Send Orders for Reprints to reprints@benthamscience.net

The Open Automation and Control Systems Journal, **2013,** *5,* **45-50**

Open Access

The Operation of Computer Technology in Corpus-based Spoken Language

Chen Jie[1] and Chen Jing[2,*]

[1]*College English Teaching and Research Department, Harbin Normal University, Harbin, P.R. 150025, China*

[2]*Computer Center, No. 1 Middle School of Qiqihar, Qiqihar, P.R. 161005, China*

Abstract: Corpus refers to the database of language materials. Cool Edit Pro is a media edit software. This paper explores how to construct spoken language corpus, how to use cool edit pro 2 to make waveforms for L2 learners at the different stages, make contrast among them and try to provide for the experimenters an intuitive observation from their own speech waveforms. The key is to offer the obvious waveforms contrast among the sampling waveform of the native speaker, the L2 learners' more times waveforms before and after necessary modifications and instructions, resulting in making the oral autonomous learning more effective and efficient. This paper follows the modern education technology methods and quantitative and qualitative analysis to make experiments on pronunciations of speech, aiming to get a satisfactory result and further promote this method to the other fields of language teaching and learning.

Keywords: Computer-assisted, Corpus, Spoken Language.

1. INTRODUCTION

Nowadays large-scale corpus with millions of the words becomes more popular and gains more concern. There are some internationally famous language materials database like London-Lund Corpus with 1/2 million words(LLC), Lancaster/ IBM Spoken English Corpus with 5000 words(SEC), Corpus of London Teenage Language with 1/2 million words(COLT), Wellington Spoken Corpus with 1 million words, Corpus of Spoken Professional American English with 2 million words(CPSA), Long-man Britain Spoken Corpus with 10 million words, Longman Spoken American Corpus with 5 million words, The Bank of English with 450 millions words, British National Corpus with 100 million words and The East African Component of the International Corpus of English(ICE-EA) [1].

In China there are also many famous corpora-based spoken/written English, as shown Table **1**.

With the development of the computer technology, utilizing the computer to record, play and process the digitalized speech materials has changed and overcome some weaknesses like audio materials' order, continuity, speed and volume. [3] Furthermore, we find computer can also settle some practical problems like large-scale oral materials collecting, recording, editing, processing, testing and even intuitively demonstrating its characteristic value acquired during the speech.

*Address correspondence to this author at the Computer Center, No. 1 Middle School of Qiqihar, Qiqihar, P.R. 161005, China; Tel: +86-451-88065058; Fax: +86-451-82278386; E-mail: tinahashida@163.com

2. CORPORA IN ENGLISH LANGUAGE TEACHING (ELT)

Over the past years, corpora, corpus tools and corpus evidence have not only been used as a basis for linguistic research but also in the teaching and learning of languages. Nowadays, more and more researchers and practitioners admit what corpus linguistics offers to language pedagogy, and the impressive number of recently published monographs and edited collections on the topic clearly indicate the growing popularity of pedagogical corpora use and the need for research in this area. The practice of ELT (English Language Teaching) to date, at least, seems to be largely unaffected by the advances of corpus research, and comparatively few teachers and learners know about the availability of useful resources.

ELT corpora mainly focus on three main dimensions: the first one deals with the compilation and exploitation of learner corpora; the second explores error analysis and make comparisons using learner corpora and native speaker corpora; and the third deals with the use of corpora to create teaching materials. The compilation (See Fig. **1**) and use of corpora has greatly enriched language learning practices. When we consider the issue of corpus compilation purposes, another interesting feature stands out: how the texts are obtained, i.e. the compilation methodology. Thus, we think that an important point is that learner corpora follow a task-based instead of a text-type based approach in their compilation and database organization. The current problem is the scarce availability of materials derived from spoken corpora.

Spoken learner corpora requires direct and indirect use and the teachers' participation that could only be carried out if the teacher is a corpus linguist or is trained specifically to

1874-4443/13

Table1. Famous Corpora-based Spoken/Written English in China [2]

Name Type	Construction Unit	Mother Language	Capacity
HKUST	Written English	HKUST	Cantonese>2500
TSLC	Written English	UHK	Cantonese 300
CLEC	Written English	GDUFS	Chinese 100
COLSEC	Spoken English	SJTU	Chinese 50
MSEE	Written/Spoken	SCNU	Chinese 87.6
SWECCL	Written/Spoken	NJU	Chinese>200（10,000 Words）

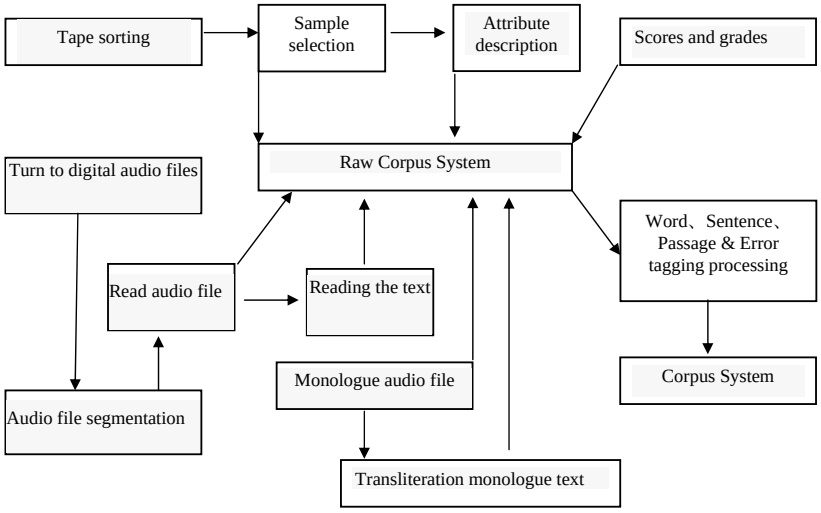


Fig. (1). The procedure of complicating corpus-based spoken language [4].

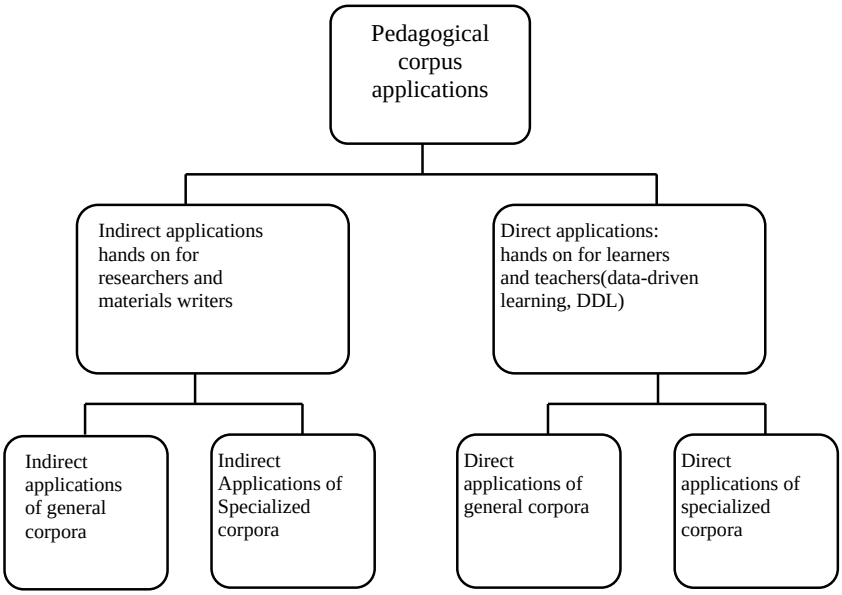
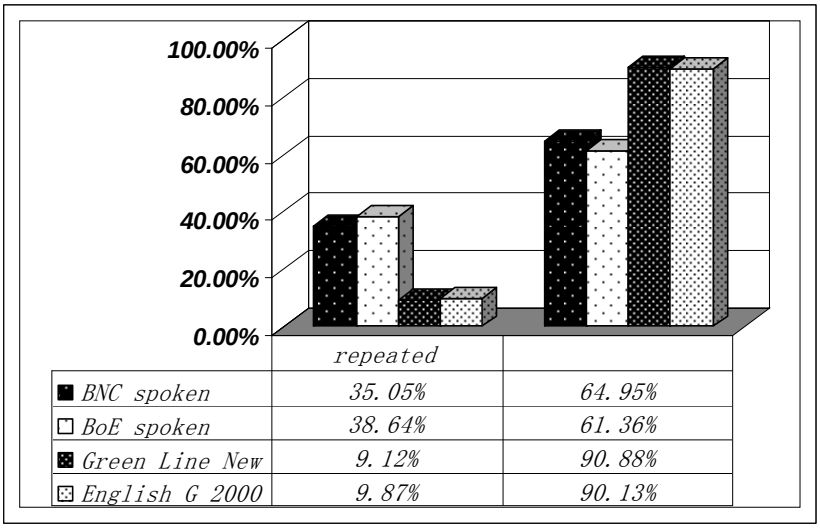


Fig. (2). The use of corpora in language learning and language teaching [5].



	repeated	
BNC spoken	35.05%	64.95%
BoE spoken	38.64%	61.36%
Green Line New	9.12%	90.88%
English G 2000	9.87%	90.13%

Fig. (3). [5].

deal with such corpora, because spoken corpora are difficult to handle in depth. See the use of corpora in language learning and language teaching in Fig. (**2**).

As Fig. (**2**) shows, direct and indirect pedagogical corpus uses both general and specialized corpora. Indirect applications involve hands-on work mainly for corpus researchers and, to a limited extent, materials writers and provide answers to questions on *what* to teach and *when* to teach it, but direct applications mainly affect *how* something is taught and actively involve the learner and teacher in the process of working with corpora and concordances [5].

Look at Fig. (**3**)[5], a case in point here are the misrepresentations of the preferred functions and contexts of the English progressive in German EFL teaching materials.

Recent pedagogically oriented studies of the indirect type that take corpus findings seriously, and use language features that are known to cause problems to language learners as their starting point, include those on reflexives, on linking adverbials, on irregular verbs, on future time expressions, on modal verbs, if-clauses, and progressives, and on the present perfect. All studies found mismatches between naturally occurring English and the English presented in English as a foreign language teaching materials. [5]

3. THE ANALYSIS OF CORPUS-BASED SPOKEN ENGLISH & COMPUTER-ASSISTED SPEECH TRAINING

Language material database includes two categories: written language database and spoken language database. The former collects the words materials after people's repeated thinking, planning and comparing, which surely misses the realness and the naturalness because the language not only refers to the word, phrase and sentence but also refers to rhythm, intonation, grammar and pragmatic remark and psychological characteristics during speaking. Therefore, the methods of recording and storing the spoken languages materials come into the use. It is commonly thought that new technologies can make a big difference in education. In Norway Professor H.Knut shows the newly-studied corpus with the appearance of voices and words simultaneously like London's COLT and American's SALT at ICAME 2002. [6] The first step of the method is to record, and then to make the videos for later, analyzes the video by voice analysis software PRAAT so that we can hear, watch and even write the voice figures (pitch, length and pause) to realize the corpus showing step by step.

This paper will further explore in the corpus-based spoken language how the cool edit pro 2.1 works in the speech to help the second language learners (L2 learners) make more oral improvements by watching their waveforms in the different stages. The speech chain shows how a wave is formed. Based on the sound wave, we set up the spoken task to make language training. By building Databases, contrasting waveforms of the native speakers, the L2 learners' waveforms before and after training and using the sound digital multimedia technology cool edit 2, we follow the quantitative approach about the phonemes, accent, rhythm, intonation and discourse reading to prove the experimenter's oral performance has been significantly improved.

Cool Edit is a set of digital audio software mixing synthesis in one with sound recording and editing. This software is not only limited to digital music production; it can also be used for making CAI assisted courseware. In addition, the software provides powerful sound analysis function and dynamic processing function to transform its superior sound information into digital audio signal. If we try to transplant this function to the field of qualitative and quantitative study in the phonetics teaching, through making an intuitive comparison between the audio waveforms of standard pronunciations and the audio waveforms of students' pronunciations, then we realize the quantitative assessment for the students' oral training to provide the most effective reference. I n this paper, we will utilize this software to make some spoken language experiments.

4. CORPUS SET-UP & COOL EDIT PRO EXPERIMENT

Make L2 learners choose their favorite videos like famous speakers' speeches and interviews etc, based on which we set up the learners' video database and the corresponding native speakers' video database via cool edit

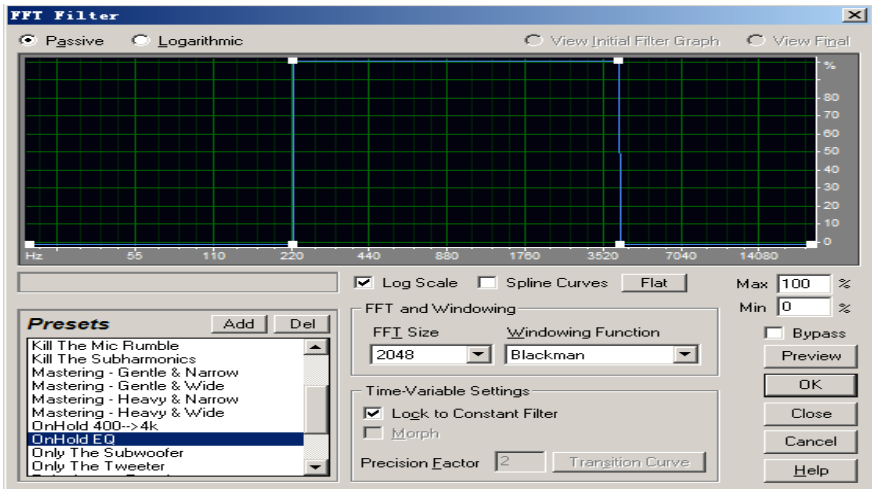


Fig. (4). choose the FFT On hold EQ.

Fig. (5). The contrast among the three waves.

2. Open and set the sampling frequency by 44100Hz and set the accuracy by mono mode16. Remove the noise, separate by every practice and store the file ending with wav. The key point is to set the parameter in order not to distort the sound when remove the noise. See Fig. (**4**).

Then separately make the two groups sound waves database (Database 1 and Database 2), train L2 learners to imitate the chosen video and set up a new database of L2 learners after training (Database 3). Through this recording we get the experiment data after training, with which we make the sound wave to compare with the wave of the native speakers, make the videos and store the achievement in the database.

Use the cool edit to open the recorded three-type videos: the native speaker's, the experimenter's before the training and the experimenter's video after training. Then transform them into three sound waves and insert into multi-track editing window.

In Fig. (**5**)**,** the top track is the L2 learner's acoustic sound waveform without imitation and training. The middle track is the sample of the native speaker. The bottom one is the mimic waveform of the L2 learner after training and imitation for the second time record. The original text to the waves is that "There's Jewish saying that I love. What is truer than truth? Answer the story". See the waveforms expression way of linguistic in Fig. (**5**), Fig. (**6**) and Fig. (**7**).

In Fig. (**6**, **7** and **8**), it is obvious to find the third waveform is better than the first one, which means the L2 learner's oral level has been improved a lot and nearly tend to the native speaker's waveform at the 2nd time recording after imitation. This method is proved effective and remarkable which can be promoting to the oral teaching especially for the L2 learner's autonomous learning and evaluation.

4. RESULTS AND DISCUSSIONS

By using statistic software, we get the results as in Table **2**. Sounds are measured by amplitude (dB), frequency (Hz) and duration (sec). The relationship among the three is that 1 Hertz is equal to 1 vibration per second (or 1 cycle per sec) [7].

By Table **2**, it is obvious to see that recording waveforms from different sound sources are different although under the same circumstances of using the same filter status and mono mode16. RMS in Table **1** refers to a square root acquired by a total average of a group of statistic datum, which is $\sqrt{\left(a1^2 + a2^2 + ... + an^2\right)/_n}$ and that when the sound gets to

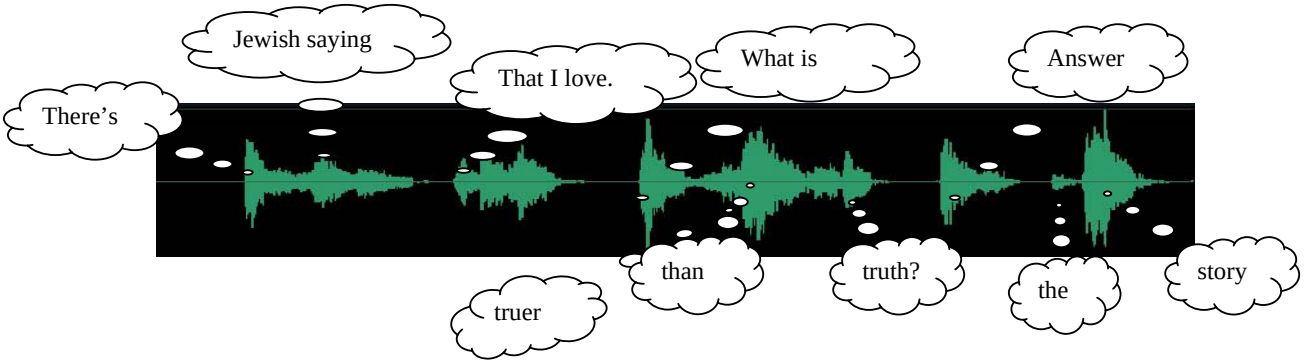
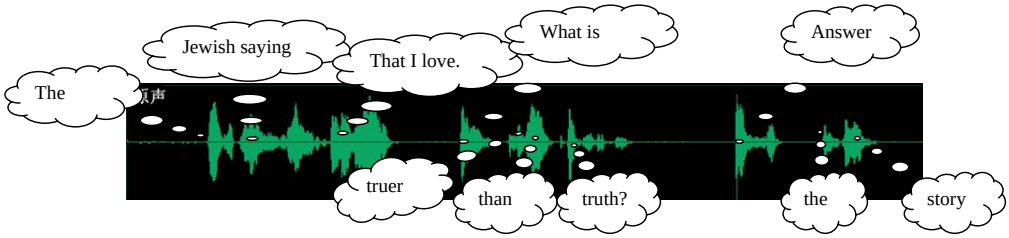


Fig. (6). The original waveform of the linguistic without imitation.



Fig, (7). The sample waveform of the linguistic by the native speaker.

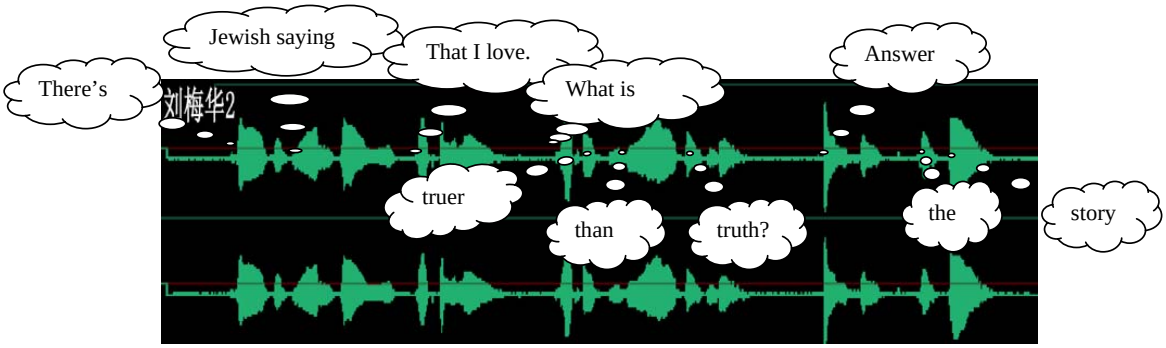


Fig. (8). The L2 learner's 2nd time waveform of the linguistic after imitation.

Table 2. The Statistic Data of the above Three Cases of Sound Waveform

The Statistic Datum of the Three	The Acoustic Waveform	The Ample Waveform	The Mimic Waveform
The mini-sampling value	65536	-32768	-32767
The maxi-sampling value	-65536	27714	26936
The mini-value of RMS excitation	-55.61 dB	-58.6 dB	-59.83 dB
The maxi-value of RMS excitation	-10.19 dB	-8.93 dB	-8.74 dB
The average of RMS excitation	-20.6 dB	-25.57 dB	-23.6 dB
A total of RMS excitation	-18.31 dB	-20.79 dB	-20.5 dB
The bit depth	16 Bits	16 Bits	16 Bits
The width of the window is 5000ms			

0.707 times than the positive and negative peaks during the transforming from the analog signal to digital signal, the waveform of the acoustic wave formed by the different periods of the vibration of the sound digital signal.

Before the imitation, it is obvious to see the offsets of the maximum and minimum sampled values and the standard soundtrack sampled values, in which the size of the maximum and minimum sampled values determine the loudness and the frequency of the sound. It is approved that the speech speed of the acoustic sound without imitation and training is so fast that it makes noises, while after imitation and training the offset of the waveform is less and the sound nearly arrive to the nice and pleasant degree.

5. CONCLUSIONS

It is very abstract and difficult to observe the L2 oral learners' progress change in the regular class teaching and learning. Also it is not real and persuasive that the teachers' assessments are based on the one-time impressions and the personal intuitive perceptions, which is short of the objective and qualitative references. If we apply the digital technology of the sound wave in the oral teaching, the L2 learners can easily observe their own voice status repeatedly and make corrections for the wrong pronunciation to the utmost degree. If the width of waveform is too narrow, it reminds the L2 learners of speaking too fast without rhythms. If the length of waveform is too far, it reminds the L2 learners of speaking too sharp without smoothness. If the size of waveform is smaller than the sample, there is something wrong with the L2 learners' accent [8]. This technology will offer the practical and effective way for the L2 learners.

Modern technology like E-learning, also termed computer-based training (CBT), internet-based training (IBT) or web-based training (WBT), which largely extends to other research fields and includes all forms of electronically supported learning and teaching, and educational technology. For example, in the research field of writing, through E-learning, we collect the raw texts in the L2 learning's' writing and discuss the inter-language roles and the construction of corpus-based written language, trying to find the regularities and the effective factors to improve the L2 learners' academic writing abilities. E-learning often involves both out-of-classroom and in-classroom educational experiences via technology applications and processes such as web-based learning, computer-based learning, virtual education opportunities and digital collaboration. Content is delivered via the Internet, intranet/extranet, audio or video tape, satellite TV, and CD-ROM. It can be self-paced or instructor-led and includes media in the form of text, image, animation, streaming video and audio. It is commonly thought that new technologies can make a big difference in education.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work was financially supported by the Heilongjiang Philosophical and Social Science Program (12H007), High school Education Teaching and Innovation Program of Harbin Normal University and the Youth Exclusive Program of Heilongjiang Science and Plan (GBD1211030) and here I am gratefully acknowledged.

REFERENCES

[1] L. Chen. and H. Wei, "The using of english spoken language corpus in the english spoken teaching", *Computer-Assisted Foreign Language Education*, vol. 103, pp. 23-24, 2005.

[2] W. Lifei and S. Xiaokun, "Current developments in learner english corpus in and outside China", *Computer-assisted Foreign Language Education*, vol. 105, pp. 19-24, 2005.

[3] Coleman J, *Introduction speech and language processing*, Cambridge University Press: Cambridge, 2005.

[4] Y. Yi, "The basic structure of Chinese learners to set up corpus-based spoken English", *Chinese Language Learning*, no. 3, pp. 61-63, 2006.

[5] M.C. Campoy, *Corpus-based approaches to english*, Language Teaching Continuum Publishing Corporation: Australia, pp. 7-27, 2012.

[6] Available from http:www.hit.uib.no/knut/hi/lydfoid.jpg

[7] L.F. Wang, *Computer-aided second language research methods and their applications*, Foreign Language Teaching and Researching Press: Beijing China, 2007.

[8] B. Tu, *A Practical Course of English Pronunciation*, Foreign Language Teaching and Researching Press: Beijing, China 2005.

Received: January 20, 2013 Revised: July 10, 2013 Accepted: October 01, 2013