Journal of Literature and Art Studies, March 2021, Vol. 11, No. 3, 183-189
doi: 10.17265/2159-5836/2021.03.009

The Referential Significance of French Interactive Spoken Corpus to the Construction of Spoken Corpus in China*

WANG Peng-fei
China Foreign Affairs University, Beijing 100037, China

In the aspect of interactive spoken corpus construction, this paper will summarize the development, research background, characteristics and experience of CLAPI interactive spoken corpus which is jointly developed by Université Lumière Lyon II and École Normale Supérieure de Lyon in the aspect of interactive spoken corpus construction. Combined with the characteristics of interactive spoken corpus construction, it proposes suggestions for the construction of spoken corpus in China by focusing on the former period of corpus collection, the mid period of corpus annotation as well as the later period of corpus research and promotion. It is predictable that interactive communication spoken corpus should be incorporated into spoken corpus, which not only enriches and improves the corpus itself, but also makes positive exploration on achieving the fact that corpus serve for language teaching and research.

Keywords: French CLAPI corpora, interactive communication spoken corpus, characteristics, referential significance

In recent years, great progress has been made in China's corpus construction, and various corpus studies have also flourished, among which the construction of spoken corpus has also received great attention. However, in the current process of building spoken corpus in China, corpus collection mainly comes from the oral language test, but rarely involves real interactive communication context. In terms of the construction of interactive spoken corpus, CLAPI, the interactive spoken corpus, jointly developed by two French University, Université Lumière Lyon II and École Normale Supérieure de Lyon, can provide us with useful reference.

I. Corpus Development and Research Background in France

France is one of the first countries to study and apply the concept of corpus, French corpus has been developed for a long time as well. The first French dictionary, *Le Dictionnaire François*, appeared in 1680, quotes a large number of literary citations, thus it can be regarded as the first French application of "corpus" (Petrequin, 2006, pp. 45-64). The study of contemporary French corpus and corpus linguistics has gained outstanding achievements with a solid foundation of research. A series of corpus with great significance and

* **Acknowledgements**: This paper is a supported by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province "Construction, Alignment and Application of Chinese-French Parallel Corpus" (Project Number: UNPYSCT-2020147).

WANG Peng-fei, Ph.D.. lecturer in Department of Foreign Languages and International Studies, China Foreign Affairs University.

important research value have been established successively. For example, Frantext, the corpus established under the auspices of National French language Corpus Institute (Institut National de la Langue Française), has collected more than 5503 texts and more than 260 million form signs by March 2021[1]. In the field of corpus research, the three major laboratories of French corpus research supported by CNRS[2], namely The Lattice Laboratory, LiLPA laboratory, and ICAR laboratory, have played important roles.

Lattice laboratory, located in Paris, France, with its full name as "Language, Texts, Information treatment and Cognition Laboratory (Langues, Textes, Traitements Informatiques, Cognition)", is a place where experts and scholars from University Sorbonne Nouvelle-Paris III and École Normale Supérieure can conduct collaborative research. It is mainly focusing on the development and research of language visual analysis tools. The representative results of it are text annotation database (La Base EIOMSIT) and loanwords annotation database in news discourse (BSP). LiLPA laboratory located in Strasbourg, with its full name as "Linguistics, Language and Parole laboratory (Linguistique Langues, Parole)", mainly rely on a scientist team from the University of Strasbourg to carry out research. Its researches focus on the description, implementation approach and the length of text reference chains, and distribution study of annotation chain in different genres of texts. The representative results are the research and development of automatic annotation tool ANALEC and annotation database of news text.

The researches of the above two laboratories mainly focus on the development, annotation, and research of written language database (namely Corpus), while the ICAR[3] Laboratory emphasizes the development and research of spoken language Corpus more. Located in Lyon, ICAR is a joint laboratory of Université Lumière Lyon II and École Normale Supérieure de Lyon, aiming to conduct multi-dimensional analysis and research on the language using in oral interactive communication, and its representative achievement is a multimodal interactive spoken language corpus: CLAPI (Corpus de Langue Parleeen Interaction). The following part will introduce the features and advantages of this French spoken corpus.

II. Features and Advantages of CLAPI

As an interactive spoken corpus, CLAPI Corpus has its own distinctive features and uniqueness in corpus scale, corpus collection, corpus retrieval, corpus annotation, etc. We shall analyze these four advantages with specific features in the following part.

2.1 Rich Corpus and Complete Information

As is known to all, the collection of spoken language corpus is difficult, and many famous spoken language corpora in the world are built on a scale of about 100 hours. CLAPI Corpus contains more than 170 hours, including 61 audio/video materials and 140 corresponding text transliteration materials. The large scale of spoken language corpus ensures its rich corpus reserve. Meanwhile, in terms of corpus information, CLAPI also tries to keep the integrity of corpus information. For example, in terms of the storage format of corpus, CLAPI provides both original audio/video data and transcribed text data. In terms of corpus types, CLAPI provides both raw and

[1] https://www.frantext.fr/
[2] Centre national de la recherche scientifique.
[3] Interactions, Corpus, Apprenticeship and Representation.

annotated corpus for researchers to use. To sum up, the sheer scale and complete corpus information in CLAPI Corpus become a major advantage of this spoken corpus, which convenients the users a lot.

2.2 Extensive Sources and Various Contents

The basic content of CLAPI is interactive spoken corpus, which is collected from various types of interactive communication scenes, thus it has a wide range of corpus sources and various contents. According to the types of interactive communication scenes, the corpus sources of CLAPI corpus can be divided into three categories: first, formal communication activities, such as classroom communication, chamber of commerce discussion, etc.; second, informal communication activities with themes, such as film discussion, dialogue in Subway inquiry office, etc.; third, informal communication activities without themes, such as chat in bar, parent-child interaction dialogue and so on. As Hu Zhuanglin (1994, p. IV) pointed out: “the more corpora are involved, the more explicit it can describe a problem”. CLAPI’s multi-subject interactive spoken corpora enrich both its own content and the applicable research field, such as interactive linguistics, social linguistics, foreign language teaching, which greatly improves its own value and research value.

2.3 Complete and Clear Corpus Annotation

CLAPI’s spoken corpus is mostly interactive spoken corpus, which occurs in the real context of daily communication. As a result, all kinds of natural language phenomena and non-language phenomena can be commonly seen. To preserve these “special” language phenomena as completely as possible, it is necessary to annotate these phenomena through related annotation means.

Based on that truth, the research team form ICAR Laboratory have designed a set of effective spoken corpus transcription and annotation convention, which can be used to mark the natural language phenomenon, non-language phenomenon and conversational turns that may appear in the interactive spoken corpus. The ICAR (2013) research team collated these conventions into normative annotation rules, as shown in Figure 1:

	Phénomène	Conventions en format liste	Exemples en format liste
1- Identité du participant (tire de transcription par locuteur)			
1a	Participant identifié	Identifiant en début de paragraphe du tour (voir 2a). L'identifiant est composé d'un, de deux ou de trois caractères (en majuscules). Il est suivi d'une tabulation. *Attention : c'est le seul endroit dans une transcription ICOR où l'usage de la tabulation est admis.*	COR salut
1b	Participant incertain	L'identifiant supposé est suivi d'un point d'interrogation *Attention* : dans le travail sous Clan ou d'autres logiciels d'alignement, l'utilisation du « ? » n'est pas acceptée.	COR? salut
1c	Hésitation sur les participants	Porter en les séparant par un « ? » les identifiants des participants entre lesquels on hésite. *Attention* : dans le travail sous Clan ou d'autres logiciels d'alignement, l'utilisation du « ? » n'est pas acceptée.	DAR?COR salut
1d	Participant non identifié	L'identifiant est composé de trois « X ».	XXX salut
		ou de deux « X » suivis d'un « F » ou d'un « H » selon l'identification du sexe du participant.	XXF salut
			XXH salut
1e	Evénement non attribuable à un participant	Description entre doubles parenthèses sans identifiant.	((un coup de tonnerre retentit))

Figure 1. Spoken Corpus Transcription and Annotation Conventions in CLAPI.

As shown in Figure 1, the annotation convention is presented in order in four columns, which are namely code number, special language phenomenon that needs to be annotated, an exact convention used for this certain language phenomenon, and annotation examples. At the same time, according to the different annotation parameters (such as communication participants’ identity, conversational turns, etc.), the annotation conventions can be classified into different sectors, the highlighted content in the fourth column from the left shows the

specific examples. The annotation conventions show the special language phenomenons need to be annotated, corresponding annotation conventions and exact examples in a visual form, which present the whole picture of the annotation conventions clearly and legibly. Meanwhile, these three bright spots are also convenient for users to search the concrete content of each convention according to different annotation parameters.

2.4 Multiple Available Retrieval Modes

In terms of corpus retrieval, CLAPI Corpus provides multiple retrieval modes for researchers based on different purposes, involving the following three types: (1) corpus sample retrieval, such as retrieving original audio/video data, retrieving raw corpus samples, retrieving annotated corpus samples, etc.; (2) Lexical retrieval, such as form sign/type sign retrieval, word frequency retrieval, root/compound word retrieval, etc.; (3) multi-standard retrieval, such as specified spacing retrieval, repeat segment retrieval, etc.

CLAPI development team insists on innovation while constantly proposing new retrieval modes to meet researchers' retrieval goals and research objectives which are more diverse and specific.

To sum up, CLAPI corpus shows its features mainly in corpus scale, collection, retrieval, and annotation, etc. These features have great referential significance for the construction of Chinese spoken language corpus. In the next part, the author will elaborate on the referential significance for the construction of Chinese as Interlanguage Spoken Corpus from three main aspects, which are the corpus collection, corpus annotation and corpus research.

III. CLAPI's Corpus Collection Experience in Earlier Stages

The collection ofthe interactive spoken corpus is more complicated and involves more factors than that derived from oral language test. To sum up, the aspects that have be paid attention to are types, scenes and ethics of corpus collection.

3.1 Type of Corpus Collection: Interactive Communication Spoken Corpus

The spoken corpus collected by commonly spoken corpora is mostly from the recording of oral language test, which is characterized by that a testee who completes an independent spoken paragraph coherently without interference. For example, "Spoken English Corpus of Chinese Learners (SECCL)" and "Chinese Learner Spoken Corpus (CLSC)" are two important domestic spoken language corpora, where the corpora are all from various English and Chinese test recordings (Yang Yi et al., 2006; Wang & Wen, 2007). However, underneath the special and unique collection scene of oral language test, spoken language corpus was only produced under this specific flat scene with the interference of "*Pseudo-real*" communication factor without interaction. Based on this fact, to obtain the real language use of Chinese learners more comprehensively, the author suggests that an interactive communication corpus can be incorporated into the collection scope, emulating the CLAPI corpus. The comparison of the characteristics between oral language test corpus and interactive communication corpus are as follows in Table 1:

Table 1

Comparison between Oral Test Corpus and Interactive Communication Corpus

Corpus Parameter	**Oral Test Corpus**	**Interactive Communication Corpus**
Scene	Unique scene	Multiple scenes
Corpus type	Independent language paragraph	Dialogue
Communicative Object	No	Yes
Learner's status	Nervous	Relaxed

The comparison shows that the interactive communicative corpus involves more diverse scenes than the oral language test corpus, and most of the corpus forms are dialogues. Owing to there are communicative objects, the learners' status during corpus collection can be more relaxed. Inevitably, more diverse and complex corpus parameters will bring a difficulty to the collection, but at the same time, they can greatly increase the availability and research value of the corpus. This requires researchers to overcome the technical difficulties and try to collect the interactive communication corpus actively.

3.2 Corpus Collection Scene

As mentioned above, the CLAPI corpus can be divided into three categories: formal communication activities, informal communication activities with themes and informal communication activities without themes. Specific to the construction of China's domestic spoken corpus, we can combine the purpose of corpus construction with the particularity of corpus collection objects, and propose different types of collection scenes. For instance, the corpus construction target is "French learners' corpus", and the collection object is Chinese learner major in French in China's domestic universities. Therefore, the following three types of corpus collection scenes can be directly proposed:

- **Class interactive activities**. Specifically the interaction with teachers and classmates;
- **Daily communication activities with themes**. For example, Chinese learners participate in French language salon with prescribed themes;
- **Daily communication activities without themes**. For example, Chinese learners chat with friends in a cafe.

By classifying corpus collection scenes, sufficient corpora can be collected in each type of communication scene, which not only ensures the objectivity and representativeness of corpus construction but also helps researchers to carry out relevant studies in the future.

3.3 Corpus Collection Ethics

During spoken corpus collection (especially interactive spoken corpus), the corpus collection objects involved in this process are normally natural persons, therefore some rights and freedoms of collection objects also need to be concerned, such as the right of information, privacy and portrait, etc. This is not only an ethical issue but also a legal issue. If it was treated casually, it would not only affect the construction of corpus and the development of related research but also provoke legal disputes. Therefore, various measures should be taken to ensure that corpus collection ethics are respected and protected. To be more specific, we propose to ensure the ethics of corpus collection from the following three aspects:

First, **to ensure the objects' right to information.** Before corpus collection begins, the objects shall be informed of the subject and purpose of the study as well as the usage of the corpus. After obtaining the objects' consent, the researchers shall let the objects sign a declaration document in agreement with video recording or audio recording. Only after that, the corpus can be collected.

Second, **to protect the privacy of the objects.** In the original collection of audio/video data, the identity information of the collected objects shall not appear or be disclosed. Meanwhile, when speakers and communication objects are involved in the transliteration text, code names shall be used for reference.

Third, **toget the right to use corpus**. The collector shall guarantee to the objects that the collected corpus is only used for academic research, and write the relevant terms into the signed statement. The researchers shall

request the objects to sign the statement, and only after obtaining the permission can the collected corpus be used for corpus development and follow-up research.

IV. Experience in the Research and Application of CLAPI Corpus

To apply the experience of CLAPI corpus to the construction and research of Chinese as Interlanguage Spoken Corpus successfully, it is necessary to fully draw on the experience of CLAPI corpus in research after corpus collection and corpus annotation, especially from the two aspects: **composition of researchers** and **promote experience**.

4.1 Composition of Team Members

According to the experience of ICAR laboratory, corpus construction requires three teams to cooperate:

Linguistics expert team, the leader and designer of the corpus construction. It is in charge of the creation, planning and coordination of corpus construction, the formulation of corpus annotation standards, corpus selection, the following research, the application and the promotion of the corpus;

Computer expert team, the creator of interdisciplinary integration application between computer technology and the corpus establishment construction. In particular,computer expert team takes charge of the storage of original corpus data, the realization of visual annotation, the construction of an online retrieval system and the publication and maintenance of related achievements of the corpus.

Implementation team, the specific performer of corpus construction. It is in charge of corpus collection, input, transliteration, basic annotation, online maintenance and other work. This team can be made up of mainly postgraduate students in linguistics and computer science majors.

4.2 Promote Experience

Propagating widely and expanding cooperation are two effective ways to realize corpus construction and promotion. At present, the French teaching circles in China and the French research institutions in France have cooperated more and more frequently. They have established an extensive exchange mechanism as well. As an example, the French Department of Beijing Language and Culture University has cooperated with Lattice and LiLPA, and signed an exchange agreement with the ICAR Laboratory which belongs to École Normale Supérieure de Lyon. They have established a good cooperation platform and layed a solid foundation of communication. The author sincerely welcomes the Chinese language research team to join the cooperation framework. We believe that the construction and development of Chinese/French corpora canstrive for further improvement while realizing linguistically interlingual communication and collaboration.

To enlarge the influence and application value of a corpus, the promotion work in post-research period is necessary. According to the experience of CLAPI corpus, promoting **research results** and **resource sharing** are two significant parts:

First of all, attention should be paid to the output based on corpus research results. The research team should promote the research results from two aspects: **research project promotion** and **research paper promotion**. Take CLAPI corpus as an example, based on this corpus, there are currently three projects under research (one French national scientific project, covering three sub-projects conducted by three research groups to tackle key and difficult problems; and two optional items); and 4 completed optional projects. In terms of paper promotion,

CLAPI corpus has become an important corpus source for the doctoral dissertation of École Normale Supérieure de Lyon and Université Lumière Lyon II. Up to now, 48 doctors or doctoral students have completed the opening of their doctoral dissertation based on CLAPI corpus, which not only tested the practicability of CLAPI corpus, but also increases its awareness.

Secondly, corpus resources should be widely shared. Laying it aside would reduce its value greatly. Therefore, it is imperative to share the resource of corpus from three aspects: **data sharing, standard sharing and retrieval sharing**. Data sharing is to provide researchers with open access to original audio/video data and transcribed text data so that researchers can carry out their own research independently. Standard sharing refers to provide the annotation conventions of spoken corpus to researchers. It can yield twice the result with half the effort by facilitating their attention to existing research, and helping them refine the annotation conventions based on the characteristics of proposing research. Retrieval sharing is to share the corpus retrieval system which has been mentioned above. On the other hand, sharing the corpus resources sufficiently will certainly expand the scale of corpus construction, improve the deficiencies in it and complete the overall corpus construction and research.

V. Conclusion

Mentioned above, the author has introduced the development and research background of CLAPI interactive spoken corpus, as well as its characteristics and use. At the same time, combined with CLAPI's characteristics, the author has put forward suggestions on the construction of Chinese as Interlanguage Spoken Corpus. We hope that more fellow researchers in Chinese and foreign language fields can participate in the creation of corpus, especially spoken corpus. Predictably, a large number of spoken corpora of various languages and different types will surge; a large number of spoken corpora will accumulate; more standardized reasonable data processing approaches and retrieval systems will be established. The series of progress will not only improve the corpus building but also make the corpus to serve empirical language teaching and research. Thus, it will eventually achieve the goal of interpreting language. Like French, Chinese is a language with rich corpora. The construction and usage of corpus is also a bridge for us to inherit the tradition and face the future. Only when we build such a solid bridge like that, can the "era of transformation and regeneration" mentioned by John Sinclair (1997) come.

References

Groupe ICOR. (2013). *Convention ICOR*. UMR 5191ICAR. CNRS-Lyon 2-ENS de Lyon.

Hu, Z. (1994). *The cohesion and coherence in discourse*. Shanghai: Shanghai Foreign Language Education Press.

Petrequin, G. (2006). La "langue littéraire" dans le Dictionnaire françois de Richelet (1680). In F. Berlan (Ed.), *Langue littéraire et changements linguistiques*. Paris: Presses Paris Sorbonne.

Sinclair, J. (1997). Corpus Linguistics at the Millennium. In H. Z. Yang (Ed.), *An introduction to corpus linguistics*. Shanghai: Shanghai Foreign Language Education Press.

Wang, L., & Wen, Q. (2007). A view on the construction and research of spoken and written English corpus of Chinese learners. *Foreign Language World*, (1).

Yang, Y. et al. (2006). Tentative ideas of constructing Chinese learners' spoken corpus. *Chinese Language Learning*, (3).