Corpora and English Teaching: Retrospect and Prospect

He Qiong
School of Foreign Studies
Wuhan University
Wuhan, Hubei, China
Tel: +86 15172532595

Abstract

As a whole system of methods and principles of how to apply corpora in language study, corpus linguistics has revolutionized nearly all branches of linguistics. In the wake of this revolution, people began to rethink language pedagogy from corpus perspective in early 1990s. However, today, although Corpus Linguistics has contributed much to English education, difficulties do exist, especially in Chinese context, where in reality, there is big gap between the wealth of corpus linguistics and teaching practice. The present paper first takes a retrospective look at the current situation of the research on corpora and English teaching, and then sketches out the prospects for future research in the context of Chinese College English education.

Key Words: Corpora, English education, Chinese College English Education

1. Literary Review

1.1 Introduction

corpus (plural corpora) is “a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety ”(McEnery, Xiao & Tono, forthcoming). From this definition, it’s not difficult for us to find the four salient features of corpora, i.e. (1) It must be a large body of authentic texts (including transcripts of spoken data); (2) It needs to be representative of a particular language or language variety; (3) It must be in machine-readable form (e.g. txt files on a computer); (4) It can act as a standard reference about what’s typical in language. Since corpora can “provide better descriptions of a language”(Aston & Burnard 1998:5), to be more specific, they can “provide large databases of naturally-occurring discourse so that analyses can be based on real structures and patterns of use rather than perceptions and intuitions”(Zhang 2000: 9), hence they have become the first-hand data in linguistic analysis, lexicography, natural language processing as well as language learning and teaching. Nowadays, with the combination of rich authentic data in corpora and the power of concordance programs, quantitative analyses can be conducted to reveal previously unsuspected linguistic phenomena. Consequently, corpora have become “a necessary acknowledged resource for students, linguistics, language professionals (teachers, translators, technical writers, lexicographers etc)” (Bernardini, 32). To some extent, it is no exaggeration for us to say that “corpora have revolutionized nearly all branches of linguistics” ( McEnery, Xiao & Tono, forthcoming).

In the wake of this revolution, the interest in integrating computer corpora with foreign language learning and teaching has dramatically increased in the 1990s, though “in initial stages of development, corpora might be used for test selection, test construction, and in the scoring process” (Alderson 1996:253). This interest, As Aston (1998) notes, “[…] derives from the tendency for contemporary language teaching to concern itself with
the learner and the learning process, a tendency above all associated with the “strong” version of the communicative approach, which holds that syllabus and materials design should specify communication tasks through which learning can take place, rather than the linguistic contents to be acquired”(8). Such a perspective puts forward the question as to how computer corpora can help with communication tasks of English language teaching, allowing learners to get competence in English Language in a relatively independent way, as autonomy and self-direction in language learning have generally been acknowledged as an important objective in language pedagogy today, since they are “[...] conducive to increased motivation to learn and consequently to increased learning effectiveness” (Bernardini 2004: 28). According to studies conducted over the last ten or so years, computer corpora improve language pedagogy in two main aspects, i.e. what to teach (i.e. syllabus design and materials development) and how to teach (i.e. teaching methodology).

1.2. What to Teach

It’s well-known that one of the characteristics of communicative language teaching is its preference for authentic texts. The use of authentic data is used to illustrate how real language patterns are actually used to communicate in context. For these reasons, they are also generally considered to be more motivating for learners (cf. Aston 1998). As the major advantage of corpora is to be able to provide authentic language data and statistical data such as frequency data, of which even native-speaker have very little intuitive awareness, they would improve pedagogy by providing more informed choice of which lexical items and grammatical structures to include in the syllabus(Aston 1995; Sinclair & Renouf 1988). Nowadays, teachers take greater interest in using concordance programs to get lists of lexical frequency, syntactical frequency and many other statistical data from a corpus. In turn, these data can be used to inform language learning and teaching. According to the analyses of large general corpora of contemporary English, “[...] traditional language teaching syllabi and materials ignore many linguistic features which are frequent in native-speaker data, and emphasize ones which are relatively rare”(Aston 1998:8). This finding is in accordance with the result of Mindt’s (1996) observation of different use of grammatical structures between textbooks for teaching English and L1 English. Thus he concludes that one common failure of English textbooks is that they teach “a kind of school English which does not seem to exist outside the foreign language classroom” (1996: 232). He then further argues that teaching syllabuses should be based on empirical evidence rather than tradition and intuition with frequency of usage as a guide [...] (cf. 1996: 245-246). Later on, researchers come up with plenty of ways in which frequency information and many other statistical data extracted from corpora might inform materials development and syllabi design. Coxhead (2000), for example, describes the development of an academic wordlist, which could be used to inform the design of EAP vocabulary, reading and writing materials. Clear(2000) presents a pattern grammar approach to ESL materials and syllabi design, to name just a few.

1.3. How to Teach

In addition to informing what to teach, computer corpora can also provide insights into how to teach. In 1991, Johns, who is among the first to realize the potential of corpora for language learners, proposed the “data-driven learning” approach to language learning and teaching, which “engages learners in the analysis of concordance lines that have been selected, arranged, and possibly edited by the teacher in order to draw the attention to patterns of language use”(Keck 2004: 94), thus it can “focus[...] on the role of corpus use in the development of learning capacities and in the establishment of a non-authoritarian learning environment”
According to Johns, “research is too serious to be left to the researchers”: that the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data” (1991: 2). In his opinion, learners should be directed by teachers to discover the foreign language, almost in the same way as corpus linguistics discover facts of their own language that had been ignored in the past. He further put forward three stages of inductive reasoning with corpora in the DDL approach: observation (of concordanced evidence), classification (of salient features) and generalization (of rules). As Bernardini (2004) sees it, “This shift of emphasis from deductive to inductive learning routine has wide-ranging effects on: (a) the teacher, who becomes a coordinator of research, or facilitator; (b) the learner, who learns how to learn through exercises that involve the observation and interpretation of patterns of use…”(16). Inspired by the “data-driven learning” approach favored by Johns, another approach known as the “discovery learning” has been explored, which encourages learners to generate their own concordance data, hence be more autonomous and responsible for their studies. This “learning as discovery” view regards the learner as traveler instead of researcher. It allows learners to follow their own interests in learning from corpora, thus “their searches become better focused, their interpretation of results more precise, their understanding of corpus use and their language awareness sharper” (Bernardini 2004: 23). Moreover, a non-authoritarian environment is created and the democratization of the learning setting is facilitated by corpus-based discovery learning. Therefore, it is considered to be more advantageous than the DDL approach in that first, learners are in charge of their own learning, and thus motivation is increased (Aston 2001); second, learners feel empowered to challenge textbook explanations and native–speaker intuitions (Lorenz 2000; Mair2000); last but not least, learners make ‘serendipitous’ discoveries about language use that, without the use of corpora, would not have been possible (Burnardini 2001).

1.4. Computer Learner Corpus

Though data from native-speaker corpora is substantial, to further apply corpora to language learning and teaching, teachers, especially teachers who emphasize learner-centeredness and active-learning, should become familiar with findings from both large native-speaker corpora (reference corpora) and target learner corpora which, according to Granger (2004), (a) have been gathered in authentic contexts (including classroom activities); (b) consist of ‘continuous stretches of discourse, not isolated sentences or words’; (c) are ‘compiled using strict design criteria’; (d) and are ‘collected for a specific SAL or FLT purpose (7-10). As noted by Aston (2000), “

It is [...] debatable how far the analysis of native-speaker behavior can provide an adequate syllabus design [...] because such specifications provide no information as to the relative difficulty and learnability of particular features to be taught: research into interlanguage grammars suggest that second language acquisition follows a relatively fixed sequence of progressive approximation to native-speaker competence, a sequence which teaching can do little to modify (Ellis, R). Relevant information on interlanguage development may be provided by corpora of non-native data, which can be used to code and classify recurrent errors, along with over and under-uses, with a view to identifying features which teaching should perhaps emphasize, and to evaluating their difficulty.[...] (11)

Therefore, attention should be paid to developing learner corpora. As Granger (Granger) argues, the benefits of learner corpora are as follows:

The efficiency of EFL tools could be improved if material designers had access not only to authentic native data but also ... to authentic learner data, with the NS (native speaker) data giving information on what
is typical in English, and the NNS (non-native speaker) data highlighting what is difficult for learners in general and for specific groups of learners. (176)

To date, learner corpora have mainly been integrated with language teaching methodology and materials and syllabus design. According to Bernardili (2004:19), The assumption behind these attempts lies in the fact that

...the learning process may be aided by form-focused instruction and access to focused negative evidence. In other words, if learners are presented with concordances showing the typical errors they statistically appear to make, and with similar textual environments where the same structure is used appropriately, they may find it easier to become aware of more or less fossilized characteristics of their interlanguage, thus potentially initiating a process of knowledge restructuring.

1.5. Difficulties and Problems

From the above illustration, we have every reason to acknowledge that “corpora constitute resources which, placed in the hands of teachers and learners who are aware of their potential and limits, can significantly enrich the pedagogic environment”(Aston, 1995, 261). However, at present, although Corpus Linguistics has contributed a lot to English education, difficulties do exist and there are still many unanswered questions in this area. As for materials design, since in most corpus studies, “the implications for pedagogy are not developed in any great detail with the consequence that the findings have had little influence on ESP […] materials design”(Flowerdew 1998:550), we would agree with Granger (2004: 136) that “with the exception of EFL dictionaries, the number of concrete corpus-based achievements is not proportional to the number of publications advocating the use of corpora to inform pedagogical practice.” Similarly, as for classroom use of corpus data, as Tribble (2000: 31) admits, “not many teachers are using corpora in their classrooms.” What is even worse, it seems that teachers are reluctant to introduce “discovery learning” in their everyday teaching practices (cf Mukherjee 2003). Therefore, much is to be explored on the extent to which English language teachers have received training in corpus linguistics, and on teacher and student attitudes towards learning about and using corpus methods in the classroom (cf Keck 2004: 93-96). Moreover, as Aston (2001) points out, few empirical studies have been conducted to investigate the effectiveness of DDL or discovery learning in promoting language acquisition (Keck, 2004: 95-96). All in all, undoubtedly, more spadework should be done to bridge the gap between corpus linguistics and the reality of English Education, so as to make English teaching more effective.

2. Prospect of Using Corpora in College English Education in China

In the light of those difficulties mentioned above, to help bridge the gap between applied corpus linguistics and the real practice of learning and teaching English as a foreign language, especially in the context of Chinese college English Education, where the influence of corpus methods is very limited, in-depth research should be conducted on the research questions as follows,

Research question 1: How popular is corpus linguistics among college English language teachers and how do to popularize it?

Research Question 2: How to design corpus-informed teaching materials for college students in China?

Research question 3: What’s the effectiveness of DDL in promoting learning and teaching English as a foreign language in Chinese college English education and what’s students’ attitude towards using corpora in the college English classroom?
2.1 Suggested Research Methods

2.1.1 Workshops and questionnaire

As a matter of fact, corpora in China is not as popular as in the UK, even at tertiary level. As far as I know, few college English teachers in China have the basic knowledge of corpus linguistics, not to mention to conduct corpus-based research, which makes it impossible for students to access corpora by themselves or even guided by teachers. Therefore, there’s urgent need to popularise corpora among College English teachers in China. To popularise corpora among college English teachers in China and to investigate the extent to which they know about corpora, researchers can adopt the methodology Mukherjee (2003) employed in German context i.e. introductory workshops and questionnaire.

2.1.1.a. Introductory Workshops on Corpus Linguistics

Researchers can help run five 2-hour introductory workshops for English language teachers. Each two-hour workshop consists of one-hour lecture and another one-hour seminar. The overall structure of the ten-hour workshops can roughly be designed as follows,

- The participants are to spend four hours auditing 4 lectures which are to familiarize them with some key concepts in corpus linguistics (e.g., the history and the current development of corpus linguistics, major corpora of present-day English, concordances, collocations, word-lists and keywords, corpus annotation, etc).
- The participants are to spend four hours taking part in 4 seminars in the lab to put what they’ve learned in the lectures into practice. They’re to be introduced how to use the BNCweb and Wordsmith to get concordances, collocations, wordlists and keywords from either plain corpora or annotated corpora.
- The participants are to spend one hour in the fifth lecture on the pedagogical application of corpora. Then in the fifth seminar, they are to devote themselves to practically learning some applications of corpus data in the classroom (e.g. using a corpus to correct learners’ errors and using a learner corpus to teach grammatical patterns, etc.).

2.1.1.b. Questionaire

To empirically assess how popular corpora are among English language teachers in China, Researchers can ask each teacher in the workshops to fill in a questionnaire. Some questions are about teachers’ knowledge of corpus linguistics before the workshop, others are designed to investigate the teachers’ attitudes towards corpora and the role of corpora in pedagogy after the workshop. Since Mukherjee (2003) has already conducted a successful survey on the current situation of corpus linguistics in Germany, researchers can adopt the form of the questionnaire in his study. Some questions in the questionnaire are as follows,

**Questionaire: The role of corpus linguistics in English language teaching in China**

1) Before the workshop: Are you familiar with corpus linguistics?
   - Yes, I am familiar with corpus linguistics (> university studies)
   - No, I am not familiar with corpus linguistics but I have already heard of it (> colleagues, books/articles, conferences, etc.)
   - No, I don’t know anything about corpus linguistics

2) After the workshop I: Do you think that teachers and/or learners may profit from corpus data?
3) After the workshop II: In which particular fields would you consider consulting or using corpus data in the future? (multiple answers possible)

- Creation of concordance-based teaching material (> teaching of collocations, patterns, spoken/written differences, etc.)
- Correction of class tests (> acceptability/idiomaticity of collocations, patterns, phrases, etc.)
- Creation of word/phrase lists for individual text collections (> set books, texts in `bilingual subjects' such as history and geography in English medium, etc.)
- Learner-centred activities (> consultation of corpus data, small-scale corpus studies, corpus browsing, large-scale term papers, etc.)

2.1.2 Designing Corpus-informed Teaching Materials

With the increase of the number of college students in every college and university China, there’s urgent need for corpus-informed teaching materials for them, as most College English teaching materials used in the past 10 years are either too old-fashioned or intuition-based, which fail to identify the real problem areas of our present students, especially in those universities where teachers are required to write teaching materials for their own students.

2.1.2.a. Building “DIY” English learner corpora for College English students

To help design corpus-informed teaching materials for future College English Students in China, relevant written English learner corpora must be built as in the Chinese learner English corpus, there are only two sub-corpora of Chinese college students, i.e. band 4 corpus and band 6 corpus. In building their “DIY” corpora, teachers may ask each student to write an argumentative essay on the same topic timed or untimed. Students are allowed to use language reference tools such as dictionaries and grammar books, but should entirely work on their own. They can’t copy others’ articles or ask a native speaker for help. When they’ve finished, to save time and labor, teachers may ask them to hand in their essays with their own individual information such as age, gender and geographical region on disk (in text format). After that, teachers can POS-tag or error-tag the whole corpus according to the need of teaching or research.

2.1.2.b. Contrastive Interlanguage Analysis

The ”Contrastive Interlanguage Analysis” (CIA) proposed by Granger (2002) compares learner corpora with native-speaker corpora, or compare different varieties of interlanguage. After building and POS tagging their DIY learner corpora, teachers may try the CIA to compare frequency lists from the learner corpora and such control English native corpus as the BNC or FLOB by using the “compare list” facility in Wordsmith Tools (Scott, 1999) and the “key category”(obtained by the comparison with the BNC Sampler) facility in Wmatrix (Rayson, 2003), so as to highlight the words, phrases, grammatical items and structures either strikingly under- or overused by learners and accordingly emphasize them in the new teaching materials.
2.1.2.c. Keywords Analysis to Identify Spelling Errors of Learners

According to Scott (1999), keywords are “those whose frequency is unusually high in comparison with some norm.” With Wordsmith Tools, the keywords are worked out by first making a wordlist for the corpus under study (the one shown on the left) and a wordlist in the reference corpus (also called reference wordlist), then comparing the frequency of each word in the wordlist on the left with the frequency of the same word in the reference wordlist. A word becomes key if it’s unusually frequent or infrequent in comparison with what one expects on the basis of a reference wordlist. Keywords are important because they provide a useful way to characterize a corpus or genre. They help find out which words characterize the corpus under study. Therefore, they can be regarded as lexical signposts, revealing the lexical features in a corpus. Teachers can compare the learner corpora they’ve built against the 100 million-word BNC or the one-million-word FLOB, the reference corpus, and two wordlists containing the frequency of each word in both corpora are generated. Then by cross-tabulating frequencies of all lexical items against each other and the number of running words in the two corpora, a list of keywords are created. As words with spelling and grammatical errors only occur in the learner corpus, therefore they are in the keyword list. If a word is spelled incorrectly more than 5 times in a corpus, it can be thought of as problematic and should be brought to the attention to the students. So through the keyword analysis, we can find out words with spelling and grammatical errors which occur more than five times, and then include them in the vocabulary list of the teaching materials.

2.1.3 Top-down and bottom-up approach in English Classroom

In corpus linguistics, the so called “top-down” approach starts 'downwards' from the data in an English native corpus, helping learners to generalize the lexico-grammatical rules lying behind the data. On the contrary, the so called “bottom-up” approach which was first described by Tribble in 1988, goes 'upwards' from the data in a learner corpus, making learners aware of their errors. So far data-driven learning, defined by Johns and King (1991:iii) as “the use in the classroom of computer generated concordances to get students to explore regularities of patterning in the target language, and the development of activities and exercises based on concordance output”, has exclusively adopted the “top-down” approach. However, it’s undeniable that this approach fails to identify the errors and difficult areas for learners. Since error identification and correction help to enhance learners’ awareness of aspects of the grammatical system of the target language, it is therefore extremely fruitful to combine the two approaches and use native data as well as learner data in DDL. To investigate empirically the effectiveness of DDL in Chinese college English classroom, teachers may combine the “top-down” approach and the “bottom-up” approach in a randomly selected class. In this class, Teachers can make use of the collective electronic data which the students hand in as assignments, guiding them to compare the "bottom-up" data in their own interlanguage productions with the "top-down" data in such English native corpus as the Flob or BNC through various corpus-based exercises, so as to enable them to better notice gaps between their own uses and those in the target language and to help them correct errors in their own interlanguage. And at the same time, surveys can be conducted by questionnaires among students instructed by DDL to investigate students’ attitudes towards corpus methods in their college English classroom.
3. Concluding Remarks
The above has taken a retrospective look at the current situation of the research on corpora and English Teaching and sketched out how to bridge the large gap between the wealth of corpus linguistics and the teaching practice in China, where corpus linguistics hasn’t yet found its way into the reality of college English learning and teaching. Bridging this gap is very challenging to College English education in China, however, I believe as more and more college English teachers get familiar with the basics, implications and language-pedagogical applications of corpus linguistics through such training as introductory workshops, Corpus-informed teaching materials can be designed and truly learner-centered corpus activities can be implemented and popularized in the college English classrooms in China in the near future.

References


**Acknowledgement**

This paper is sponsored by the CSC and the School of Foreign Studies of Wuhan University.