal of Liaoning Administration College]*, 27(7), 92–93. <https://doi.org/10.13972/j.cnki.cn21-1500/g4.2010.07.016>
- Shi, Z. (1998). 现代汉字的几何性质及其在汉字教学中的意义 [The geometric properties of modern Chinese characters and their pedagogical implications]. *语言文字应用 [Applied Linguistics]*, (4), 62–68.
- Shi, Z. Y. (1999). 外国留学生形符书写偏误分析 [An analysis of logographic writing errors of foreign students]. *世界汉语教学 [Chinese Teaching in the World]*, (01), 45–53.
- Shu, H., & Anderson, R. C. (1997). Role of radical awareness in the character and word acquisition of Chinese children. *Reading Research Quarterly*, 32(1), 78–89.
- Sinclair, J. (1996). EAGLES. Preliminary recommendations on corpus typology. *EAGLES Document EAG TCWG-CTYP/P*.
- Skoog, P. (2006). Native Language Interference: A study of interference patterns in Swedish students' English writing.
- Song, X., & Zhou, C. (2010). 符号“三分说”与汉字“六书” [The “tripartite theory of signs”

- and the Six Principles of Chinese characters] *浙江大学学报 (人文社会科学版)*[*Journal of Zhejiang University (Humanities and Social Sciences)*], 40(1), 135.
- Stymne, S., Pettersson, E., Megyesi, B., & Palmér, A. (2017). Annotating errors in student texts: First experiences and experiments. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 47–60).
- Su, P. (1994). *现代汉字学纲要* [An outline of modern Chinese character studies]. 北京大学出版社[Peking University Press].
- Su, P. (1995). 现代汉字的部件切分 [Component segmentation of modern Chinese characters]. *语言文字应用*[*Applied Linguistics*], (3), 52–55.
- Su, P. (2017). 汉字的性质 [The nature of Chinese characters]. *中国文字系统学报*[*Journal of Chinese Writing Systems*], 1(2), 109–115.
- Sun, D. (1993). 外国学生现代汉语“了 (le)”的习得过程初步分析 [A preliminary analysis of foreign learners' acquisition of the particle le]. *语言教学与研究*[*Language Teaching and Linguistic Studies*], (2), 65–75.
- Sun, D. K. (1993). Interlanguage theory and Chinese acquisition research. *Language Teaching and Research*, 4, 82–92.
- Sun, J., Ni, H., Mo, F., & Li, C. (1996). 普通话声调及其组合的统计特性[Statistical characteristics of Mandarin tones and their combinations]. *声学学报*[*Acta Acustica*], 21(1), 72–83.
- Sun, M. Y. (2021). 意大利华裔初级汉语学习者汉字习得偏误调查研究 [A study on Chinese character acquisition errors of Italian-Chinese beginner learners] 辽宁师范大学硕士学位论文[Master's thesis, Liaoning Normal University].
- Sun, Y. (2012). 国别教材的适用性研究——以《意大利人学汉语》为例 [A study of the applicability of country-specific textbooks: A case study of Italiani imparano il cinese]. *湖南师范大学教育科学学报*[*Journal of Education Science of Hunan Normal University*], 11(6), 46–52.
- Sun, Y. (2012). 国别教材的适用性研究——以《意大利人学汉语》为例 [A study on the applicability of country-specific textbooks: The case of Italiani imparano il cinese]. *湖南师范大学教育科学学报*[*Journal of Education Science of Hunan Normal University*], 11(6), 46–52.
- Tan, J. (2007). 外国学生学习汉语普通话声调和声母方面容易出现的问题[Common problems of foreign students in learning Mandarin tones and initials]. *黑龙江教育学院学报*[*Journal of Heilongjiang College of Education*], (11), 122–123.
- Tan, X., & Zhao, X. (2005). 汉语病句的识别与修改策略 [Identification and correction

- strategies for erroneous Chinese sentences]. *云南师范大学学报 [Journal of Yunnan Normal University]*, (5), 70–72. <https://doi.org/10.16802/j.cnki.ynsddw.2005.05.016>
- Tang, J. (2025). 第二语言学习者普通话单字调声学实验研究[An acoustic experimental study of Mandarin lexical tones by L2 learners]. *语言与翻译 [Language and Translation]*, (1), 73–79.
- Tarone, E. (1994). Interlanguage. In R. E. Asher (Ed.), *The encyclopedia of language and linguistics* (pp. 1715–1719). Elsevier.
- Tarone, E., Frauenfelder, U., & Selinker, L. (1976). Systematicity/variability and stability/instability in interlanguage systems. *Language Learning*, 26, 93–134.
- Tasker, I. (2018). Timeline analysis of complex language learning trajectories: Data visualisation as conceptual tool and method. *Applied Linguistics Review*, 9(2–3), 449–473.
- Tian, B. (1979). 浅谈集中识字教学[On intensive character teaching]. *天津教育 [Tianjin Education]*, (3), 5.
- Tian, M. (2022). 中级阶段留学生汉字书写偏误分析 [An analysis of Chinese character writing errors among intermediate-level international students]. *汉字文化 [Chinese Character Culture]*, (13), 152–155. <https://doi.org/10.14014/j.cnki.cn11-2597/g2.2022.13.004>
- Tian, W. (2018). 汉字学习者别字偏误共性分析 [A commonality analysis of wrongly used character errors among Chinese learners]. *汉字文化 [Chinese Character Culture]*, (21), 3–4. <https://doi.org/10.14014/j.cnki.cn11-2597/g2.2018.21.002>
- Tono, Y. (2009). The potential of learner corpora for pedagogical lexicography. In K. Matsuda (Ed.), *Perspectives in Lexicography: Asia and Beyond* (pp. 105–115). Tsinghua University Press.
- Towell, R., & Hawkins, R. (1994). *Approaches to second language acquisition*. Multilingual Matters.
- Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *Journal of the Acoustical Society of America*, 120(4), 2285–2294.
- Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., & Wirén, M. (2019). The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6, 67–104.
- Wan, Y. (1999). 汉字字符分工与部件教学 [Division of labour among Chinese characters and component teaching]. *语言教学与研究 [Language Teaching and Linguistic Studie]*,

- (4), 32–41.
- Wan, Y. (2000). 略论形声字声旁与对外汉字教学 [On phonetic components in phono-semantic characters and TCFL character teaching]. *世界汉语教学[Chinese Teaching in the World]*, (1), 62–69.
- Wan, Y. (2004). 从汉字研究到汉字教学 [From character studies to character teaching]. *世界汉语教学[Chinese Teaching in the World]*, (2), 40–48.
- Wan, Y. (2012). 略论汉语拼音和汉字在对外汉语教学中的位置和关系 [On the roles and relationship of Hanyu Pinyin and characters in TCFL]. *世界汉语教学[Chinese Teaching in the World]*, (3), 409–418.
- Wang, A., & Koo, M. (2014). 语音同化与韩国学生汉语普通话声母偏误分析[Phonetic assimilation and initial errors of Korean students learning Mandarin]. *世界语言教学[World Chinese Teaching]*, 28(4), 566–574. <https://doi.org/10.13724/j.cnki.ctiw.2014.04.014>
- Wang, H. (2005a). “字本位”与汉语二语教学 [The character-centered approach and L2 Chinese teaching]. *汉语教学学刊[Journal of Chinese Language Studies]*, 3, 58.
- Wang, J., & Zhang, L. (2008). 多媒体语料数据库在大学体验英语教学上的应用研究 [Application of multimedia corpus databases in college experiential English teaching]. *中国电化教育[China Educational Technology]*, (5), 64–67.
- Wang, L. (2005b). 《汉语拼音方案》与世界汉语语音教学[The Scheme for the Chinese Phonetic Alphabet and global Chinese phonetics teaching]. *世界汉语教学[Chinese Teaching in the World]*, (2), 5–11, 116. <https://doi.org/10.13724/j.cnki.ctiw.2005.02.001>
- Wang, L. (2012). *Second language acquisition of Mandarin aspect markers by native Swedish adults* (Doctoral dissertation). Acta Universitatis Upsaliensis.
- Wang, M., Malmasi, S., & Huang, M. (2015, June). The jinan chinese learner corpus. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 118-123).
- Wang, M., Malmasi, S., & Huang, M. (2015, June). The jinan chinese learner corpus. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 118-123).
- Wang, N. (1997). 汉字构形理据与现代汉字部件拆分 [Motivations for character formation and component segmentation in modern Chinese]. *语文建设[Chinese Planning]*, (3), 4–9.
- Wang, N. (2002). *汉字构形学讲座 [Lectures on Chinese character formation]*. 上海教育出版社[Shanghai Education Press (SEP)].

- Wang, R. (2013). 漢語「字本位」教學法分析 (An analysis of the character-centered teaching approach in Chinese). *山梨國際研究: 山梨県立大学国際政策学部紀要[Yamanashi glocal studies : bulletin of Faculty of Glocal Policy Management and Communications]*, 8, 76–84.
- Wang, R., Huang, S., Zhou, Y., et al. (2020). Chinese character handwriting: A large-scale behavioral study and a database. *Behavior Research Methods*, 52, 82–96. <https://doi.org/10.3758/s13428-019-01206-4>
- Wang, W. (1984). 现代汉语的短语结构和句子结构 [The phrase structure and sentence structure of modern Chinese]. *语文研究 [Language Research]*, (3), 11–20. <https://doi.org/CNKI:SUN:YWYJ.0.1984-03-001>
- Wang, X., & Wang, B. (2024). Exploring automatic methods for the construction of multimodal interpreting corpora. How to transcribe linguistic information and identify paralinguistic properties?. *Across Languages and Cultures*, 25(1), 48-70.
- Wang, W., & Wen, Q. (2002). L1 use in the L2 composing process: An exploratory study of 16 Chinese EFL writers. *Journal of Second Language Writing*, 11(3), 225–246.
- Wang, X., & Pan, X. L. (2004). Conception of establishing a logic-based interlanguage corpus. In M. Yang & Y. Zhang (Eds.), *Research on Chinese Language Teaching Theory and Methods* (pp. 525–530). Tsinghua University Press.
- Wang, Y. (2000). 韩国学生在汉语学习中常出现的语音问题与中韩语音的差异[Common pronunciation problems of Korean learners of Chinese and phonetic contrasts between Chinese and Korean]. *首都师范大学学报(社会科学版)[Journal of Capital Normal University (Social Sciences)]*, (S3), 112–116.
- Wang, Y. (2015). 日本留学生汉字书写偏误研究 [A study on Chinese character writing errors of Japanese students] 黑龙江大学硕士论文 [Master's thesis, Heilongjiang University].
- Wang, Y. D. (2010). 留学生汉字书写偏误及对策研究 [A study on Chinese character writing errors of foreign students and pedagogical countermeasures] 安徽大学硕士论文[Master's thesis, Anhui University].
- Wang, Y., & Li, J. (2001). 建立汉语中介语语音语料库的基本设想 [Preliminary ideas for constructing a Chinese interlanguage speech corpus]. *世界汉语教学 [Chinese Teaching in the World]*, (1), 87–92.
- Wang, Y., Kong, C., Yang, L., Wang, Y., Lu, X., Hu, R., & Sun, M. (2021). YACLC: A Chinese learner corpus with multidimensional annotation. *arXiv Preprint, arXiv:2112.15043*.
- Wang, Y., Sereno, J. A., Jongman, A., & Hirsch, J. (2003). fMRI evidence for cortical modification during learning of Mandarin lexical tone. *Journal of Cognitive*

- Neuroscience*, 15(7), 1019–1027.
- Wei, B. I. (2017). Issues in teaching Chinese characters and strategies for foreign learners outside the Chinese character culture circle. *Gdańskie Studia Azji Wschodniej*, (11), 98–111.
- Wei, X. (2008). Implication of IL fossilization in second language acquisition. *English Language Teaching*, 1(1), 127-131.
- Wen, Q. (2009). 二语习得跟踪研究的三个基本问题: 分类、设计与可比性 [Three fundamental issues in longitudinal research on second language acquisition: Categorization, design, and comparability]. *中国外语* [*Foreign Languages in China*], (2).
- Wen, Q. (2019). 二语习得重点问题研究 [*Key issues in second language acquisition*]. 外语教学与研究出版社[Foreign Language Teaching and Research Press].
- Wen, Q., & Hu, J. (2010). 中国大学生英语口语能力发展的规律与特点 [*Patterns and characteristics of Chinese college students' spoken English development*]. 北京: 外语教学与研究出版社 [Beijing: Foreign Language Teaching and Research Press].
- Wen, X. (1995). Second language acquisition of the Chinese particle *le*. *International Journal of Applied Linguistics*, 5(1), 45–62.
- 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.
- Wilson, A., & Thomas, J. (1997). Semantic annotation. In G. Garside et al. (Eds.), *Corpus Annotation* (pp. 53–65). London: Longman.
- Wolfersberger, M. (2003). L1 to L2 writing process and strategy: A look at lower proficiency writers. *TESL-EJ*, 7(2), 1–9. <http://www.tesl-ej.org/wordpress/issues/volume7/ej26a6/>
- Wu, M. J., Gao, D. G., & Xiao, X. Y. (2006). 欧美韩日学生汉字认读与书写习得研究 [A study on the recognition and writing acquisition of Chinese characters by European, American, Korean, and Japanese students]. *语言教学与研究* [*Language Teaching and Linguistic Studies*], (6), 64–71. <https://doi.org/CNKI:SUN:YYJX.0.2006-06-011>
- Xiang, C. C. (2012). 初级阶段留学生汉字偏误分析及汉字教学策略 [An analysis of writing errors of beginner CFL learners and teaching strategies] 上海外国语大学硕士论文[Master's thesis, Shanghai International Studies University].
- Xiao, R., & McEnery, T. (2004). Aspect in Mandarin Chinese. John Benjamins.
- Xiao, X. Q. (2002). 外国学生汉字偏误分析 [An analysis of Chinese character errors among foreign students]. *世界汉语教学* [*Chinese Teaching in the World*], (2), 79–85+4. <https://doi.org/CNKI:SUN:SJHY.0.2002-02-013>

- Xu, C. (2009). 整字速示条件下留学生汉字形误监控特点的研究 [Error monitoring of Chinese character forms under rapid whole-character presentation]. *汉语学习[Chinese Language Learning]*, (6), 71–78.
- Xu, H., Lu, X., & Brezina, V. (2019). Acquisition of the Chinese particle *le* by L2 learners: A corpus-based approach. In X. Lu & F. Myles (Eds.), *Second Language Acquisition and Corpora* (pp. 100–118). Bloomsbury Publishing.
- Xu, J. (2006). 俄罗斯留学生汉语声调偏误研究[A study on Chinese tone errors of Russian learners]. *佳木斯大学学报(社会科学版)[Journal of Jiamusi University (Social Science Edition)]*, (3).
- Xu, L. (2003). 对外汉语教学难点分析及对策 [Difficulties in TCFL and corresponding strategies]. *浙江师范大学学报 (社会科学版) [Journal of Zhejiang Normal University(Social Science Edition)]*, 28(5), 94–97.
- Xu, M., & Sun, Y. H. (2014). 韩国留学生汉字书写偏误及其成因分析 [An analysis of writing errors of Korean students and their causes]. *黄山学院学报 [Journal of Huangshan University]*, 16(04), 62–67.
- Xu, T. (1998). 说“字”——附论语言基本结构单位的鉴别标准、基本特征和它与语言理论建设的关係 [On “characters”: Criteria and features of basic linguistic units and their relevance to linguistic theory]. *语文研究[Linguistic Research]*, (3), 2–13.
- Xu, X., & Yao, L. (2014). “语文并进”和“语文分流”教学模式下学生动机调查——以非汉字文化圈学生为调查对象 [Learner motivation under integrated vs. separated language–script teaching models]. *华文教学与研究[TCSOL Studies]*, (3), 8–16.
- Xu, Z. (2000). 汉语作为外语教学的认知理论研究 [Cognitive theories of teaching Chinese as a foreign language]. 华语教学出版社[Sinolingua].
- Xu, Z., Wang, L. C., Liu, D., Chen, Y., & Tao, L. (2020). The moderation effect of processing efficiency on the relationship between visual working memory and Chinese character recognition. *Frontiers in Psychology*, 11, 1899. <https://doi.org/10.3389/fpsyg.2020.01899>
- Xü, G. (2006). 语音对比与对外汉语语音教学 [Phonetic comparison and Chinese pronunciation teaching for foreigners]. *广州大学学报(社会科学版)[Journal of Guangzhou University (Social Sciences)]*, (8), 87–89.
- Yan, X., & Lin, J. (2023). Chinese character matters!: An examination of linguistic accuracy in writing performances on the HSK test. *Assessing Writing*, 57, 100767.
- Yang, C. (2021). Categorical perception of Mandarin tones by native and second language speakers. In C. Yang (Ed.), *The Acquisition of Chinese as a Second Language*

- Pronunciation: Prosody, Phonology and Phonetics* (pp. 63–84). Singapore: Springer.
https://doi.org/10.1007/978-981-15-3809-4_3
- Yang, H. (2014). 汉语在意大利大学生中传播情况的调查研究[A survey of the spread of Chinese among Italian university students]. *文学与文化* [*Literature and Culture Studies*], (1), 141–144.
- Yang, S. (2016). “体假设”及“了”“着”的二语习得 [The Aspect Hypothesis and the acquisition of le and zhe in L2 Chinese]. *世界汉语教学* [*Chinese Teaching in the World*], 30(1), 101–118. <https://doi.org/10.13724/j.cnki.ctiw.2016.01.011>
- Yang, X. (1995). 谈多音字的成因及规范 [On the causes and standardisation of polyphonic characters]. *贵阳师专学报（社会科学版）* [*Journal of Guiyang University (Social Science)*], (3), 60–64. <https://doi.org/10.16856/j.cnki.52-1141/c.1995.03.014>
- Yang, Y., Li, S., Guo, Y., & Tian, Q. (2006). 建立汉语学习者口语语料库的基本设想[Basic proposals for constructing a spoken corpus of learners of Chinese]. *汉语学习* [*Chinese Language Learning*], (3), 58–64.
- Yang, Y., Zhang, J., Meng, Z. L., Qin, L., Liu, Y. F., & Bi, H. Y. (2018). Neural correlates of orthographic access in Mandarin Chinese writing: An fMRI study of the word-frequency effect. *Frontiers in Behavioural Neuroscience*, 12, 288. <https://doi.org/10.3389/fnbeh.2018.00288>
- Yi, A. (2012). On the Factors Influencing L1 Transfer. *Theory & Practice in Language Studies (TPLS)*, 2(11).
- Yi, B., & Liang, J. (2010). 作为第二语言的汉语声调习得研究回望[A retrospective review of research on tone acquisition in Chinese as a second language]. *天津师范大学学报（社会科学版）* [*Journal of Tianjin Normal University (Social Science Edition)*], (2), 77–80.
- Yip, V. (1995). *Interlanguage and learnability: From Chinese to English*. John Benjamins.
- You, H. (2003). 笔画数、部件数和拓扑结构类型对非汉字文化圈学习者汉字掌握的影响 [Effects of stroke count, component count and topological structure on Chinese character acquisition by learners from non-character-based cultures]. *世界汉语教学* [*Chinese Teaching in the World*], (2), 72–81.
- Zeng, J. (2008). 華語語音資料庫及數位學習應用 [*Chinese speech database and its application to digital learning*]. 新學林 [Xin Xue Lin].
- Zhang, B. (2011). 通用型汉语中介语语料库的标注模式 [Annotation model for a general-purpose Chinese interlanguage corpus]. 第七届全国语言文字应用研讨会（湖南）论文集 [In *Proceedings of the 7th National Symposium on the Application of Language and*

Script (Hunan)].

- Zhang, B., & Cui, X. (2013). “全球汉语中介语语料库建设和研究”的设计理念 [The design concept of the Global Chinese Interlanguage Corpus]. *语言教学与研究 [Language Teaching and Linguistic Studies]*, (5), 27–34.
- Zhang, B., & Cui, X. (2015). 谈汉语中介语语料库的建设标准 [On the construction standards of Chinese interlanguage corpora]. *语言文字应用 [Applied Linguistics]*, (2).
- Zhang, B., & Cui, X. (2016). 关于汉语中介语语料库标注规范研究的新思考——兼谈“全球汉语中介语语料库”标注规范的研制 [New thoughts on annotation standards for Chinese interlanguage corpora]. 在第四届汉语中介语语料库构建与应用国际研讨会论文集（扬州）中 [In *Proceedings of the 4th International Symposium on Chinese Interlanguage Corpus Construction and Application (Yangzhou)*].
- Zhang, B., & Cui, X. (2022). “全球汉语中介语语料库”的特点与功能 [Features and functions of the Global Chinese Interlanguage Corpus]. *世界汉语教学 [Chinese Teaching in the World]*, (1), 90–100.
- Zhang, H. (2012). 对外汉语声调习得研究综述 [A review of tone acquisition research in CFL]. *外国语言文学 [Foreign Languages and Literature]*, 28(S1), 66–68.
- Zhang, H. W., & Xie, J. (2014). Interlanguage fossilization in Chinese EFL writing: An empirical study of 20 English major students. *Sino-US English Teaching*, 11(4), 248–258.
- Zhang, H., Chang, X., Liu, W., Zhang, Y., & Wang, N. (2025). Identification of Mandarin tones in loud speech for native speakers and second language learners. *Behavioral Sciences*, 15(8), 1062.
- Zhang, J. J., & Li, M. (2010). The effects of Chinese input methods on the links among the orthography, phonology and meaning of Chinese characters. *Psychological Science*, 33(4), 835–838.
- Zhang, J., Peng, G., & Wang, W. S.-Y. (2016). Effects of language experience on categorical perception of Mandarin tones and consonants. In *Proceedings of INTERSPEECH 2016* (pp. 2433–2437).
- Zhang, K. J., Chen, L. M., & Lee, N. C. (2006, September). Tonal errors of Japanese students learning Chinese: A study of disyllabic words. *International Journal of Computational Linguistics & Chinese Language Processing*, 11(3), 281–296.
- Zhang, L. (2010). 字的本质和对外汉字教学 [The nature of Chinese characters and their role in TCFL]. *湖北师范学院学报（哲学社会科学版） [Journal of Hubei Normal University (Philosophy and Social Sciences)]*, (4), 16–21.

- Zhang, L. (2025). “字本位”与国际中文教育的元语言目标 [The character-centered approach and metalanguage goals in international Chinese education]. *Language Education*, 13(3).
- Zhang, L., & Xing, H. (2023). The interaction of orthography, phonology and semantics in the process of second language learners' Chinese character production. *Frontiers in Psychology*, 14, 1076810.
- Zhang, M., & Lu, X. (2010). 汉字的特点与对外汉语汉字教学 [Characteristics of Chinese characters and their teaching in TCFL]. *厦门广播电视大学学报*[*Journal of Xiamen Radio-television University*], 13(4), 34–38.
- Zhang, Q., & Lu, Z. (2014). The writing of Chinese characters by CFL learners: Can writing on Facebook and using machine translation help?. *Language Learning in Higher Education*, 4(2).
- Zhang, Q., & Min, G. (2019). Chinese writing composition among CFL learners: A comparison between handwriting and typewriting. *Computers and Composition*, 54, 102522.
- Zhang, R. (2014). 语料库汉字偏误分类和标注体系研究 [A study on the classification and annotation system of Chinese character errors in learner corpora]. *云南师范大学学报 (对外汉语教学与研究版)* [*Journal of Yunnan Normal University (Teaching and Research on TCFL)*], 12(1), 48–53. <https://doi.org/10.16802/j.cnki.ynsddw.2014.01.009>
- Zhang, R. (2016). 汉语中介语语料库中的汉字偏误处理研究 [A study on the handling of Chinese character errors in Chinese interlanguage corpora]. *语料库语言学* [*Corpus Linguistics*], 3(2), 50–59, 115. <https://doi.org/CNKI:SUN:YLYY.0.2016-02-004>
- Zhang, R. (2025). 留学生准别字的偏误特点和形成原因分析 [An analysis of quasi-wrongly-used character errors among international students and their formation reasons]. *语言教学与研究* [*Language Teaching and Linguistic Studies*], (3), 30–40. <https://doi.org/CNKI:SUN:YYJX.0.2025-03-003>
- Zhang, R., & Lei, Q. (2021). 英美高级水平汉语学习者别字书写研究及其理论蕴含 [A study on wrongly used character writing by advanced English and American learners of Chinese and its theoretical implications]. *语言教学与研究* [*Language Teaching and Linguistic Studies*], (6), 24–34. <https://doi.org/CNKI:SUN:YYJX.0.2021-06-004>
- Zhang, W. (1990). 从汉字部件到汉字结构——谈对外汉字教学 [From character components to character structure: On teaching Chinese characters as a foreign language]. *世界汉语教学*[*Chinese Teaching in the World*], (2), 112–120.
- Zhang, X. (2007). 多模态语料库：抢救濒危语言的有效途径[Multimodal corpora: An effective approach to documenting endangered languages]. In R. D. Ga, W. S. Slamu, &

- D. X. Jiacuo (Eds.), *民族语言文字信息技术研究——第十一届全国民族语言文字信息学术研讨会论文集*. 西苑出版社[*Research on Information Technology for Ethnic Minority Languages and Scripts—Proceedings of the 11th National Symposium on Ethnic Minority Languages and Scripts*. Xiyuan Press.].
- Zhao, D. (2013). 留学生汉字习得过程中“别字”类型分析 [An analysis of types of wrongly used characters in the process of Chinese character acquisition by international students]. *柳州职业技术学院学报* [*Journal of Liuzhou Vocational and Technical College*], 13(4), 53–58. <https://doi.org/10.16221/j.cnki.issn1671-1084.2013.04.012>
- Zhao, L. (1996). 外国留学生使用“了”的情况调查与分析 [A survey and analysis of foreign students' use of le]. 在 *第五届国际汉语教学讨论会论文选中*[*Selected Papers from the 5th International Conference on Chinese Language Teaching*].
- Zhou, Q., & Yu, S. (1996). 汉语短语标注标记集的确定 [The determination of label sets for Chinese phrase annotation]. *中文信息学报* [*Journal of Chinese Information Processing*], (4), 1–11. <https://doi.org/CNKI:SUN:MESS.0.1996-04-000>
- Zhou, X., Zhu, Q., & Deng, X. (2007). *外国人学汉语语法偏误研究* [*A study of grammatical errors made by foreigners learning Chinese*]. 北京语言大学出版社[Beijing Language and Culture University Press].
- Zhou, Y. (1957). 文字演进的一般规律 [General laws of writing system evolution]. *中国语文*[*Studies of The Chinese Language*],(7).
- Zhou, Y. (1979). 现代汉字中的多音字问题 (Polyphonic characters in modern Chinese). *中国语文*[*Studies of The Chinese Language*], (6).
- Zhou, Y. (1989). 汉字文化圈 [The Chinese character cultural sphere]. *中国文化*[*Chinese Culture*], (1), 47–54.
- Zhu, W. (2014). 非目的语环境中韩国大学生汉语声母习得偏误及教学策略研究 [Acquisition of Mandarin initials by Korean university students in non-target-language environments]. *长春师范大学学报*[*Journal of Changchun Normal University*], 33(9), 145–149.
- Zhu, Y., & Sun, Y. (2014). 意大利汉语教育的现状、问题与对策[The status quo, problems and countermeasures of Chinese education in Italy]. *云南师范大学学报（对外汉语教学与研究版）* [*Journal of Yunan Normal University(Teaching and Research on Chinese as A Foreign Language Edition)*], 12(4), 87–92.
- Zhu, Z. (2002). 汉字构形学说与对外汉字教学 [Chinese character formation theories and TCFL character instruction]. *语言教学与研究*[*Language Teaching and Linguistic Studies*], (4), 35–41.

- Zhu, Z. X., Liu, L., Ding, G. S., & Peng, D. L. (2009). The influence of Pinyin typewriting experience on orthographic and phonological processing of Chinese characters. *Acta Psychologica Sinica*.
- Zhu, Z., & Halina. (1999). 波兰学生暨欧美学生汉字习得的考察、分析和思考 [An investigation into Chinese character acquisition among Polish and Western learners]. *北京师范大学学报（社会科学版）* [*Journal of Beijing Normal University(Social Sciences)*], (6), 88–94.
- Zimmer, H. D., & Fischer, B. (2020). Visual working memory of Chinese characters and expertise: The expert's memory advantage is based on long-term knowledge of visual word forms. *Frontiers in Psychology*, 11, 516. <https://doi.org/10.3389/fpsyg.2020.00516>

Appendices

Appendix A

Metadata collection form used in MICICL main study

Metadata Collection Form

We would like to ask you to help us by answering the following questions concerning your language history and proficiency. The survey consists of 8 questions and will take less than 10 minutes to complete. This is not a test, so there are no right or wrong answers. Please answer every question and give your answers sincerely. Thank you very much for your help.

I. Biographical Information

Name and Surname		Gender	☐ male ☐ female ☐ other
Country of birth		Mother tongue(s)	
Current course level	☐ Bachelor ☐ Master ☐ PhD	Year of study	☐ First year ☐ Second year ☐ Third year
Degree course			
Home language (the most commonly used language in daily life)			
Which other languages do you know? (and at what level)			

II. Language History

In this section, we would like you to answer some factual questions about your language history by placing a check in the appropriate box.

1. How many years of classes (grammar, history, literature etc.) have you had in Chinese (primary school through university)?

☐ 0
 ☐ 1
 ☐ 2
 ☐ 3
 ☐ 4
 ☐ 5
 ☐ 6
 ☐ 7
 ☐ 8
 ☐ 9
 ☐ 10
 ☐ 11
 ☐ 12
 ☐ 13
 ☐ 14
 ☐ 15
 ☐ 16
 ☐ 17
 ☐ 18
 ☐ 19
 ☐ 20+

2. In an average week, how many hours do you spend on learning/exposure to Chinese? (Including taking classes, self-study, watching Chinese movies, etc.)

☐ 0
 ☐ 1
 ☐ 2
 ☐ 3
 ☐ 4
 ☐ 5
 ☐ 6
 ☐ 7
 ☐ 8
 ☐ 9
 ☐ 10
 ☐ 11
 ☐ 12
 ☐ 13
 ☐ 14
 ☐ 15
 ☐ 16
 ☐ 17
 ☐ 18
 ☐ 19
 ☐ 20+

3. How many months have you spent in a country/region where Chinese is spoken?

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17 ☐ 18 ☐ 19 ☐ 20+

III. Language Proficiency

In this section, we would like you to rate your language proficiency by giving marks from 0 to 6.

4. How well do you speak Chinese?

0=not well at all 6=very well
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

5. How well do you understand Chinese?

0=not well at all 6=very well
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

6. How well do you read Chinese?

0=not well at all 6=very well
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

7. How well do you write Chinese?

0=not well at all 6=very well
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

8. What's your HSK test level?

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

*(Each number represents a different HSK level.
 0 means that you have never taken the HSK test.)*

Appendix B

Chinese proficiency level of each student in MICICL

Student Code	Profeciency Level	Student Code	Profeciency Level
01	Intermediate	21	Advanced
02	Advanced	22	Advanced
03	Beginner	23	Intermediate
04	Advanced	24	Intermediate
05	Advanced	25	Advanced
06	Intermediate	26	Beginner
07	Intermediate	27	Advanced
08	Intermediate	28	Intermediate
09	Advanced	29	Intermediate
10	Intermediate	30	Intermediate
11	Intermediate	31	Intermediate
12	Advanced	32	Intermediate
13	Beginner	33	Intermediate
14	Intermediate	34	Advanced
15	Intermediate	35	Intermediate
16	Beginner	36	Intermediate
17	Beginner	37	Intermediate
18	Advanced	38	Intermediate
19	Beginner	39	Intermediate
20	Advanced		

Appendix C

Number of samples provided by each student

Student code	No. of writing samples	No. of reading samples
01	6	6
02	5	5
03	3	3
04	6	6
05	8	7
06	8	8
07	8	8
08	6	6
09	8	8
10	8	8
11	10	10
12	6	6
13	9	9
14	2	2
15	2	2
16	1	0
17	1	1
18	1	1
19	9	9
20	2	2
21	9	9
22	1	1
23	4	4
24	7	7
25	4	4
26	2	2
27	2	2
28	2	2
29	2	2
30	2	2
31	2	2
32	2	2
33	1	1
34	1	1
35	1	1
36	1	1
37	1	1
38	1	1
39	1	1

Appendix D

39 students' incorrectly written character ratio

Student code	No. of incorrectly written characters	Total number of characters	Incorrectly written ratio
No.1	10	1458	0.69%
No.2	42	986	4.26%
No.3	32	275	11.64%
No.4	62	1450	4.28%
No.5	14	2423	0.58%
No.6	44	1546	2.85%
No.7	65	1889	3.44%
No.8	16	1478	1.08%
No.9	39	2193	1.78%
No.10	15	2552	0.59%
No.11	52	1787	2.91%
No.12	0	989	0.00%
No.13	1	1001	0.10%
No.14	3	339	0.88%
No.15	4	308	1.30%
No.16	10	56	17.86%
No.17	0	64	0.00%
No.18	0	152	0.00%
No.19	44	1030	4.27%
No.20	3	354	0.85%
No.21	37	2530	1.46%
No.22	1	128	0.78%
No.23	17	1063	1.60%
No.24	24	1308	1.83%
No.25	10	863	1.16%
No.26	12	407	2.95%
No.27	4	546	0.73%
No.28	14	295	4.75%
No.29	5	508	0.98%

No.30	1	354	0.28%
No.31	3	500	0.60%
No.32	17	452	3.76%
No.33	0	276	0.00%
No.34	0	346	0.00%
No.35	0	399	0.00%
No.36	0	270	0.00%
No.37	1	176	0.57%
No.38	4	143	2.80%
No.39	1	137	0.73%

Appendix E

Total incorrectly written target character types in MICICL and their frequency

Character Type	Frequency	Character Type	Frequency	Character Type	Frequency
喜	47	觉	2	亲	1
候	23	者	2	搬	1
或	17	虽	2	喝	1
春	16	英	2	差	1
京	14	带	2	校	1
帮	12	聊	2	练	1
夏	11	逛	2	桥	1
住	10	们	2	照	1
知	8	观	2	集	1
看	7	原	2	束	1
书	7	怎	2	暖	1
爱	6	浪	2	府	1
假	6	赢	2	择	1
季	6	变	2	息	1
雪	6	我	2	茶	1
找	5	都	2	矮	1
庭	5	机	2	当	1
影	5	除	2	圣	1
高	5	认	2	首	1
言	5	识	2	更	1
争	5	音	2	聪	1
试	5	国	2	静	1
早	5	忘	2	能	1
事	5	真	2	以	1
博	5	楼	2	面	1
望	4	州	2	笔	1
现	4	复	2	馆	1
便	4	堂	2	身	1

店	4	尝	2	眼	1
课	4	厅	2	半	1
旅	4	友	2	少	1
漂	4	道	2	算	1
架	4	相	2	贵	1
你	4	费	1	的	1
因	4	助	1	蛋	1
谢	4	买	1	舞	1
趣	3	空	1	典	1
新	3	品	1	梦	1
值	3	间	1	跟	1
感	3	晚	1	写	1
想	3	德	1	活	1
参	3	饭	1	富	1
希	3	难	1	洗	1
最	3	等	1	数	1
来	3	明	1	爬	1
年	3	腿	1	傅	1
妈	3	未	1	努	1
重	3	冬	1	洛	1
睡	3	颜	1	屋	1
教	3	再	1	杭	1
室	3	那	1	终	1
猫	3	欧	1	智	1
奶	3	光	1	曲	1
但	3	度	1	么	1
飞	3	树	1	切	1
第	3	抱	1	糕	1
脑	3	发	1	酷	1
秋	3	荐	1	温	1
特	3	情	1	迁	1
打	3	挣	1	点	1
餐	3	邻	1	孕	1
钱	3	弃	1	判	1
常	3	题	1	举	1

欢	3	随	1	定	1
咱	3	市	1	北	1
游	3	食	1	港	1
萨	2	果	1	然	1
得	2	次	1	统	1
幸	2	准	1	境	1
般	2	瘦	1	同	1
所	2	每	1	商	1
对	2	胖	1	错	1
它	2	决	1	实	1
说	2	男	1	加	1
园	2	婚	1	宜	1
热	2	班	1	电	1
爸	2	故	1	凯	1
懂	2	赛	1	永	1
净	2	到	1	心	1
铁	2	房	1	建	1
结	2	介	1	亮	1
样	2	美	1		

Appendix F

39 Students' incorrectly used character ratio

Student Code	No. of incorrectly used characters	Total number of characters	Incorrectly used character ratio
No.1	11	1458	0.75%
No.2	21	986	2.13%
No.3	6	275	2.18%
No.4	23	1450	1.59%
No.5	17	2423	0.70%
No.6	32	1546	2.07%
No.7	40	1889	2.12%
No.8	32	1478	2.17%
No.9	25	2193	1.14%
No.10	21	2552	0.82%
No.11	27	1787	1.51%
No.12	11	989	1.11%
No.13	0	1001	0.00%
No.14	4	339	1.18%
No.15	4	308	1.30%
No.16	0	56	0.00%
No.17	0	64	0.00%
No.18	1	152	0.66%
No.19	23	1030	2.23%
No.20	1	354	0.28%
No.21	35	2530	1.38%
No.22	0	128	0.00%
No.23	22	1063	2.07%
No.24	16	1308	1.22%

No.25	15	863	1.74%
No.26	8	407	1.97%
No.27	2	546	0.37%
No.28	3	295	1.02%
No.29	6	508	1.18%
No.30	4	354	1.13%
No.31	12	500	2.40%
No.32	11	452	2.43%
No.33	0	276	0.00%
No.34	1	346	0.29%
No.35	0	399	0.00%
No.36	2	270	0.74%
No.37	0	176	0.00%
No.38	0	143	0.00%
No.39	1	137	0.73%

Appendix G

Total incorrectly used target character types in MICICL and their frequency

Character type	Frequency	Character type	Frequency	Character type	Frequency
菜	13	没	2	教	1
观	9	吵	2	喜	1
弗	8	架	2	站	1
力	8	俩	2	极	1
猫	8	近	2	亮	1
但	7	注	2	自	1
夏	7	食	2	组	1
字	6	能	2	果	1
她	6	得	2	店	1
园	6	虽	2	咖	1
酒	5	较	2	真	1
公	5	同	2	牛	1
活	5	因	2	期	1
是	5	常	2	铁	1
或	5	季	2	通	1
哪	5	机	2	景	1
气	5	弟	2	第	1
告	5	干	1	明	1
西	4	远	1	直	1
毕	4	利	1	知	1
共	4	斯	1	目	1
每	4	式	1	会	1
便	4	钟	1	披	1
汽	4	认	1	文	1

孩	4	雄	1	化	1
介	4	水	1	错	1
馆	4	界	1	玩	1
它	4	乡	1	运	1
般	3	诉	1	送	1
择	3	历	1	处	1
那	3	刻	1	怀	1
们	3	当	1	轻	1
乐	3	饭	1	手	1
术	3	花	1	叫	1
坏	3	使	1	暖	1
更	3	想	1	选	1
担	3	意	1	起	1
父	3	词	1	我	1
城	3	作	1	有	1
以	3	保	1	清	1
决	3	济	1	确	1
部	3	助	1	幹	1
像	3	博	1	电	1
友	3	钱	1	拿	1
室	3	闹	1	位	1
白	3	午	1	罗	1
简	2	鸟	1	亲	1
写	2	昨	1	己	1
房	2	特	1	请	1
痛	2	睡	1	闻	1
说	2	绍	1	网	1
已	2	夫	1	话	1

他	2	士	1	厅	1
母	2	吧	1	啦	1
蓝	2	物	1	外	1
方	2	性	1	舒	1
来	2	格	1	购	1
惊	2	谊	1	校	1
聪	2	到	1	伦	1
回	2	刑	1	境	1
兴	2	未	1	间	1
议	2	映	1	止	1
及	2	游	1	需	1
赞	2	换	1	标	1
给	2	值	1	觉	1
课	2	之	1	建	1
成	2	种	1	正	1
此	2	剧	1	交	1
幸	2	理	1	球	1
雪	2	问	1	啡	1
北	2	康	1	扶	1
筑	2	都	1	许	1
冷	2	住	1	始	1
具	2	两	1	净	1
东	2	安	1	欢	1
