Some Pragmatic Issues in Learner Corpus: A CSL Perspective

Weiping Wu

Abstract Unlike elements in language structure (phonology, semantics, and syntax), factors related to language use are much more difficult to handle and are often neglected, or simply ignored, in the construction of a corpus. For example, how to design the tasks and prompts while gathering oral samples so that pragmatic factors become an integrated part of the data collected? Instead of treating pragmatic issues as some "extra elements" to be identified after the samples are collected while building the corpus, the author presents a systematic approach in which pragmatic factors are treated as part of the design before the construction of the corpus. Both theoretical framework and specific steps taken in the implementation are discussed in this paper in the context of understanding and using pragmatic knowledge in oral communication. All examples used are from the Language Acquisition Corpus constructed with oral productions by CSL learners of various language and cultural backgrounds.

Keywords Learner corpus · Corpus construction · Pragmatic factors · CSL learning

1 Introduction

This paper explores the role of pragmatic factors in the construction of corpus, focusing on how a systematic approach can be applied in obtaining oral productions of CSL learners from different language and cultural backgrounds and how to organize the data obtained in the corpus. Because of the availability of pragmatic clues associated with the data, such a corpus can then provide opportunities for studies related to L2 production beyond the structure of the language.

To provide a larger context in which we discuss this CSL learner corpus, let's take a closer look at this area of linguistic research. In terms of language, the majority

W. Wu (✉)
Center for Linguistics and Applied Linguistics (CLAL), Guangdong University of Foreign Studies (GDUFS), Guangzhou, China
e-mail: 202070007@oamail.gdufs.edu.cn

H. H.-J. Chen et al. (eds.), *Learner Corpora: Construction and Explorations in Chinese and Related Languages*, Chinese Language Learning Sciences,
https://doi.org/10.1007/978-981-19-5731-4_3

of the corpora and related studies so far are predominantly centered around English, including some of the most widely used online corpora, such as the Global Web-based English, the Corpus of Contemporary American English, Corpus of Historical American English, the TV Corpus, the Movie Corpus, and the British National Corpus. (cf. https://www.english-corpora.org/). When it comes to learner corpus, studies reported in the Journal of Learner Corpus Research, among others, can provide us with a glimpse of what is going on in this field, especially the special issue in which the editors (Brezina & Flowerdew, 2019) put together some of the impressive studies related to the Trinity Lancaster Corpus.

Studies related to the Chinese language, on the other hand, are still few and far between compared with what has been achieved in English, even though rapid progress can be seen in recent years. Among some of the popular ones are various corpora as listed online (cf. https://www.cncorpus.org), those maintained by the Academic Sonica in Taiwan (cf. http://www.sinica.edu.tw/SinicaCorpus/), as well as others that seem to focus on specific areas of language use, like the MLC (by the Chinese University of Communication, cf. http://ling.cuc.edu.cn/RawPub/). Due to the availability of data from large-scale proficiency tests in the past decades (e.g., the HSK, which is a proficiency test taken by hundreds and thousands of CSL learners from all over the world who want to enter Chinese language programs in universities in China), learner corpus for CSL has been developing very quickly (Chen & Tao, 2019; Tao, 2017; Tao et al., 2020; Zhang & Tao, 2018; Zhang et al., 2019). Other learner corpora similar in nature include the BCC Corpus by Beijing Language and Culture University and the CCL Corpus by Beijing University. The Language Acquisition Corpus focusing on Spoken Chinese (a.k.a. LAC/SC) to be discussed below is unique because of the availability of information related to pragmatic factors for each oral sample in the corpus.

In reference to the corpora mentioned above, the CSL learner corpus based on oral production is still at the very initial stage of its development path compared with those based on data from ESL. The LAC/SC now provides direct access to the original sound files for each of the oral productions as well as a clean version of the written transcription in Chinese characters. It is hoped that such a model with built-in pragmatic factors can contribute to narrowing the gap between ESL and CSL studies based on corpora.

The discussion below will be divided into four parts, each of which is briefly described here to provide an overall picture. The next part will explain two concepts behind the construction of the LAC/SC, one being the distinction between language structure (LS) and language use (LU), the other, whether the final goal of all learning activities is "appropriate culturally" or just "correct structurally". The third part of this paper describes the structure of the LAC/SC, including what we mean by pragmatic factors and how they are identified and dealt with in the process of corpus construction. Problems met, and possible solutions applied, are discussed in the fourth part, covering data eliciting procedure, task design, and measures taken to guarantee adequate comprehension of tasks by L2 learners. The final part of this paper offers some concluding remarks, representing our current understanding of creating and implementing the pragmatic framework in building a CSL spoken corpus.

2 Two Fundamental Concepts Behind the CSL Learner Corpus

Understanding the concepts behind the two distinctions discussed here is key to understanding the logic and reasoning behind the construction of the LAC/SC. In the distinction we make between language structure (LS) versus language use (LU) in L2 teaching and learning, we propose that LU be viewed as a system of systems, consisting of three key components: Interlocutors, Setting, and the Timing of the communication event (or Purpose if clues for timing are not available). Each of these components can be further divided into sub-categories. We argue that such a view is comparable to the way we view LS, which is also a system of systems consisting of phonology (Sound), semantics (Words), and syntax (Grammar). For the convenience of discussion, we will use the following abbreviations and equations to represent these two systems:

$$\mathrm{LU} = \mathrm{I} + \mathrm{S} + \mathrm{T/P}$$

$$\mathrm{LS} = \mathrm{Pho} + \mathrm{Sem} + \mathrm{Syn}$$

In the second distinction we make between being "structurally correct" versus "culturally appropriate" in reference to the final goal of language learning, we believe that all CSL learners' production for communication purposes should be the latter and not the former. That means one step is missing between the final goal and most of our current curricula and teaching practices, most of which seem to stop when students "understand" what is being taught and their productions are correct in terms of LS.

It is not surprising to find such a reality in the CSL field because, in various subfields under the general heading of language teaching and learning, the focus of attention has been overwhelmingly on the structure of language (Chao, 1968; Lado, 1957; Wang, 2010; Wu, 1993). In recent years, we started to hear calls for attention to pragmatic ability in discussions related to CSL teaching, second language acquisition, and pedagogy (Ran, 2004; Rose & Kasper, 2001; Wang, 2006; Wu, 2006, 2016). Common sense would tell us that people call for attention means there is a lack of attention. As pointed out by Li in a recent interview (Li, 2021), most of the attention in linguistic studies in China was on the research on language structure and, by comparison, neglecting the real situations in actual language use. Teaching materials preparation with various vocabulary lists, grammar points, and sentence pattern lists can serve as typical examples of such focused attention. For many years, teaching activities within any language learning program tend to center around the explanation of grammar points, which also indicates that the point of attention is on language structure. Various tests in the teaching and learning process, such as the common practice of a "quiz" on grammar after each lesson, as well as many proficiency tests (Clark & Li, 1986; Ke, 1994; Li, 1997; Liu, 2008; Xiong et al.,

2002) that are not supposed to be closely related with any particular curriculum, are often designed with three key components of the LS: pronunciation, vocabulary, and grammar.

Corpus construction in recent years, similarly, follows the same general direction, with tagging of grammatical categories and errors based on deviation from standard pronunciation and grammar rules. This is certainly understandable because research on language structure has been long and many. Moreover, the basic structures of any language are always the starting point of a learning program. How can CSL learners use Chinese if they don't know the pronunciation of a word, what it means, and how to use relevant grammatical rules to put word strings together when they speak?

Once we are out of the classroom and out of the school, once L2 learners get into real communication in real life with real people for meaningful exchange of ideas, however, problems arise. When we come face to face with scenarios in our daily life, we realize what we need to have meaningful and smooth communication is way beyond the knowledge of language structure. We have to consider and remember, unlike native speakers who usually do that without thinking, who we are and to whom we are talking, where we are, and why we are talking at that particular moment. These are the basic elements of communication. Proper understanding and application of such elements will contribute to the communicative ability of the language user. Careful analysis of any communicative event tells us that issues related to these can be grouped into categories, which can be related to LS but are not part of the LS. It reiterates the points we made above, and somewhere else (Wu, 2006, 2008a, b, 2019, 2020), that Pragmatic Factors are what native speakers can intuitively make use of when they talk, but L2 learners cannot due to the lack of such intuition. So telling CSL learners what these factors are is a duty that teachers cannot avoid.

Now let's return to the final goal for all L2 learners, being appropriate culturally versus being correct structurally. Obviously, the former must include the latter but not vice versa. To use a metaphor here, where is the finishing line in the school language program if we treat all L2 learners as athletes participating in the marathon? Although no one would openly deny that all our language teaching activities should aim at the application of knowledge for real communication, and not just "finishing the teaching tasks" as required by the curriculum, it is also hard for any of us to deny that the reality in most cases is still "doing the teaching job" as required by the curriculum, which is unfortunately still largely if not totally based on structure. To go the extra mile from being correct to being appropriate requires too much extra efforts and too much resources.

As a result, it is not uncommon to see CSL learners at the higher end of the proficiency level produce utterances that are correct in terms of pronunciation and grammar, but culturally not appropriate in real communication with real people, thus failing the very purpose of communication. There are many examples from the data collected for the LAC/SC to illustrate this. On the discourse level, the absence of a formal greeting to show respect in a formal setting, for instance, is a case in point. As cited in a research based on the corpus (Fan, 2018), out of 15 oral productions

by advanced CSL learners in a formal setting, 10 of them (2/3) did not start properly when they made a speech as a representative on behalf of a delegation, here is one of them:

e.g.1 (Note: First utterance of the speech, absence of any greeting)
我們-也-知道-我們-來-到-這邊-會-麻煩-你們-啊 ……
Women-ye-zhidao-women-lai-dao-zhebian-hui-mafan-nimen-ah
We-also-know-we-come-particle-here-will-bother-you-particle……
We also knew that we would bother you when we came here

(LAC/SC sample id: Kw0129-SS008)

This is of course a very polite way of saying things, but certainly, it should not be the first utterance when you start talking! More examples of similar nature and relevant discussions along this line can be found in the research reported in a Ph.D. dissertation based on the LAC/SC (Fan, 2018).

3 Pragmatic Framework and Its Application in Corpus Construction

How to implement pragmatic factors in the construction of the corpus? Earlier, we have identified the three essential categories (I, S, and T/P) under LU, which jointly contribute to the appropriateness of oral production by CSL learners. We recognize that not every communicative event has obvious clues to these categories. For example, clues for the timing of the communication, or timing as a factor, are sometimes missing if it is not crucial in that particular communication event. In such a case, P (purpose) can often be used alternatively to fill in the gap. Findings from sociolinguistic research tell us that appropriateness in communication by native speakers is not by chance, but by the speaker's thorough understanding of these essential factors in communication and the social rules, most of which are oblivious to L2 learners. As teachers, we need to tell our students, like what we tell them in their learning process about the structure of the language, and let them know what these elements are. One way to do this, as we did in the construction of the LAC/SC, is to make explicit relevant information about all the three categories, which in most cases tend to be "understood or inferred" by native speakers.

Like grammar rules governing sentence formation and word selection in LS, there are rules governing the choices we make in LU, including pronunciation, vocabulary, and grammar. To make available clues for these three categories for CSL learners will therefore help them make the right choice like native speakers. The use of the polite form "nin" in Chinese instead of "ni", for example, is governed by the LU rule related to the I category. It is a two-way distinction in which the polite form "nin" is used when the speaker knows he is "talking up". We use L→H in the Pragmatic Framework to indicate such a relationship, which can be further clarified as follows:

Relationship among Interlocutors:

We can use L to stand for Low, and H for High, as an indication of status. For CSL learners, we can introduce three types of relationships, among friends (L→L or H→H, in which both parties are equal), from subordinate to superior (L→H), or vice versa (H→L), in which the speaker should be aware if he is talking up or talking down, and choose the polite form in the former case. In the Chinese culture, there are some common examples in the L→H relationship:

Age: to someone of your parent's or grandparent's age.

Social status: to someone with a higher social status, such as your boss, your teacher, someone with a higher rank in the official or social hierarchy, etc.

In the S category, we can introduce a three-way distinction: informal, formal, and ceremonial or ritual. To borrow the example from another study (Feng, 2018), the choice of words in each situation may differ even if the meaning to be expressed is the same, as indicated below:

Informal: use "pian", as in "pianren (騙人)"

Formal: use "qipian", as in "qipian laoshi ren (欺騙老實人)"

Ceremonial: use "qi", as in "chengbuqiwo (誠不欺我)"

In the T/P category, the situation is a bit subtle in comparison to the other two categories, and harder for CSL learners to grasp. We can call it a two-way distinction because, in contrast to "the right moment", there is the "wrong moment". Even if you observe rules on interlocutors and settings, what you say may still be inappropriate if you choose the wrong moment to talk. If you want to propose a toast at the dinner table, for example, you will have to wait for your turn. Doing it too early or too late may render your toast inappropriate in the Chinese culture, no matter how polite you may be. This is perhaps most difficult for CSL learners because it is not something we can spell out for them as we do in the I and S categories. Being able to do this requires the knowledge and skill that even native speakers are not sure of from time to time. This is an area that is waiting to be explored and, before we can identify and find a way to explain and label what we "feel", the best way of doing things at the present is to raise a flag in the mind of all CSL learners, with the hope that such a flag will help them understand what may go wrong in their communication.

In each of the three categories given above, there are of course many more layers in each of the sub-categories. The Setting category may have varieties in different situations, such as semi-formal between informal and formal. To make it easy for CSL learners who participated in the data-collecting process, however, we limited the variations and just draw their attention to the existence of pragmatic factors known as I, S, and T/P.

To practice what we preach, we made every effort to include the I, S, and T/P clues while designing the tasks, which cover a wide variety of content areas with calculated degrees of difficulties. Responses to these tasks were obtained from participants and rated with confirmation to LU rules in mind, in addition to factors covered under

language structure. For instance, two very similar response samples from the same task to "express thanks at the farewell party" would be rated differently simply because one of them has no greetings at the beginning, even all other aspects (ideas expressed, complexity of vocabulary used and grammatical structure employed, etc.) are very similar.

Most of the oral productions were collected using the testing format. That means learners provided their responses either while taking the exam for real or in the situation in which a test is simulated. As mentioned above, pragmatic factors were used as the key criteria in the assessment of the proficiency level. Inclusion of pragmatic factors is actually a common practice in many well-recognized oral proficiency assessment tools, such as the Oral Proficiency Interview (OPI) by the American Council on the Teaching of Foreign Languages (ACTFL), or the Simulated Oral Proficiency Interview (SOPI) by the Center for Applied Linguistics (CAL), which pioneered large-scale oral proficiency assessment in the early 1980s.

Now let's take a closer look at some specific tasks used in the data-collecting process for the LAC/SC and see how information related to I, S, and T/P was included. Pasted below is an example of a language task used as part of the oral proficiency test. It is a task to elicit the spoken production from CSL learners with English L1 background, aiming at CSL learners whose proficiency level is expected to be at the Superior level according to the ACTFL proficiency guidelines (https://www.actfl.org/resources/actfl-proficiency-guidelines-2012). Please note that all the pragmatic factors related to LU are underlined.

e.g.2 Task description in English (for learners whose L1 is English)

You are at a farewell party given by your host organization in China for a group of teachers from your school, of which you are the leader. After the host makes a speech thanking you for the job you have done, you are invited to say a few words of thanks on behalf of all the teachers. During your one semester teaching in China, the host organization has been very helpful in many ways, making arrangements for accommodation, providing opportunities for teacher-student communication, doing the best they could to facilitate your teaching, and so on. Now think about what you want to say in this formal situation. After your Chinese host's introduction, respond on behalf of your group, expressing your appreciation for the hospitality of your host organization, acknowledging any inconvenience your group may have caused, and offering to reciprocate their hospitality. As in a formal speech, end your talk with a toast.

Prompt in Standard Chinese (Mandarin):

各位來賓, 現在我們請貴方的代表給我們講話。

Referring to the underlined parts above, which provide the pragmatic factors that are also part of the assessment criteria for the oral productions elicited by this type of task, we can now fill in the content of what I, S, and P stand for:

Interlocutors (clues for the I category):

Who you are: leader of a teacher delegation.
To whom you are speaking: the host and his/her team.

Setting (clues for the S category):

Formal situation (farewell party)
In public

Purposes of communication, with reference to content (clues for the T/P category):

Wait for your turn and
Speak on behalf of your team

This is what we mean by making explicit clues for LU, so that the speaker will use such a framework in their oral production, with appropriateness as one of the aims. In addition to the specific description telling you that this is a formal situation where former words and certain ritual in public speech are expected, clues of a formal situation are also given in the Chinese prompt (such as "gewei" and "guifang"), which serves as the final reminder before the speaker starts talking. We all understand that any prompt can elicit a response in such a situation, even if you simply say "now start talking". Providing contextual clues in language assessment, nevertheless, has now become the hallmark of the communicative approach in testing, also called Stage III in the history of assessment (Li, 1997; Douglas, 2000, Wu, 2008b). Test items or prompts in the Stage I period (coincided with a focus on LS in language teaching) will not bother to provide any contextual clues, such as quoted below:

e.g.3 Sample of prompt for oral test in Stage I assessment

"Talk about the most unforgettable person in your life"

Giving a speech in public in formal settings is designed for CLS learners at the advanced level. For those with a lower proficiency level, the task below will be more appropriate for eliciting their oral production.

e.g.4 Task description in English (for learners whose L1 is English)

You have a friend from Xiamen who likes reading a lot. Please tell him what types of book you like and what books are worth reading. Please think about it and answer after listening to the question in Mandarin.

Prompt in Standard Chinese (Mandarin):

你呢?你喜歡看什麼書呢?

We can see here that, similarly, pragmatic factors are also provided and can be summarized using the same categories above:

Interlocutors:

Who you are: a friend of somebody you talk to
To whom you are speaking: a friend to whom you speak

Setting:

informal situation

Purposes of communication, with reference to content:

Small talk on personal hobbies

Comparing the two task descriptions above, we start to see how a systematic approach in providing pragmatic factors in LU is implemented from the very beginning in the process of corpus construction before oral productions were collected. It was expected that the conditions set in each of these tasks under each of the three categories would produce data more appropriate for studies focusing on LU. If more and more research in this direction could be carried out from time to time, it would eventually contribute to a better understanding of LU as a system similar to LS.

For the past decades, studies on each of the subsystems under LS have produced a multitudinous amount of literature, most of which are somewhat related to theories in structuralism as a school in linguistics. Language teaching and learning as a field has benefited tremendously from such studies. Although impossible to quantify, we can speculate based on common sense that, if only one percent of the efforts for linguistic research from now on could focus on LU and its subsystems, we would soon understand much more about how language teaching and learning can benefit from sociolinguistic research.

The second step taken was to deal with the challenge that all task descriptions should be clearly understood by the speaker. Using L1 of the CSL learner in the testing formal is a practice in SOPI. Such a format has been recognized as one of the solutions to cover all participants, including those at the lower end of the proficiency spectrum. Over the years, this approach has attracted many controversies. For learners with advanced proficiency levels, it would be more efficient to use L2 in the whole process, involving no translation and thus avoiding possible misunderstanding of the requirement regarding pragmatic factors. Referring to the different stages in the history of assessment as an academic field, we can see that using L1 is determined by the development of testing theories and practices. Most of the assessment tools now are, or claim to be, communicative in nature. The aim is therefore to assess the ability of the learner in using language for communication, focusing on LU, rather than taking stocks of their knowledge of the language, focusing on LS. In order to achieve this goal, a simple request such as “circle the right answer” will not work.

The complexity of test instructions focusing on LU, and the description of the tasks to be performed, requires much more in terms of the grammatical structure, the difficulty level of the words involved, and the discourse features related to cultural beliefs and practices. Learners at the lower end of the proficiency level will not be able to understand the task descriptions with all the contextual clues in the language they are learning. Looking back, we can see the use of L1 became the alternative in

L2 tests at the beginning, and later the norm of large-scale assessment, especially in assessment tools that cover the whole spectrum of proficiency level (from novice to distinguished according to the ACTFL proficiency guidelines, or A1 to C2 according to the CEFR).

With the two steps taken, we found in the deep briefing after the data-collecting that the requirement of the tasks was well understood by the participants. At least they were aware that the speaker was "talking up" or "talking down", the setting was formal or casual. Whether or not they could adjust their oral production to match such pragmatic factors, however, would then depend on their own proficiency level.

4 Structure of a CSL Learner Corpus with Built-In Pragmatic Factors

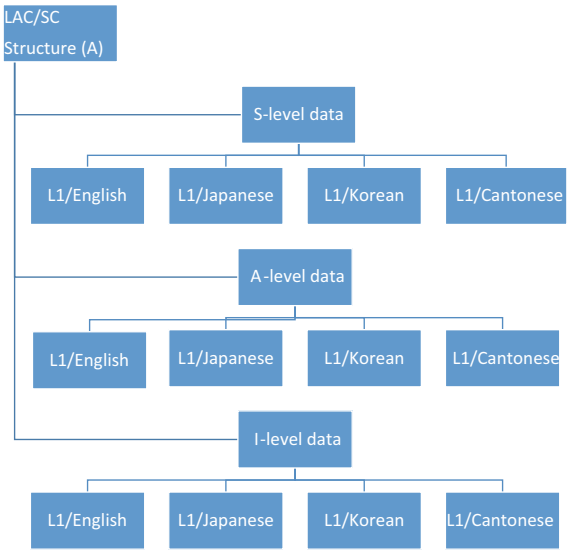
Once all the CSL leaner contributions were collected, two trained assessors would cross-rate each and every sample to determine the proficiency level. If there was a major difference in the ratings, a third assessor would be called in to have the final rating. All rated samples are then put under the LAC/SC structure and transcribed.

Bearing in mind the two distinctions mentioned previously (LS vs. LU, correct vs. appropriate), we can go on to provide an overall picture of the LAC/SC with reference to these distinctions. Since most of the CSL learners at that time were from Japan, Korea, and English-speaking countries, data collection was conveniently grouped according to the L1 of each group. For comparison, similar data were also collected from local students whose L1 is Cantonese. One of the reasons for adding this group was to see the differences and similarities in pragmatic ability between this group, which is a subset of the Chinese language and culture, and the other three groups, which were not part of the Chinese language and culture family. Such an addition turned out to be very helpful later on because it provides one more dimension in L2 acquisition studies: the comparison of the in-group versus the out-group in their understanding of Chinese cultural concepts and practices, as demonstrated by the degree of appropriateness in their oral productions. It also shed light on why pragmatic factors are important and should not be neglected in the corpus construction process.

To fully explain the organization of the corpus, we can trace the path of our construction process from the very beginning, when we were still trying to decide on the top layer of the corpus. There were two possibilities, as illustrated by the two figures below. In structure A, the top layer is the proficiency level of the oral production data, while learners with different L1 backgrounds are grouped beneath, under each proficiency level.

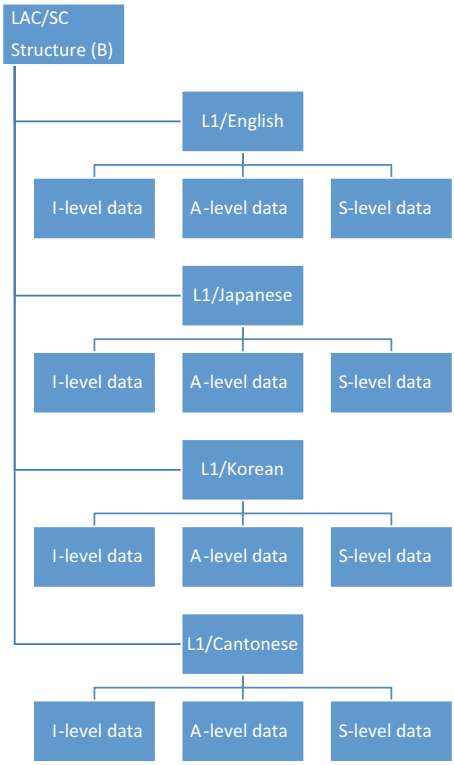
Structure A: CSL Learner Corpus based on the proficiency level

(Note: Proficiency levels used here in the corpus are based on the ACTFL Proficiency Guidelines, where "I" stands for Intermediate, "A" for Advanced, and "S", Superior).



As shown in Structure A, there are three levels in the chart, each covering all oral production data from different L1 groups at a specific level. These three levels are fixed and cannot be changed and is therefore a closed system. Under each of them, however, it is an open system that allows block building. In this chart, we have data from CSL learners whose L1 is English, Japanese, Korean, and Cantonese, respectively, but we can see from the structure that, should we have data from other L1 learners (e.g. Russian and Thai according to the plan), we can easily add more blocks so that, instead of 4 L1 groups, we will have 5 or 6, or more as we continue. With reference to the systematic implementation of pragmatic factors, which tend to have the same features at the same proficiency level, this structure was very attractive because it can also make things easier for pragmatic annotation and tagging down the road. This allows the necessary flexibility for an ongoing research project of a similar nature.

Structure B, on the other hand, used the learner group as the fixed layer, commanding all the oral production data at different proficiency levels from the same L1 group.



The advantage of such an arrangement is ready access to any study focusing on one L1 learner group and to any comparison study within the same L1 group, regardless of the proficiency level in their oral production. From the perspective of language acquisition studies, such a structure offers conveniences in following the development of acquisition within a particular L1 group. Once the collection of data started, however, it was discovered that getting a sizeable group of learners at the same proficiency level within any L1 group was more difficult than expected, especially those at the higher end of the proficiency level. That means space allocated to that particular L1 group at a particular level would remain unfilled for an unknown period of time. Such uncertainty is hard to tolerate in most research projects and, moreover, it may also lead to inconveniences in any attempt to do research within the same proficiency level because of the lack of data.

Given the fact that the systematic implementation of pragmatic factors depends on a sizeable population of advanced learners, and with consideration of the difficulties involved in obtaining enough oral production of CSL learners at the high end of the proficiency, Structure A was finally adopted and all data collected were grouped accordingly. With the most recent update, the LAC/SC now has the following data:

CSL learners with English L1 (90 + samples):

transcribed and checked, in 3 proficiency levels;

CSL learners with Japanese L1 (45 + samples):
transcribed, in 3 proficiency levels;

CSL learners with Korean L1 (45 + samples):
transcribed, in 3 proficiency levels;

CSL learners with Cantonese L1 (600 + samples):
with 90 + of them transcribed and checked in 3 proficiency levels.

By design, each of the speaking samples has approximately 11 min of speaking time, covering 12 different content areas in various settings under three categories: informal, formal, and ceremonial or ritual.

As discussed earlier, three pragmatic factors were built in at the very beginning of the data-collecting process, when the tasks for participants were created. A pragmatic frame that includes information about the Interlocutors, the Setting, and the Timing (or Purpose) of the speech sample obtained and used in the corpus is available at two levels: the task level for all speakers, as well as the individual speaker level for all his/her tasks. The advantage of the availability of such information is obvious: it is now possible to conduct studies related to the appropriateness of oral production, either at the task level across L1 groups to find the similarities and differences, or at the speaker level to find the unique features associated with a certain L1 group.

Due to resource limitations, tagging and annotation for LAC/SC are still waiting to be completed. Studies have been done, nevertheless, based on clean copies of the transcription with sound files, including studies of prosodic features based on sound files, and sociolinguistic research focusing on advanced CSL learners' oral production based on the transcription and the sound files. Compared with other common corpus-based research focusing on phonological, semantic, and syntactic studies, one outstanding feature of the LAC/SC is the possibility to do research on the pragmatic ability of CLS learners. It is expected that, once the tagging of pragmatic features is completed and made searchable (e.g. presence/absence of greeting at beginning of a speech in a formal setting would be very useful for studies at a discourse level), more research can be done focusing on the pragmatic ability of the CSL learner at different proficiency levels; the salient features related to the appropriateness of a particular L1 group or proficiency level; and their understanding and use of words, phrases, and grammatical devices to show modesty as the native speakers tend to do, among others.

5 Concluding Remarks

Like all ongoing research, it is impossible, nor is it responsible, to draw any conclusion at this stage because LU as a system is still new to many, and too many questions remain unanswered. Based on the experience in the problem-finding and solution-seeking process while building the LAC/SC so far, however, it is reasonable to point out the following in reference to the pragmatic issues discussed in this paper.

1. In the construction of a learner corpus, data related to language structure (LS) is a given but information related to language use (LU) should be included. Moreover, it should be an integrated part of the corpus, starting from the very beginning as part of the overall design, not just an add-on later on. Framing the task of eliciting oral responses with real communication settings is a positive step in the right direction.
2. While LS is a system of systems consisting of phonology, semantics, and syntax (LS = Pho + Sem + Syn), LU can also be treated as a system of systems consisting of interlocutors (I), setting (S), and timing, or purpose of the communicative event (T/P) if clues for timing is not available or not important (LU = I + S + T/P). Employment of both systems, and not just one of them, could serve as a solid foundation in the construction of the CSL learner corpus, or any corpus for that matter.
3. Each of the subsystems in LU, like those in LS, can be further divided into categories and sub-categories, (e.g. the three categories under Setting: informal, formal, and ceremonial or ritual). Each of the sub-categories certainly has layers that allow further division for research purposes, such as semi-formal between informal and formal.
4. Compared with studies in LS, research is badly needed for LU as a system with reference to development in sociolinguistics. We must admit that only very little is understood about LU at this stage and there are many more factors in this system than what we have discussed here. Findings from more research in this direction, however, are expected to eventually contribute to the goal of culturally appropriate productions from CSL learners.
5. For tagging and annotation of the data in the LAC/SC down the road, both information for LS and LU should be included, starting with those related to I, S, and T/P at this stage. It would be impossible to study the appropriateness of language use if no information about LU is available.

Looking back and looking around, we must reiterate that studies on LS have been long and many, while those on LU are still sporadic by comparison. That means it is natural for us to see many more questions and challenges for any research focusing on LU. However tentative the concluding remarks above may seem to be, and no matter how big a hole we can see in various aspects of the LAC/SC as reported here, or similar projects reported somewhere else, what we have discussed in this paper helps us see that the study of pragmatic factors in corpus construction can contribute to the development of our field, even if some of us feel that our discussion has led to many questions and offered few answers.

References

Publications

Brezina, V., & Flowerdew, L. (Eds.) (2019). *Learner corpus research: new perspectives and applications*. Bloomsbury Academic.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Chen, H., & Tao, H. (2019). Academic Chinese: from corpora to language teaching. In X. Lu, & B. Chen (Eds.), *Computational and corpus linguistic approaches to Chinese language teaching and learning* (pp. 57–79). Berlin & Singapore: Springer.

Clark, J. L. D., & Li, Y. C. (1986). *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages*. Center for Applied Linguistics.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.

Fan, L. (2018). Pragmatic competence of advanced CSL learners in spoken Chinese: A comparison of native speakers of English and of Cantonese. Ph.D. Dissertation in Linguistics, The Hong Kong Polytechnic University, Hong Kong.

Feng, S. (2018). A sketch of Chinese *Yuti* Grammar. Beijing Language and Culture University Press.

Ke, C. (1994). An empirical investigation of the relationship between a simulated oral proficiency interview and the ACTFL oral proficiency interview. *Selecta*, *15*, 6–10

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.

Li, X. (1997). The science and art of language assessment. Hunan Education Press.

Li, Y. (2021). Responsibilities and concerns of linguistics. *Journal of North China University (Social Science Edition)*.

Liu, J. (2008). Research on pragmatic ability: Current status, problem and revelations. *Foreign Language Research*, *4*.

Ran, Y. (2004). New achievement in the interdisciplinary research on pragmatics and second language acquisition: Review on pragmatic development in a second language. *Foreign Language Teaching and Research*, *2*.

Rose, K., & Kasper, G. (Eds.). (2001). *Pragmatics in language teaching*. Cambridge University Press.

Tao, H. (2017). Spoken Chinese corpora: Construction and sample applications in research and language pedagogy. *Bulletin of the Chinese Linguistic Society of Japan., 2017*(264), 25–43.

Tao, H., Jin, H., & Zhang, J. (2020). A corpus-based investigation of manner/state complement constructions in Mandarin Chinese. *Sinica Venetiana, 6*, 1–40. https://doi.org/10.30687/978-88-6969-406-6/001

Wang, H. (2006). Pragmatics in foreign language teaching and learning: Reflections on the teaching of Chinese in China. In W. Chan et al. (eds.). *Foreign Language teaching in Asia and beyond: Current perspectives and future directions*. Center for Language Studies, National University of Singapore.

Wang, C. (2010). How we learn a foreign language. Foreign Language Teaching and Research Press.

Wu, W. (1993). *Towards a theory of teaching chinese as a second language*. Springfield, VA: ERIC Document Reproduction Service. ED 366 216.

Wu, W. (2006). Pragmatic points in teaching Chinese: A practical approach. *Chinese Teaching in the World, 1*, 91–96.

Wu, W. (2008a). Pragmatic framework and its role in language learning: With special reference to Chinese. In W. Chan et al (eds.), *Processes and process-orientation in foreign language teaching and learning*. Germany: De Gruyter Mouton. (Reprinted 2011).

Wu, W. (2008b). Teaching Chinese as a foreign language: theory and practice in proficiency test with a pragmatic approach. *Journal of Chinese Language Teaching, Beijing University Press*, *4*.

Wu, W. (2016). Chinese language pedagogy. In S. Chan, J. Minett, & F. Li (Eds.), *The Routledge Encyclopedia of the Chinese Language* (pp. 137–151). Routledge.
Wu, W. (2019). Language structure vs language use in TMP: Focusing on pragmatic ability of learners. *TCSOL Quarterly*, (1).
Wu, W. (2020). Implementation of the pragmatic framework in teaching: a systematic approach for CSL. *TCSOL Quarterly*, (2).
Xiong, D., et al. (2002) Research on large-scale recorded oral assessment for College English. *Foreign Language Teaching and Research*, *4*.
Zhang, J., & Tao, H. (2018). Corpus-based research in Chinese as a second language. In C. Ke (Ed.), *The Routledge Handbook of Chinese Second Language Acquisition* (pp. 48–62). Routledge.
Zhang, B., et al. (2019). *Research on tagging of Chinese interlanguage corpus*. Beijing University Press.

(on-line)

ACTFL https://www.actfl.org/resources/actfl-proficiency-guidelines-2012.
CEFR https://www.commoneuropeanframework.org.
English Corpora https://www.english-corpora.org/.
Learner Corpus Association https://www.learnercorpusassociation.org/.
Learner corpora around a world https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.
Media Language Corpus http://ling.cuc.edu.cn/RawPub/.
On-line Corpora www.cncorpus.org.
Sinica Corpus http://www.sinica.edu.tw/SinicaCorpus/.