A Learner Corpus - ESCCL*

Chen Hua[1] *Wen Qiufang*[2] *Li Aijun*[3]

[1]School of Foreign Studies, Nantong University, PRC
[2]National Research Center of Foreign Language Education, BFSU, PRC
[3]Institute of Linguistics, Chinese Academy of Social Sciences, PRC
jenniferb@vip.sina.com

Abstract

This paper describes the reason, the method and the implication of compiling a learner corpus - ESCCL (English Speech Corpus of Chinese Learners). As for the main reason, the existing spoken corpora of Chinese EFL learners in China are completely text-based, and not suitable for phonetic analysis because of the poor quality of recording. The subjects at 4 different educational backgrounds were asked to fulfill 2 tasks ¨ reading-aloud and topic-based spontaneous dialogue. The recordings were collected from different parts of China and dialectal areas. The annotation system employed in the corpus combines the British system and the American one. The corpus-based research findings have important implications for China¡s EFL pedagogy, and will be helpful for the improvement of rating rubrics for China¡s oral English tests.

1. Need

The development of the computer science calls for corpus-based linguistic research. Further, in order to provide a comprehensive picture of speakers¡ patterns of English, studies on spoken English should cover not only the grammatical, lexical and syntactic aspects, but the phonological features as well. These motivate the occurrence of spoken corpus.

Taken native speakers as subjects, BNC (the British National Corpus: 10 million tokens in spoken section), Bank of English (the largest one so far: 15 million tokens), and London-Lund Corpus of Spoken English (the first computerized corpus with prosodic annotation: 0.435 million tokens) are worth mentioning.

In China, two corpora are related to students of English at tertiary level. COLSEC (College Learners¡ Spoken English Corpus) was compiled based on the standard English oral tests ¨ CET (College English Test) Band 4 and Band 6, in which the subjects are non-English majors. SECCL was similarly designed but for English majors in China. Based on these corpora, large quantities of research papers made the understanding of Chinese EFL learners¡ oral English patterns available to some extent, which are on lexical, syntactic, textual, pragmatics and other aspects.

However, at least several things need improving. For one thing, these studies were almost all error analysis, lack of description of acceptable features. Further, the studies have been conducted on the basis of transcribed text-based corpora and the corpora were tagged with POS tags, and for errors, which lack prosodic annotation.

What¡s more, since phonetic analysis has to depend on acoustic cues, the requirement for sound quality is critical. But existing spoken corpora in China were not aimed at phonetic analysis, in which subjects were recorded without the control of noise, so they are not suitable for phonetic analysis (see Figure 1).

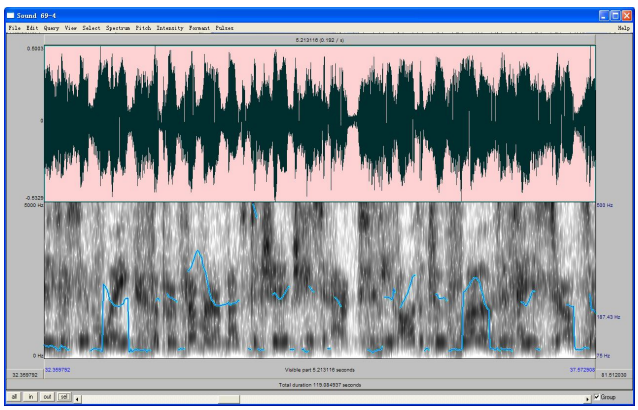

Figure 1: *Quality of sound in existing spoken corpus*

Studies on Chinese learners of English have revealed that Chinese EFL learners do have some problems which affect the conveying of information on prosodic level (Chen, 2006a; 2006b). While some of the problems may only cause their speech to sound foreign, or inaccurate, some others may put at stake the intelligibility of their speech, as certain aspects of English intonation are indispensable to the successful conveying of information.

Since these studies are not corpus-based ones, a speech corpus of Chinese EFL learners is necessary in order to confirm these findings.

* The authors gratefully acknowledge the support of both a project sponsored by Ministry of Education, PRC (06JA740031), and the MOE Project of Key Research Institute of Humanities and Social Sciences in Universities.

2. Variables

In compiling this speech corpus, the data are categorized according to two kinds of variables: learner variables and task variable.

2.1 Learner Variables

Learner variables can be classified according to the following perspectives: geographical, dialectal, and educational backgrounds (Wang & Li, 2001).

Subjects in the corpus come from different parts of China (see Figure 2), and have covered 10 major dialect areas. The subjects were also randomly sampled from different educational backgrounds.

The purpose of collecting data in different dialectal districts is to provide a platform for the potential research topics on dialectal transfer. According to Chinese Academy of Social Science¡s Dialect Map, dialectal districts are Northern Dialect (mainly in North-East, North China, and North-West), Wu Dialect (mainly in East China), Xiang Dialect (mainly in Mid-China), Gan Dialect (mainly in Mid-China), Hakka Dialect (mainly in South China), Yue Dialect (mainly in South China), Min Dialect (mainly in South China), JinDialect (mainly in North China), Ping Dialect (mainly in South China) and Hui Dialects (mainly in East China).

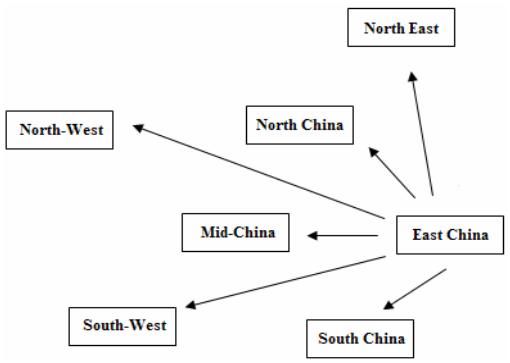

Figure 2: *Geographical distribution of subjects*

The subjects in the corpus cover almost all learners under formal classroom instruction at different educational backgrounds, with an interval of three years between adjacent groups. In each part of China (also in each dialectal district), at least 30 junior middle school students, 30 senior middle school students, 30 college English majors, and 30 English majors for master degree were willingly recorded. In each group, the number of male and female students was well balanced.

2.2 Task Variable

All subjects were asked to fulfill only one task (dialogue reading-aloud) except college level students when the corpus was being compiled.

The subjects at middle school levels were asked to read aloud 16 dialogues, while the students at college levels were asked to read aloud 10. The content is quite familiar to the subjects without any new words. About average of 3 minutes¡ read speech was recorded for each middle school student and 2 for college student. The reading material is full of syntactic sentence types and prosodic features (see Table 1 for details).

Table 1 : *Syntactic and Prosodic properties of reading material*

Features	Examples
Simple Declarative	*Something wrong with my computer.*
subordinate clause	*I¡ve heard that that film ¡*
Relative clause	*¡ the shop where you bought it.*
Selective Interrogative	*Do you want to ¡ , or have you ¡*
Yes-No Interrogative	*Will they make the man sick?*
Wh-Interrogative	*What do you want buy?*
Tag Interrogative	*¡ , didn¡t they?*
Imperative	*Close the window.*
Listing	*spring, summer, autumn and winter*
Final reporting phrase	*¡¡¡ , Betty asked.*
Final vocative	*¡ , Mummy.*
Temporal / spatial adverbial	*¡ last night ¡ over there*
Contrastive Feature	*¡ clever ¡ ¡ foolish ¡*
Emphatic Feature	*Mary does like swimming.*
Negation	*¡ don¡t*
Compound nouns	*New York*
Nominal phrase	*a famous tragedy*

College English majors were further asked to choose one topic randomly out of 14, and then conduct a topic-based dialogue for each pair. Still, the task is easy enough for average student to accomplish. They have 10 minutes to prepare and the dialogue should last about 2 minutes.

The recording was done in language labs by MP3-H06 at the sample rate of 16000 (16kHz, 16 bit mono PCM) (Zu, 1998). The noise was under strict control (see Figure 3).

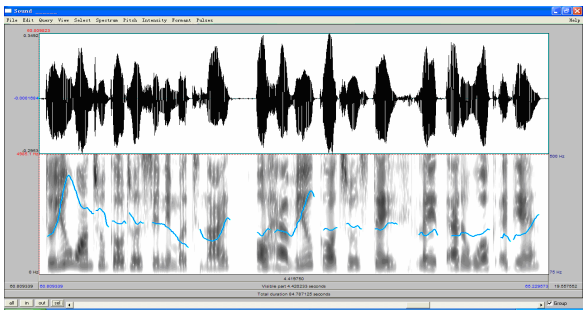

Figure 3: *The quality of recording in this corpus*

The recorded data were annotated (read speech at the present time) on the computer with Praat (http://www.praat.org) by 15 college English researchers. All the data were cross-checked by three phoneticians in China.

3. Annotation

The annotation system in this corpus combines the British one (Prehead ¨ Head ¨ Nucleus - Tail) and the American one (ToBI). Therefore, 6 tiers were done totally (see Figure 4).

The first two tiers are corresponding to the Orthographic Tier in ToBI system. The first tier describes the actual pronunciation in terms of word. The second tier describes both the standard pronunciation and the actual pronunciation of the speaker in terms of IPA. The standard pronunciation is referred to ***Longman English Pronunciation Dictionary*** (Wells, 2005). The IPA phonemes are realized by the codes on Praat.

The third tier matches the Break Index Tier in ToBI system, in which 4 stands for intonation group boundary, 3 for intermediate group boundary, 1 for prosodic word boundary, and 0 for clitic group.

The Fourth tier is for sentence accentuation indicated by H* (high pitch prominent) and L* (low pitch prominent). The nucleus (British version) is marked with H**.

The fifth and sixth tiers match the Intonation Tier in ToBI system. The fifth one is a British one and the sixth is the American one. The British system describes English intonation in terms of 2 types of preheads, 7 types of heads, and 7 types of nuclear tones (see Table 2 for details).

After being annotated, the data are to be categorized according to some factors, like task types, parts of China, dialectal areas, etc. In addition, building up recording diary and speakers¡ files are also very important.

Table 2: *Annotation coding samples in the corpus*

American	British	Coding
H* L L%	Fall	<NTHF>
H* L H%	Fall-Rise	<NTFR>
H* H H%	High Rise	<NTHR>
L* L L%	Low Fall	<NTLF>
L* H H%	Low Rise	<NTLR>
L+H* L L%	Rise-Fall	<NTRF>
L+H* H H%	High Rise (with low head)	<HL> <NTHR>
L*+H L L%	Rise-Fall (emphatic)	<sliding> <NTRF>
H+L* L L%	Low Fall (with high head)	<HH> <NTLF>
H+L* L H%	Fall-Rise (with high head)	<HH> <NTFR>
H*+L H L%	Fall-Rise (¡calling contour¡)	H% <HH> <NTFR>

4. Implications

As a special corpus, learner corpus has important implications on not only second language acquisition studies, but foreign language teaching as well.

Constructing Chinese EFL learners¡ English speech corpus with basic prosodic annotation information can be a complementary research database for existing learners¡ spoken corpora in China. The existing corpora have limitations on completely text-based tagging, lack of possibility to be checked with reference to the sound facts such as spectrum. Therefore, the implication of building a learner speech corpus lies in providing a platform for researchers and learners to learn the Chinese EFL learners¡ patterns; and make it possible to conduct contrastive analysis with native speech corpus.

Eternal recordings are able to confirm language facts that those who fail to have a field research experience to believe and test the facts.

With speech corpus, quantitative studies are available, for example, the comparative studies of intra- or inter- linguistic durational differences.

With the help of experimental approach, corpus-based study could reveal the weakness in English teaching in China, which has important implications for China¡s EFL pedagogy as well as for the improvement of rating rubrics for China¡s oral English tests.

5. References

[1] Armstrong, S., 1993. *Using large corpora.* London: A. Bradford Book, MIT Press

[2] Biber, D., Conrad, S., & R. Reppen. 1998. Corpus-based approaches to issues in applied linguistics. *Applied Linguistics*, 15(2): 179-189

[3] Chafe, W., 1992. The importance of corpus linguistics to understanding the nature of language. In: Svartvik J (ed.) *Directions in corpus linguistics.* Berlin / New York: Mouton de Gruyter

[4] Granger, S., 1998. *Learner English on computer.* London / New York: Longman

[5] Johansson, S., 1982. *Computer corpora in English language research.* Bergen: Norwegian Computer Center for the Humanities

[6] Chen Hua, 2006a. Intonation phrasing in Chinese EFL learners¡ read speech. Proceedings of the International 2006 Speech Prosody Conference, Dresden

[7] Chen Hua, 2006b. Tone patterns in Chinese EFL learners¡ read speech. Proceedings of Tonal Aspects of Languages 2006，La Rochelle

[8] He Anping, 2004. *Corpus linguistics and English Teaching*. Beijing: Beijing Foreign Language Teaching and Research Press

[9] Li Aijun, 2001. *Dialogue corpus CADCC and phonetic research*. Proceedings of the 5th National Phonetic Conference, Beijing

[10] Li Wenzhong, Corpus, learner corpus and foreign language teaching. *Foreign Language World*. 1999(1).

[11] Wang Jianxin, 2005. *Construction and Application of Computerized Corpus*. Beijing: Qinghua University Press

[12] Wang Yunjia & Li Jimei, About building Chinese interlanguage speech corpus. *The World Chinese Teaching.* 2001(1).

[13] Wei Naixing, Studies on COLSEC. *Modern Foreign Language*. 2004(2).

[14] Wells, J., 2005. *Longman English Pronunciation Dictionary*. London: Longman

[15] Wen Qiufang, Wang Lifei, Liang Maocheng, 2005. *Manual for using SWECCL*. Beijing: Beijing Foreign Language Teaching and Research

[16]Yang Huizhong, 2002. *Introduction to Corpus Linguistics*. Shanghai: Shanghai Foreign Language Education Press

[17] Yang Huizhong, & Gui Shichun, 2005. *CLEC-based Chinese EFL Learners English Analysis*. Shanghai: Shanghai Foreign Language Education Press

[18] Yuan Jiahua, 2001. *On Chinese Dialects.* Beijing: Chinese Literacy Press

[19] Zu Yiqing, Some Keys on the Scientific Speech Corpus. *Applied Language Literacy*. 1998(1).

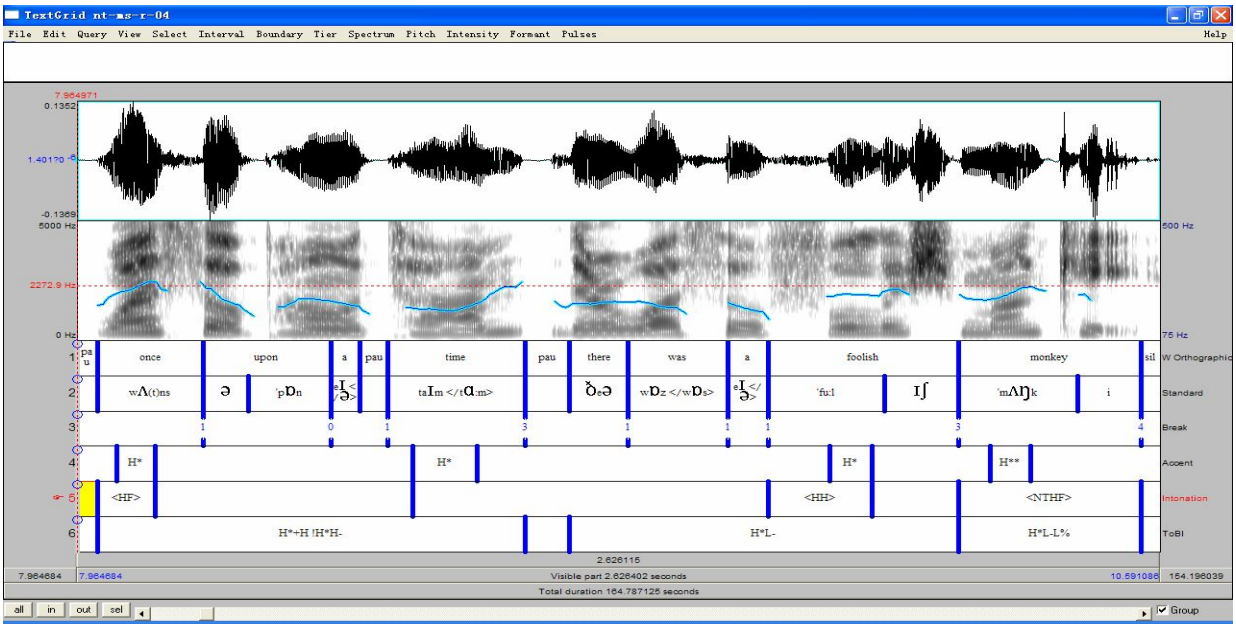

Figure 4: *Annotation sample*