The Development of the Mandarin Interlanguage Corpus (MIC)—A Preliminary Report on a Small-Scale Learner Database

Wai-lan Tsang
The University of Hong Kong; The University of Cambridge
Yuk Yeung
The University of Hong Kong

In this paper we report on the recent construction of a small-scale learner corpus with written and spoken output from pre-intermediate to intermediate Mandarin learners of different first languages—The Mandarin Interlanguage Corpus (MIC). The learners attended a 2-year certificate course on Mandarin Chinese at a tertiary institution in Hong Kong. Both their written and spoken production in the form of coursework and examinations, amounting to a total of about 50,000 characters and 60 hours of oral output, has been included in the database so far. The rationale, methodologies (i.e., collection, transcription, and annotation), and design of the database are described. Challenges in compiling the database are also addressed.

本論文は、The Mandarin Interlanguage Corpus (MIC) という、様々な母語話者から構成される初中級から中級までの中国語学習者の書きことばと話しことばの小規模学習コーパスの構築過程を報告する。学習者たちは香港の高等教育機関における 2 年間の中国語課程に参加しており、授業および試験において収集されたサンプルは、これまでに約50,000文字の書きことばと60時間に相当する話しことばが収録された。本コーパス構築における論理的根拠、方法論（収集、書き起こしと注釈の付記）、およびデータベース設計について紹介し、データベースを編纂する際の諸問題についても考察する。

JALT Journal, Vol. 34, No. 2, November 2012

Building on a foundation laid by previous projects on corpora on Mandarin Chinese (e.g., *The Lancaster Corpus of Mandarin Chinese,* McEnery & Xiao, 2004; *The Sheffield Corpus of Chinese*, Hu, Williamson, & McLaughlin, 2005; *The UCLA Chinese Corpus*, Tao & Xiao, 2007), this paper presents a project which has compiled a small-scale database with both written and spoken output from learners of different first languages (L1s) in a certificate course in Mandarin. It is intended that the database will serve as another public resource for researchers, teachers, and students of Chinese as a second or foreign language.

The project was motivated by the development of corpus linguistics in the 1980s and subsequent corpus development in China. As reviewed by Feng (2006), the notion of corpora appeared on the mainland as early as the 1920s, in the form of small-scale non-machine-readable corpora (e.g., *The Applied Glossary of Modern Chinese*). Since the development of corpus linguistics worldwide in the late 1980s, many corpora have been or are being compiled (see Feng, 2006, and Zhan, Chang, Duan, & Zhang, 2006, for their detailed reviews of the development of corpora in China). The majority of the corpora are of the general monolingual written type, serving as databases of written Mandarin Chinese in various genres (e.g., newspapers, literary texts, and textbooks, as in *The Academia Sinica Balanced Corpus of Modern Chinese*, discussed in Huang & Chen, 1992). While some spoken corpora were constructed (e.g., *The Contemporary Beijing Spoken Chinese Corpus*, as recorded in X. J. Yang, 2006), their advancement has been far behind that of the written counterpart, as shown in the aforementioned reviews. Therefore, spoken corpora certainly deserve much more investigation, as noted in Jia (2006), J. Yang (2008), X. J. Yang (2006), and Zhou (2007).

Yet to be fully developed are not only general spoken corpora but also learner corpora (or interlanguage corpora), one type of specialised corpus. Both X. J. Yang (2006) and Zhan et al. (2006) cite *The Chinese Interlanguage Corpus* (or *The Corpus of the Chinese Language as Interlanguage*) as the only example of an interlanguage corpus (disappointingly, it is not accessible to the public or available on the Internet). According to their descriptions, the corpus started in 1995 and comprises written texts from foreign students. From 1993 to 1995, the Beijing Language and Culture University (BLCU) constructed the BLCU Chinese Interlanguage Corpus, storing 1,371 compositions written by 740 students. Some years later, the International R&D Center for Chinese Education at the BLCU compiled *The Inter-Media HSK Essay Corpus* (Beijing Language and Culture University, 2003, 2009), which appears to be the most extensive one and is open to the public (HSK = Hanyu

Shuiping Kaoshi, or the Chinese Proficiency Test, a standardized test at the state level to assess the Chinese proficiency of nonnative speakers). Another recent searchable corpus, the Modern Interlanguage Chinese Corpus, comprises compositions and sentence-making tasks collected from Chinese studies students in years 2 to 4 at six Korean universities between 2004 and 2006, totalling 10,135 sentences. Another interlanguage corpus consists of written work done by elementary-level Chinese heritage learners in the US (Ming & Tao, 2008). Some other interlanguage corpora with written data are also currently being constructed, such as those by Lutong University and Shanghai Jiaotong University. Apart from these written interlanguage corpora, a recent spoken corpus was constructed by Cao and Zhang (2009) on interlanguage phonology.

Overall, the development of written Chinese interlanguage corpora is not surprising in that the techniques involved have been quite well established through the investigation of general corpora. However, this also implies that more spoken interlanguage corpora are yet to be compiled. Discussion has been ongoing (e.g., Wang & Li, 2001; Yang, Li, Guo, & Tien, 2006; Zhang, 2005), which helps emphasise the value and significance of the establishment of a spoken interlanguage corpus. Added to this is the importance of Mandarin as a second or foreign language in different parts of the world. The Asian context is a good example. In Japan, the importance and popularity of Chinese is acknowledged and Chinese is taught in some high schools (Gottlieb, 2012; Maher, 1995). Similarly, in Korea, Mandarin has been a popular foreign language (Teng & Yeh, 2001, as cited in Xing, 2006), and there are efforts to construct Chinese interlanguage corpora (such as the one mentioned above). In the certificate course reported on in the present study, Japanese and Korean native speakers formed two of the significant learner groups. All these instances in turn suggest that the development of a Mandarin learner or interlanguage corpus, be it written, spoken, or both written and spoken, is of utmost importance.

Development of Learner Corpora

In this section, the features of three of the learner interlanguage corpora mentioned above are presented so as to illustrate the existing development of Mandarin learner corpora.

The Inter-Media HSK Essay Corpus (Beijing Language and Culture University, 2003, 2009), constructed by researchers at the BLCU, is based on 11,569 compositions (4.24 million words in total) written by advanced

level students during the HSK examination in the period 1992-2005. Both original scripts produced by the students and error-tagged scripts are open to public access. Errors in the scripts are tagged in terms of five levels: character, punctuation, word, sentence, and paragraph or passage.

Ming and Tao's (2008) corpus collected written input from Chinese heritage learners whose Mandarin speaking and listening skills were at the advanced level but whose writing abilities varied. One thousand written samples (about 200,000 characters) were encoded in UTF-8, and segmented and tagged by ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) in the heritage corpus, and the errors in the samples are tagged according to 10 major categories and 30 subcategories (e.g., verb or verb phrase errors and nominal errors).

Cao and Zhang (2009) devised an interlanguage phonology corpus with reading samples collected from learners of Chinese in an experimental setting. The corpus is composed of six subcorpora (e.g., monosyllabic, disyllabic, sentence, and paragraph). In trying to tag nontarget-like pronunciations, Cao and Zhang addressed the inadequacy of using only IPA (International Phonetic Alphabet) symbols. Instead, they used SAMPA2C (a version of the Speech Assessment Methods Phonetic Alphabet) and C2ToBI (a version of the Chinese Tones and Break Indices), which are specifically designed for Mandarin Chinese.

The establishment of the three learner corpora helps show the recent development of learner corpora based on Chinese data. At the same time, they bring out certain problems or challenges in presenting as much information as possible about a given error. For example, Cao and Zhang (2009) acknowledged the need to use different symbols to represent learners' nontarget-like phonological output. In Ming and Tao's (2008) heritage corpus, the sentence errors cover a wide range from *inappropriate conjunctions* to *punctuation errors*. There is even a very general category labelled *multiple errors*. This in turn draws our attention to the need to handle learners' errors with care.

Aim of the Mandarin Interlanguage Corpus Project

In light of the possibility that an error can be subject to more than one judgement or interpretation, especially at the sentence or discourse level, the Mandarin Interlanguage Corpus (MIC) tags the errors at the character level, as explained below. This avoids the situation where errors receive different treatments by the research team and the user, and as a result do not

turn up in the search. At the same time, this also allows room for researchers and teachers to analyse the original data with their own interpretation in terms of their own focus.

With the intention of assembling a collection of written and spoken data from Mandarin learners, the MIC aims to

- Relieve the scarcity of learner corpora of Mandarin;
- Identify and track both written and spoken language patterns from Mandarin learners of different L1s who have reached the post-elementary level or above;
- Facilitate research comparing features among learners of different L1s and possibly different proficiency levels; and
- Enhance the development of teaching and assessment materials for learners of Mandarin.

In what follows, we will explain how data were collected from the learners in the certificate course and processed for the compilation of the database (e.g., tagging). We will then describe the design of the MIC and discuss the challenges in the construction of the database.

Data Source

Students from the Mandarin stream of a 2-year *Certificate Course in Chinese Language* offered by the Chinese Language Centre at a tertiary institution in Hong Kong were invited to participate in the corpus project. Both Year 1 and Year 2 students were recruited. They received 15 hours of in-class Mandarin input per week for three semesters in a year and took two final examinations, one at the end of each academic year, which assessed their reading, listening, writing, and speaking competence in Mandarin. Their level is supposed to reach intermediate on the HSK proficiency test upon completion of the course.

All students in the course have L1s other than Chinese. In total, 19 participants from two groups of Year 2 students were recruited within the 1.5-year project period (see Table 1).

As Table 1 shows, the student pool was not balanced as to L1. This is due to a number of reasons. First, the types of student taking the Chinese programme vary from year to year. In the first Year 2 group, there were students with English, Korean, and French as their L1. In the second Year 2 group, there were Japanese and Thai learners. An additional concern was students' willingness to join the project. Every participant was required to sign a consent form before their course output was recorded and processed for the

project. Otherwise, their output would not be recorded. Should they request to withdraw during the data collection period (which happened in two of the four oral classes involved), data collection had to cease and the collected data had to be discarded. Lastly, all participants were asked to complete a placement test so that the research team could ensure that their proficiency level had reached post-elementary by the time of data collection. It turned out that all Year 1 students attained very low scores in the placement test and so had to be excluded from the participant pool.[2]

Table 1. First Languages of the Participants of the MIC

First Language	Number of Participants
English	5
Korean	3
Japanese	3
German	2
French	1
Tamil	1
Indonesian	1
Spanish	1
Dutch	1
Thai	1
Total	19

Data Collection

Two kinds of output were collected during the course and from the end-of-course examination: written and spoken. The written output was in the form of short compositions ranging from 150 to 700 characters, depending on the genre. The spoken output was short presentations (1 to 2 minutes) delivered by the participants in class and during the examination. In both contexts, during the course and in the end-of-course examination, participants were given a topic to write on or talk about (e.g., *A memorable day*). While the topics were fixed by the course instructors, the output was considered to be fairly natural production by the participants because they were free to write or talk about whatever they liked in relation to the given topics.

A placement test, in the form of a fill-in-the-blank exercise and an error correction exercise, was administered to check the proficiency level of the

participants at the outset of data collection. This was to ensure that the project participants' level of Mandarin had reached at least post-elementary. The items tested were based on what they had studied in Year 1 and what they were learning or would learn in Year 2. The perfect score on the test was 50 and students attaining a score lower than 25 were not included in the database.[3]

With consent from the participants and after the completion of the placement test, photocopies of their written work were collected and digital recorders were placed in the classroom to record the short presentations. The participants were also requested to fill in a bio-data questionnaire which asked for personal information (e.g., nationality and age) and information relating to their language learning prowess (e.g., first language, other languages they speak, and how long they have learnt Mandarin). This was important in that their nationalities as recorded by the programme teacher did not necessarily predict their L1 (e.g., American nationality with Spanish as L1) or reveal any previous experience or knowledge of the Chinese language (e.g., American nationality with Chinese as the heritage language).[4]

So far, in total, 88 compositions (amounting to 50,000 characters) and 120 hours of recordings (of which 60 hours were observed to be coherent and therefore usable data, a point to be explained later) have been processed from coursework and examinations. The topics of the spoken presentations varied, usually depending on the time of year when they were recorded. For example, near the Chinese New Year, the students were asked to talk about their national festivals or the practices of the Chinese people in their home country. As to the written data, there was a balance between narration or description and exposition. Some sample topics are listed in Table 2.

Table 2. Sample Written and Spoken Topics in the MIC

Written Topics	Spoken Topics
A wedding	Festival
Autumn	A quarrel
My little treasure	Travelling
A memorable day	Making a choice
The world economy	Competition
How to nurture a kid	Role play
News censorship	On vacation

Data Processing

Based on the discussion about procedures for corpus compilation proffered in McEnery, Xiao, and Tono (2006), and Yang et al. (2006), the current project follows the typical flow of data handling. After photocopies of learners' compositions and digital audio files were collected, the *contextual* information about the data (e.g., nationality, age, and the number of languages spoken) collected via the questionnaire was also recorded. A code was assigned to each participant to keep identities anonymous (e.g., *F1* to refer to the first Mandarin learner with French as L1). Then the written and spoken raw data were coded in terms of three kinds of *textual* information: topic, mode (written or spoken), and source (coursework or examination). Each piece of output was thus coded in terms of contextual and textual information (for example, learner 1 with L1 X [X1]—written 1 [W1]—coursework composition 1 [CW1]), which corresponds to the mark-up process later on. For example, the first composition and presentation from the first French native speaker (F1) were labelled as F1CW1 and F1CO1 respectively; the first composition and the second presentation from the second French native speaker were F2CW1 and F2CO2. The number after W or O helps indicate the chronology of the learner output.

Then, the data were transformed into electronic versions: Written data were word-processed and spoken data were transcribed with traditional and simplified Chinese characters and *Hanyu Pinyin*. The data were then annotated with POS (part-of-speech) tagging, based on a tag set following the national standard in China and the HSK conventions as listed in standard HSK textbooks and references (e.g., Huang & Sun, 2000; Liu, et al., 2002-2005), and some key grammar references on Mandarin Chinese (e.g., Beijing Language and Culture University, 2000; Cheung, Liu, & Shih, 1994; Li & Thompson, 1981; Shao, 2007). Sixteen parts of speech and 40 sub-parts-of-speech were devised, totalling 52 categories (see the Appendix for the complete tag set).

After manual tagging and cross-checking of the data, the tagging was compared with that generated by the word tagging server designed by the CKIP (Chinese Knowledge and Information Processing) group. This server was chosen because it is open to the public for free. While the tag set of the CKIP server is different from that of the MIC, the broad categories (e.g., noun and verb) did help identify any improper tagging in the data. As to be explained later, the MIC administrator page houses a statistical programme to check internal tagging consistency.

In handling the written data, proper names were changed in order not to reveal the identity of the participants. As to the spoken data, the research

team read through every script to select coherent speech produced by each participant. Random utterances, chorus reading, repetition from teachers' samples, and metalinguistic discussions (e.g., about the usage of verb-object constructions) were excluded.

Regarding the nontarget-like forms (i.e., errors) in the learner output, the issue of whether and how they would be marked as in other existing interlanguage corpora was considered by the research team. In light of the possible subjective nature of correcting learner output and in order not to throw off other teachers' and researchers' searches for samples through our own interpretations of the learners' usage, error analysis was kept at the character level, including replacement of wrongly used or mispronounced characters and addition of missing characters. The character level was regarded as the place bearing the most prominent error type that was worth displaying and studying in the learner output and at the same time a more objective presentation of the learners' usage.

On the corpus search interface, whenever a nontarget-like character or missing character was found, it was inputted in brackets and placed next to the target character. The same technique was applied to pronunciation. When a character was mispronounced with a nontarget-like sound segment, it was indicated in brackets. Figures 1a and 1b illustrate what the tagging looks like.

Figure 1a. Sample Error Tag in the MIC (Written)

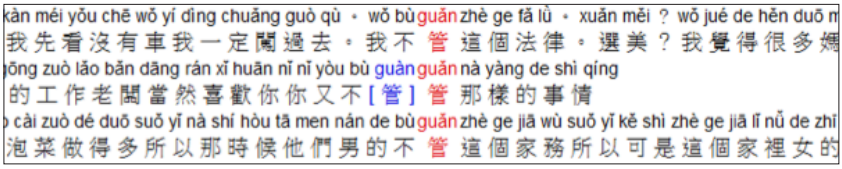


Figure 1b. Sample Error Tag in the MIC (Spoken)

In Figure 1a, the node word is 住 zhù *live* and there are instances where learners incorrectly used the character 主 zhǔ *main* for 住. On the interface, it was indicated in brackets and placed next to the node word. In Figure 1b, the character 管 guǎn *control* was mispronounced with the high falling tone instead of the falling-rising tone. The learner's mispronounced tone in brackets was placed next to the node word.

In addition to the manual encoding and cross-checking done by the research assistants and researchers, a statistical programme was written to aim at consistency of word tagging across all the data. In the programme, the analysis of the words in one text was checked against that in the database in terms of word token, text token, and the corresponding percentages of each word, as shown in Figure 2a.

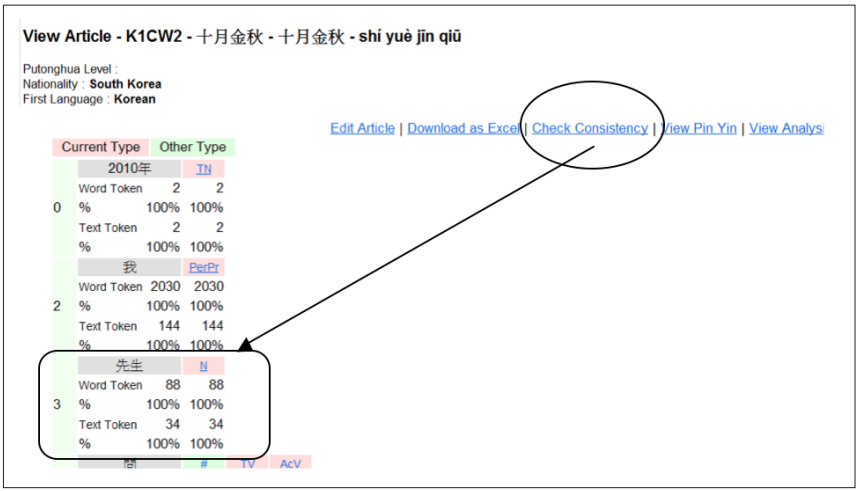


Figure 2a. Consistency Template: Word Token, Text Token, and Ratios

For every word in a text under checking, there are five rows with different pieces of tagging information: the word itself and its part of speech (first row), *WORD TOKEN* (second row), *WORD TOKEN RATIO* (third row), *TEXT TOKEN* (fourth row), and *TEXT TOKEN RATIO* (fifth row).[5] An example is the word 先生 xiānshēng, meaning *husband* in the context where it appears. In the text under checking (K1CW2), the word 先生 bears the part of speech *N* (*noun*), as indicated in the first row [先生 | N]. Under this row, we can

find the match or mismatch between this analysis with the others in the database. *WORD TOKEN* refers to the number of instances of a word in the database at the time of checking (88 tokens of 先生). Next to it is the number of instances of the word with a particular part of speech (88 tokens with the part-of-speech *N*). Given that 先生 has only one part of speech, the number of *WORD TOKEN* should be the same as that of the part of speech (i.e., 88 for 先生 vs. 88 for the part-of-speech *N*) or there was a missing tag. The third row shows the corresponding percentages: Given that all 88 tokens of 先生 were tagged *N*, the percentage is 100%, indicating consistent tagging. The next two rows present the number and percentage of texts where the word appears. At the point of checking, the word 先生 appeared in 34 texts, including the checked one, and all those tokens were labelled with the same part of speech *N*.

An alert was given to the administrator of the site when one of the following two conditions was found: (1) there was only one instance of a part of speech for a word token, or (2) there was more than one instance of a part of speech for a word token but the percentage was lower than 10%. Figures 2b and 2c help illustrate the two conditions. Condition (1) states that there is only one instance of a particular word tag and this can in turn imply wrong assignment of a part of speech to the word.

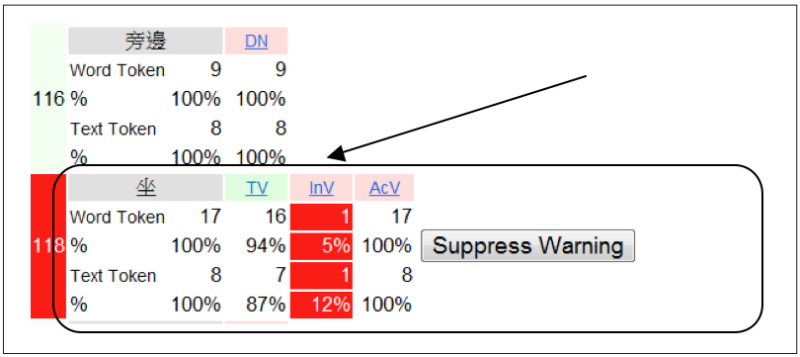


Figure 2b. Consistency Template: Word Token, Text Token, Percentages, and Inconsistency Alert—Condition (1)

As shown in Figure 2b, inconsistent tagging of the verb 坐 zuò *sit* was noted by the system. Among the 17 tokens, 16 were tagged as TV (i.e., transitive verb, 94%) and one as InV (i.e., intransitive verb, 5%), although all 17 tokens were consistently tagged as the sub-part-of-speech AcV (i.e., action verb). As a result, an alert was issued by the system. The administrators then referred to the text concerned and did subsequent editing for another round of uploading. With cross-checking done, the administrators then pressed the button *Suppress warning* to avoid confusion in any further cross-checking.

As a supplement to Condition (1), Condition (2) handles situations where there is more than one instance of the tag but the percentage is low. In Figure 2c, the word token 多 duō *many* in the text under checking was among the 18 instances in the database with the part of speech *Other particle [OtherPt]* (as in 十多個大的蟊 shíduōgè dàde měng *about 10 big locusts*). Even though there was more than one instance of 多 bearing the part of speech *Other particle* (18 instances), the percentage of the tag *OtherPt* was low (4%). Consequently, an alert was issued by the system and cross-checking ensued.

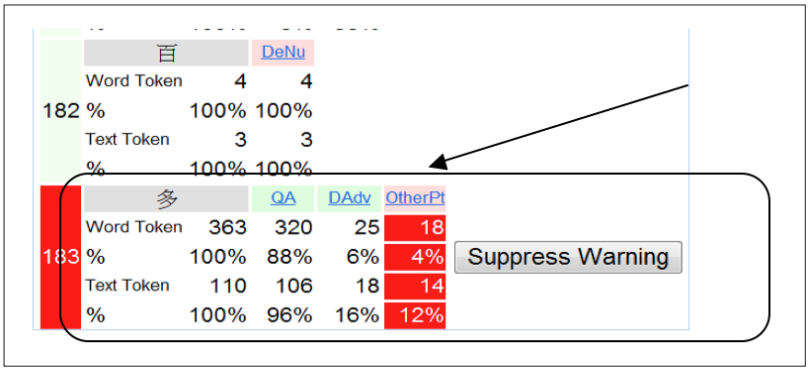


Figure 2c. Consistency Template: Word Token, Text Token, Percentages and Inconsistency Alert—Condition (2)

Database Structure of the MIC

In constructing the MIC, PHP (Hypertext Preprocessor) and MYSQL were adopted for several reasons. First, they are free of charge, which is important considering the funding and other resources available. More importantly, they can be run on every platform and provide good support to the UTF8 encoding which accommodates both Traditional and Simplified Chinese characters. Last, they are quick and light on server resource usage.

Components of the MIC Site

The MIC is made up of the two standard components of any corpus resource: the corpus and statistics. In addition to these two components, a Learners' Corner is also available for the student participants in the project. The content of the pages can be displayed in three different modes: English, traditional Chinese, and simplified Chinese.

Corpus Search Interface

The corpus search interface consists of two corpora: spoken and written. Once a user clicks on the spoken or written corpus, he or she will go into the search interface of that particular database. The search interface is designed in a user-friendly manner, with five key search options phrased in nontechnical terms: *Keyword*, *Source*, *Word category*, *First language*, and *Topic*. A user can key in any of the five options in searching for tokens in the database (see Figure 3).

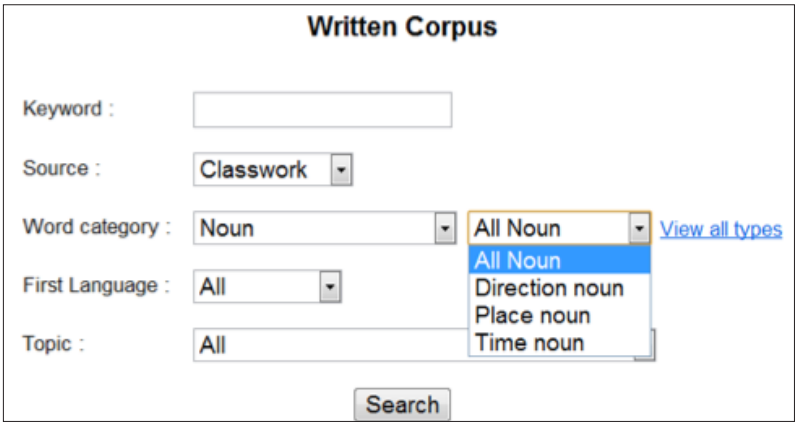


Figure 3. Search Interface of the MIC Site

Alternatively, the user can combine two or more search options so as to execute a more specific search (e.g., how action verbs [word category] are used by learners with French as the first language [first language]). Once a search is keyed in, the system will generate the output (see Figure 4).

Figure 4. Sample Output Interface of the MIC

In addition to the node word in the centre and the concordance lines, the user can refer to the original text from which the node word is taken by clicking on the *View Original Text* button next to a particular line (see Figure 5). The full text is made available to users so as to facilitate further independent and undistorted analysis of the word under examination in the context where it was used.

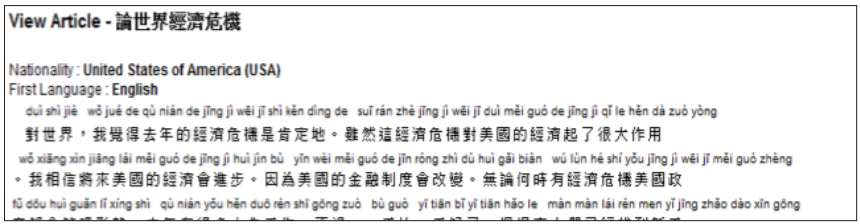
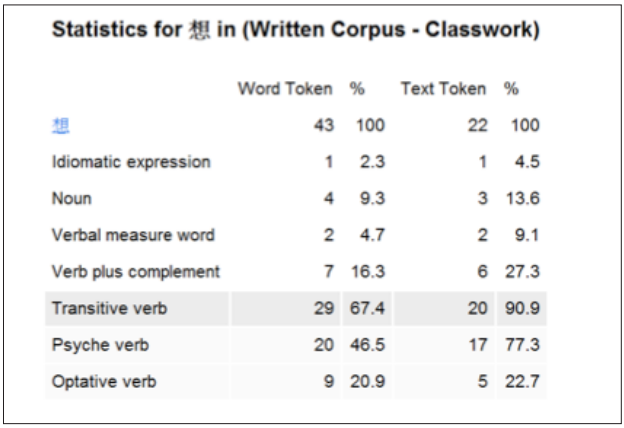


Figure 5. Sample Output Interface of the MIC (Original Text)

Statistics

On the statistics interface, the user can search for the frequency of a keyword in either the spoken or written corpus (see Figure 6). Three kinds of statistics about the word under search are given on the page: the different parts of speech assigned to the word, the *WORD TOKEN* (and ratio) and the *TEXT TOKEN* (and ratio). In the example in Figure 6 of the character 想 xiǎng, we can see the number of instances when it was used as a *transitive verb* (29 instances, 67.4%), with its sub-category *psyche verb* (20 instances, 46.5%) or *optative verb* (9 instances, 20.9%) below it, or as part of a word (e.g., in the noun 想法 xiǎngfǎ *thought*).[6]



Statistics for 想 in (Written Corpus - Classwork)

	Word Token	%	Text Token	%
想	43	100	22	100
Idiomatic expression	1	2.3	1	4.5
Noun	4	9.3	3	13.6
Verbal measure word	2	4.7	2	9.1
Verb plus complement	7	16.3	6	27.3
Transitive verb	29	67.4	20	90.9
Psyche verb	20	46.5	17	77.3
Optative verb	9	20.9	5	22.7

Figure 6. Statistics Interface of the MIC

Learners' Corner

As a way to encourage students to submit their work to the project, a Learners' Corner was constructed on this site. Unlike the corpus search interface, which is open to the public, this is a place reserved for student participants and they need to log in to get access to the exercises on Mandarin and the corresponding answers. Besides some general exercises with different types of questions, there are some questions on the specific open parts of speech (e.g., verbs and nouns) and some questions that target different first languages. So far, 100 exercises have been uploaded to the site.

Preliminary Observations

In the course of tagging the data, some preliminary observations as to the features of the learners' Chinese were made:

1. Learners, in general, tended to overuse the aspectual marker *le* and underuse the aspectual marker *guo*.
2. Classifier-language speakers and non-classifier-language speakers differed in their use of some specific classifiers, although they used the general classifier *ge* most of the time.
3. There were traces of first language (L1) influence in L2 Chinese word order, as in the position of the modifier of a place name:
   Native Chinese: Wǒ de lǎojiā shì zài Měiguó nánbù de North Carolina Charlotte.

L2 Chinese: Wǒ de lǎojiā shì Charlotte, North Carolina zài Měiguó nánbù.
English: My hometown is Charlotte, North Carolina in southern America (i.e., USA).

Some of these features have been and are being examined in more detail in Tsang (2012a, 2012b).

Challenges

In the compilation of the MIC, several challenges deserve some discussion in this report. These challenges are mainly related to data processing at the encoding level and at the technical level.

Annotation-Related Challenges

The first challenge concerns the ambiguity of the parts of speech of some words. Such ambiguity arises largely because of two difficulties. The first relates to the lack of a specific tag in the existing HSK conventions and the nature of the word as discussed in the literature. For example, 還是 háishì *still* is labelled as an adverb in the HSK system. However there is no mention of what kind of adverb it is, and different subcategories have been suggested in the existing key literature on Mandarin grammar (e.g., *scope adverb* or *mood adverb*). The other difficulty is associated with the usage of a word in the context. Take 沒有 méiyǒu *not yet* as an example. It can be one lexical unit functioning as a negative adverb as in (1) or two lexical units as a negative adverb plus an existential verb as in (2):

(1) 我沒有打瞌睡 (沒有 méiyǒu as an adverb)
wǒ méiyǒu dǎkēshuì
I not yet fall asleep
"I haven't fallen asleep yet."

(2) 我身上沒有錢 (沒有 méiyǒu as an adverb plus a noun)
wǒ shēnshang méi yǒu qián
I body on no money
"I have no money with me."

To arrive at a correct tag for tokens such as 還是 háishì *still*, the research team looked at all the contexts collected in the first 3 months and decided on one sub-part-of-speech of the word to be used consistently for future analyses. As to the second difficulty (i.e., as exemplified by 沒有 méiyǒu *not*

yet), each context in which the word appeared was carefully examined and cross-checked to find out which pattern the word fell into.

One more challenge was about marking participants' output at different periods of the course in terms of the corresponding level of Mandarin. As stated earlier, the student participants were asked to complete a placement test, which helped to identify their level of Mandarin at the outset of data collection. Besides this indicator of their Mandarin proficiency, the end-of-course examination was taken as another available indicator. While the examination output could be clearly marked with the participants' levels of Mandarin, the output produced as coursework has yet to be marked in terms of participants' proficiency. As a makeshift solution, all the data are currently marked with the time when the participants produced the work and the number in the code (e.g., CW1 earlier than CW2).

Technical Challenges

As well as the annotation-related issues, the research team needs to handle two challenges in the technical realm. First, the team is actively exploring the best way to upload actual sound files to the database site, taking into account the need for the anonymity of the participants in the audio clips, an economical budget, the size of the clips, and the availability of space on the server. The other technical challenge concerns the two conditions for the consistency check function. A statistical programme is deemed necessary to show the reliable thresholds which can further enhance the checking of tag consistency across a bulk of data.

Conclusion

With the intention to overcome the above challenges, the MIC project hopes to contribute to the learning of Mandarin Chinese as a foreign language from both theoretical and pedagogical perspectives. Theoretically speaking, the learner corpus can serve as a small-scale database that facilitates and stimulates research on both corpus linguistics and different fields of applied linguistics. Pedagogically, the database is expected to enhance the learning and teaching of Mandarin as a foreign language.

Regarding the area of corpus linguistics, the inclusion of both written and spoken output will shed light on the development of two-mode corpora, as well as that of spoken learner corpora and learner corpora in general, especially in terms of the nature of learners' output and the compilation process or technology concerned. The database can also contribute to different fields

of applied linguistics, with its systematic and detailed database of Mandarin language output from learners of different L1s. In particular, the availability of both written and spoken data in the database can help investigation or comparison of learners' output in terms of different parameters, such as L1s, levels of proficiency, language patterns (especially those contrasting Mandarin and L1s of different learners), genres, contexts (e.g., in-course vs. examination), and modes of output (written vs. spoken).

Pedagogically, students should find the Learners' Corner particularly useful. They can choose to work on some general exercises applicable to learners of any first language or some exercises targeted at different types of Chinese learners (e.g., Japanese learners of Chinese in Japan or Mainland China). Meanwhile, language teachers and course or textbook developers of Chinese as a second or foreign language will be provided with a resource where they can examine learners' Mandarin in a systematic and focused context. Drawing on information from the database, they will be in a better position to plan curricula and design teaching and assessment materials. This in turn will help explore further the role of corpora in language teaching.

Finally, it is intended that the MIC will serve as a preliminary database leading to a bigger resource with more variables covered, especially in terms of the number of language samples and participants with different L1s.

Notes

1. The authors would like to express their gratitude to one of the reviewers for information about the BLCU Chinese Interlanguage Corpus, the Modern Interlanguage Chinese Corpus, and some corpora which are currently under construction.
2. This kind of occurrence leads to potential difficulties in attaining corpus balance in different areas such as proficiency levels or observations for each language point. One possible solution to this problem is that the data be analysed in terms of proportions rather than absolute frequencies. Meanwhile, it is intended that more resources will be generated and learners be recruited in the future to arrive at a bigger and more balanced pool of learners' data.
3. At the time of submission of this report, the lowest score on the placement test was 29 and the highest 43.

4. Data from Chinese heritage learners were not included in the database because their exposure to the target language is different from that of other learners of Chinese.
5. Each word in a text was given a number as the identifier, which is stated in the left column (e.g., *3* for the word 先生 *husband*).
6. In reading the numbers under the column TEXT TOKEN, it should be noted that a word can appear more than once in a text with different parts of speech, for example, 想 as an optative verb and as a noun, in two different parts in one text. Therefore, the text token at the top of the column is not the sum of the specific text tokens beneath it.

Acknowledgement

The MIC project is supported by a grant under Small Project Funding from the University Research Committee of the University of Hong Kong (Project no.: 200907176041).

References

Beijing Language and Culture University. (2000). *汉语8000词词典* [A dictionary of Chinese usage: 8000 words]. Beijing: Author.

Beijing Language and Culture University. (2003, 2009). *HSK动态作文语料库* [The Inter-Media HSK Essay Corpus]. Beijing: Author.

Cao, W., & Zhang, J. S. (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 construction of a CAPL Chinese interlanguage corpus and its annotation]. *语言文字应用* [Applied Linguistics], *4*, 122-131.

Cheung, S. H. N., Liu, S. Y., & Shih, L. L. (1994). *A practical Chinese grammar*. Hong Kong: Chinese University Press.

Feng, Z. W. (2006). Evolution and present situation of corpus research in China. *International Journal of Corpus Linguistics*, *11*, 173-207.

Gottlieb, N. (2012). *Language policy in Japan: The challenge of change*. Cambridge: Cambridge University Press.

Hu, X., Williamson, N., & McLaughlin, J. (2005). Sheffield Corpus of Chinese for diachronic linguistic study. *Literary and Linguistic Computing*, *20*, 281-293.

Huang, C. R., & Chen, K. J. (1992). A Chinese corpus for linguistics research. In the *Proceedings of the 1992 International Conference on Computational Linguistics (COLING 1992*) (pp. 1214-1217). Nantes, France.

Huang, N. S., & Sun, D. J. (2000). *HSK词语用法详解* [A guide to the usage of HSK vocabulary]. Beijing: Beijing Language and Culture University Press.

Jia, M. (2006). 国内语料库语言学研究述评 Corpus linguistics on the mainland. *阜陽師範學院學報(社會科學版)* [Journal of Fuyang Teachers College (Social Science)], *113*(5), 65-66.

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

Liu, X., Zhang, K., Liu, S. H., Chen, X., Zuo, S. D., & Shi, J. W. (2002-2005). *新实用汉语课本 (一至五册)* [New practical Chinese reader (Books 1-5)]. Beijing: Beijing Language and Culture University Press.

Maher, J. C. (1995). The kakyo: Chinese in Japan. *Journal of Multilingual and Multicultural Development*, *16*, 125-138.

McEnery, T., & Xiao, R. (2004). The Lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In M. Lino, M. Xavier, F. Ferreire, R. Costa, & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 1175-1178). Lisbon, Portugal: Centro Cultural de Belem.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Ming, T., & Tao, H. Y. (2008). Developing a Chinese heritage language corpus: Issues and a preliminary report. In A. W. Y. He & Y. Xiao (Eds.), *Chinese as a heritage language: fostering rooted world citizenry* (pp. 167-187). Honolulu: National Foreign Language Resource Center, University of Hawai'i.

Shao, J. M. (2007). *现代汉语通论* [An introduction to modern Chinese] (2nd ed.). Shanghai: Shanghai Educational Publishing House.

Tao, H. Y., & Xiao, R. (2007). *The UCLA Chinese Corpus*. Lancaster, UH: UCREL.

Tsang, W. L. (2012a, June). *L2 acquisition of Mandarin classifiers: How distinct are classifier-language learners from non-classifier language learners?* Paper presented at The 24th North American Conference on Chinese Linguistics at the University of San Francisco, CA.

Tsang, W. L. (2012b). Aspectual marking among English and Korean learners of Mandarin Chinese. *Chinese as a Second Language Research*, *1*, 1-32.

Wang, Y. J., & Li, J. M. (2001). 建立汉语中介语语音语料库的基本设想 [Our tentative ideas concerning the establishment of a corpus of Chinese interlanguage speech]. *世界汉语教学* [Chinese Teaching in the World], *55*(1), 87-92.

Xing, J. Z. (2006). *Teaching and learning Chinese as a foreign language: A pedagogical grammar*. Hong Kong: Hong Kong University Press.

Yang, J. (2008). 近十年國內語料庫語言學研究中的若干問題綜述 Issues in corpus linguistics on the mainland in the last 10 years. *湘潭师范学院学报（社会科学版）*[Journal of Xiangtan Normal University (Social Science Edition)], *30*(1), 105-107.

Yang, X. J. (2006). Survey and prospect of China's corpus-based research. In A. Wilson, D. Archer, & P. Rayson (Eds.), *Corpus Linguistics around the World* (pp. 219-232). Amsterdam: Rodopi.

Yang, Y., Li. X. L., Guo, Y. W., & Tien, Q. Y. (2006). 建立汉语学习者口语语料库的基本设想 [Tentative ideas of constructing Chinese learners' spoken corpus]. *汉语学习* [Chinese Language Learning], *3*, 58-64.

Zhan, W. D., Chang, B. B., Duan, H. M., & Zhang, H. R. (2006). Recent developments in Chinese corpus research. In the *Proceedings of the 13th NIJL International Symposium*, Tokyo. Retrieved from http://ccl.pku.edu.cn/doubtfire/papers/2006_Corpora_NIJL_Workshop.pdf

Zhang, J. Q. (2005). 关于建立“普通话中介语语音语料库”的设想 [Ideas about constructing Putonghua interlanguage speech corpus]. *广西梧州师范高等专科学校学报* [Journal of Wuzhou Teachers College of Guangxi], *21*(2), 46-49.

Zhou, Y. J. (2007). 語料庫語言學的應用及其在中國的發展趨勢 [The application of corpus linguistics and its growing trend in China]. *齊齊哈爾大學學報. 哲學社會科學版* [Journal of Qiqihar University (Phi & Soc Sci)], *3*, 138-140.

Appendix

MIC Tag Set

A	Adjective	形容詞	形容词
QA	Qualifying adjective	性質形容詞	性质形容词
DA	Descriptive adjective	狀態形容詞	状态形容词
NPA	Non-predicate adjective	非謂語形容词	非谓语形容词
Adv	Adverb	副詞	副词
DAdv	Degree adverb	程度副詞	程度副词
ScAdv	Scope adverb	範圍副詞	范围副词
TAdv	Time adverb	時間副詞	时间副词
NegAdv	Negative adverb	否定副詞	否定副词
MdAdv	Mood adverb	語氣副詞	语气副词
Conj	Conjunction	連詞	连词
IE	Idiomatic expression	習慣用語	习惯用语
Int	Interjection	歎詞	叹词
M	Measure word	量詞	量词
Cl	Classifier	個體量詞	个体量词

UM	Unit measure word	度量詞	度量词
VM	Verbal measure word	動量詞	动量词
N	Noun	名詞	名词
DN	Direction noun	方位詞	方位词
PN	Place noun	處所詞	处所词
TN	Time noun	時間詞	时间词
Nu	Numeral	數詞	数词
CaNu	Cardinal number	系數詞	系數詞
DeNu	Decimal number	位數詞	位數詞
OrNu	Ordinal number	序數詞	序數詞
Ono	Onomatopoeia	象聲詞	象声词
Pr	Pronoun	代詞	代词
PerPr	Personal pronoun	人稱代詞	人称代词
DemPr	Demonstrative pronoun	指示代詞	指示代词
QPr	Question pronoun	疑問代詞	疑问代词
Pref	Prefix	詞頭	词头
Prep	Preposition	介詞	介词
Pt	Particle	助詞	助词
AsPt	Aspect particle	動態助詞	动态助词
StPt	Structural particle	結構助詞	结构助词
MPt	Modal particle	語氣助詞	语气助词
OtherPt	Other particle	其他助詞	其它助词
SFPt	Sentence final particle	句末語氣詞	句末语气词
DeSFPt	Declarative SFPt	陳述句末語氣詞	陈述句末语气词
IntSFPt	Interrogative SFPt	疑問句末語氣詞	疑问句末语气词
ImSFPt	Imperative SFPt	祈使句末語氣詞	祈使句末语气词
ExSFPt	Exclamatory SFPt	感嘆句末語氣詞	感叹句末语气词
Suf	Suffix	詞尾	詞尾
V	Verb	動詞	動詞
TV	Transitive verb	及物動詞	及物动词
InV	Intransitive verb	不及物動詞	不及物动词
AcV	Action verb	動作動詞	动作动词
ExV	Existential verb	存在動詞	存在动词
DirV	Directional verb	趨向動詞	趋向动词
CauV	Causative verb	使令動詞	使令动词
PsyV	Psyche verb	心理動詞	心理动词
CopV	Copula verb	判斷動詞	判断动词
OpV	Optative verb	能態動詞	能态动词
VC	Verb plus complement	動補式動詞	动补式动词
VO	Verb plus object	動賓式動詞	动宾式动词