The Development of Corpus Linguistics in English and Chinese Contexts

Mcenery, Tony

Xiao, Richsrd

(Citation)
Learner Corpus Studies in Asia and the World,2:7-45

(Issue Date)
2014-05-31

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81006688

(URL)
https://hdl.handle.net/20.500.14094/81006688

The Development of Corpus Linguistics in English and Chinese Contexts

Tony MCENERY
Richard XIAO
Lancaster University

Abstract

In this paper we review the state of the art in corpus linguistics in both English and Chinese in three areas – lexicography, grammar development and learner corpora. In doing so, we also contrast and compare the development of these areas for both languages. We discover that, while the lexicographic tradition in China means that work in English corpus based lexicography and Chinese corpus based lexicography bear comparison, in the other two areas the development of work in English is markedly more advanced than that in Chinese. However, were relevant we also look to future developments that hold the promise to change this situation.

Keywords

Corpora, Chinese, English, Lexicography, Grammars, Learner corpora

1 Introduction

In this paper we will survey progress in English corpus linguistics (henceforth ECL)[1] and contrast that with the development of Chinese corpus linguistics (henceforth CCL). In doing so we will be comparing the relative state of development of the two, highlighting areas where the development of one may present a profitable avenue of development to the other. We will also identify challenges which are unique to either, while considering also areas where generalising from one to the other may lead to problems. Such a review could become unwieldy if every area of linguistics were to be reviewed. While in passing some areas of linguistic theory will be considered, this review will focus principally upon language description and applied linguistics, areas where, arguably, corpus linguistics has made its greatest impact to date. The areas considered by this paper are as follows: lexicography, descriptive grammars and learner corpora. In each case we will review the development of these areas in ECL, outlining key debates and changes brought about by the application of corpora in the area in

question. In all cases we will take a broadly historical approach, tracing the impact of corpora upon the research area by identifying key points in the development of ECL as well as key researchers who brought about that development. We will then contrast the ECL approach to that area with development to date in the area by CCL. We will then consider the contrast between the two.

II English Corpus Linguistics and Lexicography

While lexicography has always typically called upon spontaneously occurring data, for a long time it did so in the form of citation slips, short records of examples of usage. While the use of such slips became increasingly sophisticated over time, the methodology itself was limiting. The observations that generated the slips were highly partial, and prone to skew based upon the observations of the lexicographer or of the observations reported to them. They were also fundamentally limited – the organisation and exploitation of such slips through manual cataloguing was slow and error prone. Even where such slips were computerised, there was still the basic issue of bias in observation to overcome. Also, citation slips were never able to produce reliable data on the frequency of the words in question and were not well suited to exploring how a word and its meaning may vary by context or genre. In ECL the process of overcoming these limitations is closely linked to research developments at the University of Birmingham where, under the leadership of John Sinclair in the 1960s and 1970s, a corpus-based approach to lexicography developed (see Sinclair *et al.* 1970, for instance). While earlier corpora such as the Brown corpus (Francis and Kucera, 1964) had been consulted by dictionary builders (the Brown corpus was consulted for the first edition of the *American Heritage Dictionary* in 1969), the impact of the corpora upon the practice of lexicography and the content of dictionaries was initially minimal. In large part this was because of the limited size of such early corpora. One million words, while impressive by the standards of the time, is an insufficient sample on which to build a dictionary with any pretensions to be comprehensive.

The real change in the use of corpora in lexicography came in 1980 when Sinclair started a partnership with the publishing company Collins to set up COBUILD (the Collins-Birmingham University International Lexical Database) as a research facility. As part of this a very large corpus, the Bank Of English, was established which could be used as the basis for a comprehensive dictionary.[2] COBUILD provided data, ideas and analyses for Collins, to help them develop a new corpus-based dictionary (the Collins COBUILD dictionary, 1987). To get a sense of the advances that corpus lexicography made possible, consider the following example from Hanks (2009: 216), relating to insights gained by use of the COBUILD corpora in the 1980s into *-ly* adverbs:

"It was widely assumed by pre-corpus lexicographers that all (or almost all) *-ly* adverbs

were adverbs of manner modifying the sense of a verb, an adjective, or another adverb and that the meaning of the adverb was always (or almost always) systematically derivable from the root adjective. There is some truth in this, of course: *walking slowly* is walking in a slow manner. But some *-ly* adverbs in English have special functions or constraints, which were not always well reported in the pre-corpus dictionaries. The *Oxford Advanced Learner's Dictionary* (OALD1-4) says nothing about the use of words like *broadly*, *sadly*, *unfortunately*, *luckily* and *hopefully* as sentence adverbs – linguistic devices that enable speakers and writers to express an opinion about the semantic content of what they are saying. Those pre-corpus dictionaries which did notice sentence adverbs did not succeed in noticing all of them systematically."

The focus upon Birmingham here is appropriate as it was there that the revolution in ECL lexicography began. However, an exclusive focus on Birmingham runs the risk of obscuring the fact that what began at Birmingham eventually became a general trend in lexicographic practice, as Hanks (2009) outlines in detail. Corpus data forced lexicographers to shift away from a focus upon regularities which were easily attested by a few examples towards what Hanks (2009: 216) describes as 'the idiosyncratic conventions that are associated with each word'. As a result of this shift, the descriptive adequacy of the dictionaries produced improved markedly, revealing pre-corpus lexicography to be 'little more than a series of stabs in the dark, often driven by historical rather than synchronic motives' (Hanks 2009: 230). Frequency, collocation, authentic illustrative examples, and contextual and genre variation are all forms of data which now commonly appear in dictionaries because such data can be derived readily from large corpora. It is now difficult to find a major dictionary of British English which is not based upon corpus data, for example, *Longman Dictionary of Contemporary English* (LDOCE, 3rd edition), *Oxford Advanced Learner's Dictionary* (OALD, 5th edition), *Cambridge International Dictionary of English*, and *Macmillan English Dictionary*. The trend begun at Birmingham, for the reasons outlined above, became very dominant, very rapidly in British dictionary publishing. The ability of the corpus to address the observation biases inherent in the citation slip method and its capacity to provide fresh insight into, and a better description of, language, relegated the previously dominant citation slip method to the role of a minor supporting method in dictionary production.

In doing so, some other trends were spawned by ECL. Sinclair's ideas regarding language in general were closely linked to his work in lexicography – he took a lexically based approach to explaining as well as describing the language system. His work emerged from earlier ideas developed by J. R. Firth (1957). Sinclair also worked as part of a network of scholars such as Michael Halliday, whose approach to data shaped and influenced Sinclair's own views (Sinclair 2004a: vii). Sinclair took a corpus-driven approach to the analysis of English grammar, in which words and grammar were

ineluctably bound in what has been called *lexicogrammar* (Halliday 1985, Sinclair 1991 – though Sinclair preferred the term *lexical grammar*). Sinclair collaborated with a number of linguists at Birmingham who, while working in the tradition that Sinclair championed, developed that approach to corpus linguistics further. For example, Susan Hunston has been central in advancing the lexicogrammatical approach to the analysis of language, developing Pattern Grammar, a model where language is built up of a series of linked sequences of fuzzy structures, within which collocation provides both structural coherence and meaning (Hunston and Francis 1999). So while the work of ECL in lexicography has led to a thorough revision of British English dictionaries, it has also spawned a distinct approach to the explanation of the linguistic system.

III Chinese Corpus Linguistics and Lexicography

Just like English, Chinese as the world's most spoken language has also benefited greatly from the development of corpus linguistics. While there is no doubt that the development of corpus linguistics initially overlapped with that of ECL, Chinese corpus linguistics developed almost at the same pace with ECL. The first study of Chinese character frequency in a modern sense dated back as early as the 1920s, with the publication of Li Jinxi's (1922) 'Statistical study of basic vocabulary in Chinese'. Also in the 1920s, Chen Heqin, a renowned educationalist in China, assembled a paper-based corpus composed of a range of materials such as children's books, after-class readings for school children, newspapers, women's magazines, classical and modern Chinese fiction, amounting to well over half a million Chinese characters. Chen and his nine assistants worked for 2-3 years to establish a list of 4,261 most frequently and widely used Chinese characters, which was first published as *The Applied Glossary of Modern Chinese* (《語體文應用字彙》) in the fifth issue of the fifth volume of *New Education* in 1922. It was revised later and republished as the book of the same title by the Commercial Press in 1928. Chen's character frequency list played an instrumental role in promoting literacy in primary education in China by providing empirical evidence for the development of primary teaching materials such as *Thousand Character Lessons for Civilians* (Tao and Zhu 1923), which was used to teach over half a million civilians basic Chinese characters in more than twenty provinces.

Chen's (1922, 1928) frequency list of Chinese characters was profoundly influenced by Thorndike's (1921) *The Teacher's Word Book*. However, the contribution of the former to Chinese is arguably more significant than the contribution of the latter to the English language. Chinese as a script language is different from phonetic languages such as English. While such phonetic languages usually have an alphabet of 20-30 characters that make up words which are typically delimited by spaces in writing, there are 6,000-7,000 Chinese characters in general use (with a possible total of over 50,000), many of which are used as monosyllabic words. With such a large character set, it is of

crucial importance to identify both a basic set of the most common Chinese characters and the words, which are most frequently and widely used as a means of promoting literacy. Chen's (1922, 1928) character list is particularly important because it was established in a timely manner, at the initial stage of the development of modern Chinese as used today. The New Culture Movement in 1919 is generally recognised as the demarcation line between classic Chinese and modern Chinese. As such, Chen's *The Applied Glossary of Modern Chinese* has not only contributed to primary education and the promotion of literacy in China, it has also helped to shape present-day Chinese.

While the data used in Chen (1922, 1928) was not computerised, his list of basic characters was nevertheless corpus-based; it is the forerunner of today's word frequency lists and frequency dictionaries of Chinese derived from computer corpora. For example, since the founding of the People's Republic of China in 1949, the central government and local authorities have also published a range of lists of Chinese words and characters, including for example, 'Register of Common Characters' published by the Ministry of Education in September 1950 which contains 1,017 Chinese characters, 'List of Common Characters' published by the Ministry of Education in June 1952 which contains 2,000 Chinese characters, 'List of Common Characters in *Putonghua* Common Speech' published by the Shandong Provincial Department of Education in August 1958 containing 3,000 Chinese characters, 'Three Thousand Common Words in *Putonghua* Common Speech' published by China's Committee for Chinese Written Language Reform in 1962, 'A List of Four Thousand Words for Foreign Students' published by the Beijing Language and Culture University in 1964, and 'List of Common Used Characters' published by the Beijing Municipal Commission of Education in March 1965 which contains 3,100 Chinese characters (cf. Wang 2006).

The examples cited above demonstrate that Chinese linguists have a long standing tradition of studying word and character frequency, because of the size of Chinese character set. This tradition has, unsurprisingly, continued, with the rapid development of corpus linguistics in general and Chinese language processing in particular, into the 1980s and 1990s as well as the new millennium. For example, Liu's (1973) *Frequency Dictionary of Chinese Words* gives statistics such as frequency, dispersion index and usage rate for 3,059 most frequently used words in Chinese on the basis of a 250,000-word corpus covering five registers (fiction, drama, essays, newspapers and periodicals, technical writing); *A Comprehensive Frequency Table of Character Usage in Modern Chinese* (《現代漢字綜合使用頻度表》) established by Project Code 748 (1976) lists 4,152 frequently used characters on the basis of a corpus of 21 million characters; Beijing Aeronautical University's (1985) *A Frequency Table of Character Usage in Modern Chinese* (《現代漢語用字頻度表》) lists frequently used characters for ten genres and technical domains on the basis of samples totalling 11.08 million characters; *A Frequency Dictionary of Modern Chinese* (《現代漢語頻率詞典》) developed by Beijing Language and Culture University (1986) lists 16,593 commonly used words extracted

from 1,315,752 word tokens (or 1.82 million characters); the National Language Committee's (1988) *Commonly Used Characters in Modern Chinese* (現代漢語常用字表) lists the most commonly used 2,500 characters and 1,000 commonly used characters on the basis of data collected by Beijing Aeronautic University covering the period 1928-1986; Hong Kong Polytechnic University's (1991-1997) *A Chinese Word Bank from Mainland China, Taiwan, and Hong Kong* (《中國大陸、臺灣、香港漢語詞庫》) lists 68,011 entries based on a six-million-character corpus of news texts published during 1990-1992 in the three Chinese speech communities. These corpus-based frequency lists and dictionaries are essentially targeted either at native speakers of Mandarin learning their mother tongue (e.g. Chen 1928; National Language Committee 1988), or at language engineers (e.g. the frequency list by Project Code 738) and expert Chinese linguists (e.g. the word bank by Hong Kong Polytechnic University for studying language variation). More recently, Xiao, Rayson and McEnery's (2009) *A Frequency Dictionary of Mandarin Chinese*, which is based on a sizeable balanced corpus of current spoken and written language from four broad categories – namely Spoken, Fiction, Non-fiction, and News, totalling approximately 50 million word tokens (or 73 million Chinese characters), provides a list of the top 5,000 Chinese words and 2,000 Chinese characters, thus defining a core vocabulary for learners of Chinese as a second or foreign language.

With regard to non-native language learning resources for Chinese, one should not fail to mention the Syllabus of Graded Words and Characters for Chinese Proficiency compiled by the Chinese government's *Hanyu Shuiping Kaoshi* (the Chinese Proficiency Test, HSK) Committee, which was published in 1992 and revised in 2001. The HSK lexical syllabus lists the words and characters required of learners of Mandarin Chinese as a second or foreign language to pass the Chinese proficiency test HSK, as indicated in Table 1. According to the HSK lexical syllabus, learners of Chinese as a foreign language who have learnt about 5,000 words will be able to express their ideas on general issues in Chinese.

Table 1. HSK graded lists and words and characters in Chinese

HSK level	Words	Characters
Level 1	1033	800
Level 2	2019	803
Level 3	2205	591
Level 4	3583	671
Levels 1-3	5257	2194
Levels 1-4	8840	2865

As can be seen in Table 1, this vocabulary is approximate to the total number of words in HSK Levels 1-3. The number of words covered in Xiao *et al.*'s (2009) frequency dictionary is roughly comparable. While it is certainly true that the larger a learner's vocabulary the better the language learner's proficiency, it is nonetheless increasingly difficult to learn new words as the learner's vocabulary grows. This is because, according to Zipf's law, the frequency of a word is reversely proportional to its rank in the frequency table. As such, there is a 9.27% increase in coverage from top 1,000 to top 2,000 words, whereas the increase in coverage drops to 0.94% from top 8,000 to top 9,000 words (see Table 2). In addition to the reference to the HSK syllabus, Xiao *et al.*'s (2009) decision to include 5,000 words is also empirically based by the sharp drop in coverage (from 3.07% to 1.77%) from top 5,000 to top 6,000 words.

Table 2. Coverage of top N words

Top N words	Coverage (%)	Increase (%)
1000	66.24	--
2000	75.51	9.27
3000	80.53	5.02
4000	83.81	3.28
5000	86.17	3.07
6000	87.94	1.77
7000	89.33	1.39
8000	90.47	1.14
9000	91.42	0.94

As can be seen in Figure 1, which shows the increase in vocabulary size and the drop in coverage resulting from each additional block of 200 characters, Zipf's law also applies to characters. Coverage grows very slowly after the top 1,200 characters. The top 2,000 characters cover nearly 98% of our whole corpus, with 4,839 characters accounting for the remaining 2% of coverage.

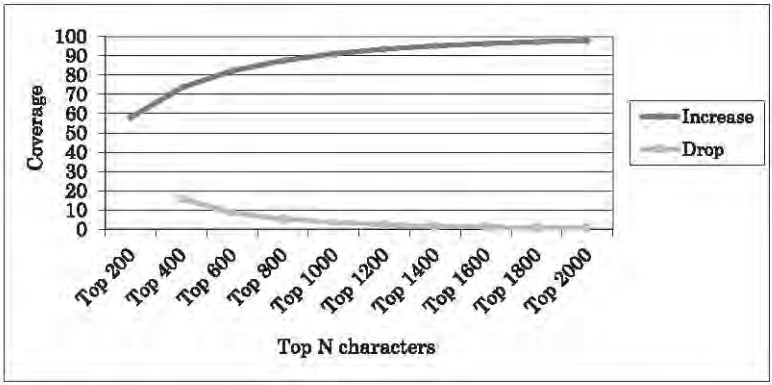

Figure 1. Coverage of top N characters

Corpora can be said to have an important role to play in all aspects of lexicography, ranging from selecting head words for inclusion in a dictionary, identifying word senses, ordering of polysemous and homograph items to determining a word's word class and providing illustrative examples of its use (cf. Wang 2010). Cui (2011: 85) provides a good summary of the various roles of corpora in compiling *New Word Dictionary for Chinese as a Foreign Language*.

Table 3. A summary of the roles of corpora in lexicography in Chinese

Dictionary microstructure	Roles of corpora
Headword	Retrieving new words
	Identifying headwords
	Identifying synonyms
	Determining primary and secondary headwords for cross reference
Pinyin pronunciation	Deciding the Pinyin gloss of words with multiple pronunciations
Word class	Determining the word class of the headword
	Decide the order of word classes of words with concurrent classes
Translation	Assisting in translation of the headword

Linguistic interpretation	Mining for pragmatic properties of the headword
	Mining for syntactic properties of the headword
Definition	Assisting in definition of the headword
	Determining the order of senses of polysemous words
Illustrative example	Providing authentic examples in context
Knowledge window	Providing background knowledge of the headword

According to Xiao (2010), a corpus that is annotated with word senses can provide a major means of verifying whether the definitions in a dictionary are reasonable and complete. Jiang (2005) observes that major Chinese dictionaries, when evaluated against large-scale corpora, suffer from a number of drawbacks including for example, omissions of some commonly words or word senses, or word senses too narrowly defined, and lack of illustrative examples.

According to Li (2002: 95), the process of identifying and arranging word senses in dictionary making can be summarised as three steps: collecting corpus data, selecting materials from the corpus according to the purpose of the dictionary, and identifying word senses. This is traditionally a subjective process relying on a combination of slip cards and the lexicographer's intuitions, which is likely to result in omissions. For example, the word 愛護 'cherish, treasure; take care of; love and protect' is defined in authoritative dictionaries as 愛惜、保護 'cherish; protect' in *A Dictionary of Chinese* and as 愛惜並保護 'cherish and protect' in *A Dictionary of Modern Chinese*. On the other hand, corpora can provide more objective criteria for identification and ordering of word senses in the dictionary on the basis of exhaustive quantitative and qualitative analysis. For example, a random selection of 1,000 concordance lines of 愛護 from a sizeable corpus of modern Chinese shows that the word is used in three senses: (a) in 651 examples (67.46%), it means 關心、關愛 'care for or show concern for' (typically seniors to juniors, or elder to young people); (b) in 221 examples (22.90%), it means 保護，使其不受損害 'protect (from being damaged or hurt)'; (c) in 93 examples (9.64%), it means 珍惜、不浪費、不毀壞 'cherish or treasure, not waste or damage'. Clearly, the first two are the major senses of the word, and they are interrelated. They should be listed as separate senses in a medium-sized dictionary because of their different collocational behaviour and grammatical properties. On the one hand, word sense (a) typically (95% of the time) collocates with nouns or pronouns denoting humans or animate objects whereas word sense (b) usually co-occurs with nouns and pronouns referring to inanimate objects (77%). On the other hand, 愛護 with sense (a) can be used either as a verb (71% of the time) or as a noun (27%) or as an adjective or adverb, while sense (b) is

normally used as a verb. It is suggested that according to frequency data from the corpus, the word senses of 愛護 should be revised as follows, in the given order: (a) verb or noun: 關心、關愛(一般對人或生命体，多用於上對下、長對幼) 'care for or show concern for' (generally human or animate object; typically seniors to juniors, or elder to young people); (b) verb: 保護，使(物體)不受損害 'protect (usually an inanimate object) from being damaged'; (3) verb: 珍惜、不浪費、不毀壞 'cherish or treasure, not waste or damage' (see Huang 2004). In this arrangement, word sense (c) is close to sense (b); therefore in a small-sized dictionary, it can be merged into sense (b).

Language keeps changing over time. Hence, an important area of lexicography is the study of neologism, which can benefit greatly from the corpus approach, because corpora provide the necessary sources of data as well as the method for reasonably identifying new words or new usage of existing words. *A Dictionary of New Words* (Kang 2003) is the largest Chinese neologism dictionary of its kind, containing 20,000 new words that have gained currency and remained relatively stable between 1978 and around 2000. The dictionary is based on a huge corpus composed of 25 years' data (1978 – 2002) from the *People's Daily* newspaper, in addition to recent data from twelve newspapers of the *People's Daily* newspaper family, nine websites of the people.com.cn family, the *Readers* magazine (previously known as *Readers' Digest*) since its founding, the *Southern Weekly* newspaper since its founding, as well as online data from newspapers including *Guangming Daily*, *China Education Daily*, and *Southern Daily*. Similarly, Tsou and You's (2010) *The Global Dictionary on Chinese Neologism* includes 1,600 Chinese neologism that have entered the Chinese language since 2000. These new words are carefully selected from the over 20,000 new words from LIVAC (Linguistic Variations Across Chinese Communities), which is a gigantic, homothematic and synchronous Chinese corpus designed to monitor Pan-Chinese language development and underlying Pan-Chinese changing cultural trends. The monitor corpus samples representative local media texts published in Chinese, covering main news, editorials, international news, local news, entertainment, sports, finance, etc. from Chinese speaking communities in Hong Kong, Taipei, Macau, Singapore, Shanghai, Beijing, Canton, Shenzhen, and Japan. As of 2010, the corpus has included over 16 years' data, amounting to 400 million Chinese character tokens, with a lexicon of 1.5 million word types (Tsou *et al.* 2011).

In addition to monolingual Chinese dictionaries, corpora, especially parallel corpora, have been used extensively in bilingual lexicography involving Chinese and a foreign language in China. Before corpora made inroads in bilingual dictionary making, the examples in such dictionaries were either invented introspectively by lexicographers (e.g. *New English-Chinese Dictionary*) or cited from famous literary works (e.g. *Chinese-English Dictionary of Idioms*) (cf. Li and Zhong 2011). The primary aim of a bilingual dictionary is to help the user to establish a link between a lexical unit in the source language and its equivalents in the target language. In a bilingual dictionary, the

headword is usually defined in the target language. In most cases, a definition is only partially equivalent to the headword because it is often an abstract generalisation of the typical meanings of the word instead of covering all of its meanings fully. For example, 妥當 is usually defined as 'appropriate, proper, suitable, sound' in a Chinese-English dictionary. However, the word can also mean 'safe; ready; wise; guaranteed' in attested language use, which is not covered in the dictionary definition. In this case, bilingual examples cited from parallel corpora can complement missing meanings, e.g. 他想把大衣放在最妥當的地方 'He wanted to put his coat in the safest possible position' (cf. Li and Zhong 2011).

One particular issue in making specialised bilingual dictionaries relates to the definition and translation of the domain-specific usage of ordinary words in specific domains. In business domain, for example, the concept of 表 'table' is conventionally expressed as 'statement' instead of 'table', as in 'financial statement' (財務報表), and 'statement of income and expenses' (財務收益與費用表). Issues such as this can be addressed readily with the help of parallel and comparable corpora of the languages involved. As Li (2006) notes, specialised corpora are ideal linguistic and knowledge resources; and corpus-based specialised dictionaries can ensure a systematic coverage of headwords of practical value, accurate definitions, and appropriate illustrative examples.

The brief review of the state of the art of corpus-based lexicography in Chinese context suggests that the advantages of using corpora in dictionary making, whether monolingual Chinese dictionaries or bilingual dictionaries, are self-evident. On the other hand, corpora are a double-edge sword. If used inappropriately, they simply mean labour lost; and worse still, they can lead to falsehood under a scientific and objective disguise (cf. Li 2008: 203). This warning echoes Sinclair's (2004b: 2) observation that 'A corpus is not a simple object, and it is just as easy to derive nonsensical conclusions from the evidence as insightful ones.'

IV Contrasting English and Chinese Corpus Based Lexicography

It can be seen from the reviews in the two sections above that corpus linguistics has had a profound influence and great impact on lexicography in both English and Chinese contexts. While corpus linguistics was virtually identical to ECL before the middle 1990s, and CCL has lagged behind ECL in some areas of research (e.g. descriptive grammars, see below for further discussion), the use of corpora in Chinese lexicography nevertheless dated back as early as in English lexicography, because of the importance of defining basic Chinese characters and words from a huge lexicon.

As the two most widely used languages in the world, there are a large and growing number of learners of English and Chinese as second or foreign languages. Therefore, a range of learner dictionaries have been published for the two languages. On the other

hand, corpus-based English lexicography appears to have been confined largely to monolingual English corpora. In contrast, CCL has helped to create both monolingual Chinese dictionaries as well as bilingual dictionaries of Chinese with a foreign language like English, French, German and Japanese. The current state of affairs in lexicography in ECL might be due to English monolingualism. The CCL experience with bilingual lexicography suggests that this undoubtedly presents a fruitful avenue of development to ECL.

One tool which is particularly useful in corpus-based lexicography is the Sketch Engine (SkE) developed by Adam Kilgarriff, which automatically provides a corpus-derived summary of a word's grammatical and collocational behaviour on a single page (Kilgarriff *et al.* 2004). This tool is now in regular use by major publishers for lexicography at Oxford University Press, Cambridge University Press, and Collins among others.[3] Sketch Engine was successfully ported to Chinese when Chinese Sketch Engine (CSE)[4] was constructed by loading the Chinese Gigaword corpus onto Sketch Engine to test its cross-linguistic robustness (Kilgarriff *et al.* 2006).

English and Chinese are two genetically distinct languages with a range of cross-linguistic differences – this leads in turn to different issues for lexicographers. For example, unlike English in which the one-to-one correspondence between orthographic and morpho-syntactic word tokens can be considered as a default with a few main exceptions,[5] a written text in Chinese contains a running string of characters without delimiting spaces. Hence, the first step in CCL is to segment the running text into legitimate word tokens, a process known as 'word segmentation' or 'tokenisation' which usually requires complex computer processing that involves lexicon matching and the use of a statistical model. Furthermore, as Chinese is not a morphologically inflectional language, there is a lack of strict correspondence between word classes and syntactic functions, which means that even a noun can be used as a predicate in the same way as a verb, and a verb can be used as a grammatical subject or object as if it were a noun. All of the properties such as these in Chinese make the application of CCL in lexicography more challenging than in ECL.

On the other hand, there are a number of common issues to be solved in corpus-based lexicography for both English and Chinese, which are largely related to the balance and representativeness of corpora used, and the accuracy of corpus annotation. The corpora used in dictionary making for English and Chinese are usually very large, for example, the Bank Of English, LIVAC for Chinese, and the Gigaword corpora in Sketch Engine for both languages. Nevertheless, such corpora are often unbalanced, with little no little spoken data - they rely heavily, rather, on newspaper or newswire text. Another aspect of the lack of representativeness of the corpora is an associated time lag of the corpus data behind the actual language change, though a monitor corpus that is constantly updated is better able to address such an issue than a sample corpus. While part-of-speech annotation is generally reliable for English and Chinese (e.g. Lancaster's

CLAWS for English, ICTCLAS developed by the Chinese Academia for Simplified Chinese, and the part-of Speech tagging system developed by the Academia Sinica for Traditional Chinese), the type of corpus annotation that is particularly relevant in lexicography, semantic annotation or word sense disambiguation, still requires an improvement in accuracy so that different senses of the same words in different usage contexts, including their word senses that belong to the same word class, can be identified and counted reliably. This is important for the more sensible ordering of different word senses under the same headword entry.

Having explored the state of the art of corpus-based lexicography in ECL and CCL, we will now shift our focus, in the next section, to the use of corpora in descriptive grammars.

V English Corpus Linguistics and Descriptive Grammars

If John Sinclair can be readily identified as a pioneer in corpus-based lexicography, Randolph Quirk is the most readily identifiable pioneer in the development of corpus-based grammars. While Quirk's work at University College London (UCL) was more broadly focused than upon grammar alone, it is his work in redefining descriptive grammars that is our principal concern here. Similarly to Sinclair's innovations, Quirk was able to innovate because he constructed a corpus, the Survey of English Usage (SEU), that allowed him to do so. Quirk's work on the Survey of English Usage (SEU), started in 1959, was the first attempt to provide an on-going collection of present-day English that would, over time, facilitate the diachronic study of British English. It was a precursor of later corpora such as the British National Corpus and the American National Corpus (Ide and Reppen 2004), as it sought to balance its approach to the English language, recording both written and spoken English and sampling them in a range of genres and contexts. The SEU was very much a pioneering venture in corpus linguistics. Initially the corpus was not stored on a computer at all. It was stored on file cards and only later converted into a computerised form, the spoken part of which is available as the London-Lund Corpus (Svartvik 1990). When computerised, the corpus contained one million words of grammatically-analysed modern British English. Given that the material in the corpus was collected over a thirty year time period (1955-1985), the corpus can reasonably be described as an early attempt to provide a resource for diachronic study of contemporary British English. Indeed, the corpus has been re-edited, and an 800,000-word dataset specifically tailored to facilitate the diachronic study of spoken English, the Diachronic Corpus of Present-Day Spoken English (DCPSE), combines and contrasts spoken material from the SEU and from the British component of the International Corpus of English (ICE-GB).

Yet Quirk also developed a second innovation that was not pursued by Sinclair and this proved to be of great assistance to the creation of a corpus informed grammar –

from its earliest days, one of the features that made the UCL contribution to ECL distinctive was its engagement with the parsing of corpus data. The materials in the SEU were manually analysed by grammarians, who introduced simple grammatical analyses into the corpus data. It was the salience of grammatical analyses in the UCL approach to ECL that enabled one of the major contributions by the UCL team to the development of ECL – the use of corpus data in grammar production. Arising from the work on the SEU, the *Grammar of Contemporary English* (Quirk *et al.* 1972) and the *Comprehensive Grammar of the English Language* (Quirk *et al.* 1985) were published. The 1985 grammar was also the first widely distributed modern corpus-informed grammar, making its publication something of a milestone in the development of corpus linguistics. While some earlier grammars had drawn on corpus evidence, notably that of Fries (1940), Quirk *et al.* (1985) set the standard for the grammars that followed and in that sense generated a paradigm shift. As Mukherjee (2006: 337) notes:

"[...] the *Comprehensive Grammar of the English Language* [...] has been widely acknowledged to be the authority on present-day English grammar, bringing together descriptive principles and methods from various traditions and schools in order to cover grammatical phenomena as comprehensively as possible [...]"

That is not to say that the *Comprehensive Grammar* remains unparalleled – the benchmark for corpus-based grammars has now been substantially raised (see Mukherjee 2006 for a discussion and comparison of corpus-based grammars). Indeed the *Comprehensive Grammar* contains features which would not be acceptable in modern corpus-informed grammars. For instance, many of the examples given appear to be, at best, modified corpus examples; statistical data is provided only sporadically, and generally focuses on very rare or very common features. Nonetheless, by comparison to what went before the *Comprehensive Grammar* provided a new model for a corpus-based grammar, a model which in a refined form continues to dominate the field of English reference grammars.

Curiously, however, the grammars produced in the 1970s and 80s by UCL all had the written language as their principal focus, in spite of the fact that i) a great deal of effort was invested in the production of spoken corpus material by the Survey of English Usage and ii) some pioneering early work, notably that of Fries (1940) as noted, had used small collections of transcribed speech in order to explore grammar in spoken English. The grammars of this period were very much rooted in the attitude to speech that casts it as a debased form of language, mired in hesitations, slips of the tongue and interruptions. Yet as spoken language corpora such as the spoken section of the British National Corpus and the Santa Barbara Corpus of Spoken American English,[6] became available a spectrum of opinion regarding grammar in speech developed quite rapidly. One extreme may be characterised as the orthodox position – that grammar in speech is

present only in some bastardised form, subject to interference from a host of irrelevant performance features. The advent of spoken corpora allowed an opposite extreme to develop – the view not only that speech is grammatical, but also that it has a grammar of its own which is quite distinct from writing. This latter position is most closely associated with Brazil (1995). Brazil argues for a linear grammar of speech, a grammar which is not sentence-oriented and which does not have 'recourse to any notion of constituency of the hierarchically organized kind' (Brazil 1995: 4) – to put it simplistically, this kind of grammar involves no tree-style parsing.

Linguists such as Ron Carter and Mike McCarthy began to move this debate to the middle of the spectrum. In their work, Carter and McCarthy (1997) used the five million word CANCODE corpus of spoken English developed for Cambridge University Press[7] to explore the nature of grammar in speech. While initially seeming to adopt a position similar to that of Brazil (Carter and McCarthy 1995), Carter and McCarthy later took an approach that did not call for a distinct grammar of speech. Instead, their approach focused on those features of speech which appear most at odds with grammars taking the written language as their starting point. By shifting the focus from arguments in favour of the uniqueness of spoken grammar towards the distinctive features of spoken grammar, McCarthy and Carter developed a useful characterisation of spoken English, given full expression in their recent *Cambridge Grammar of English* (Carter and McCarthy 2006). McCarthy and Carter have drawn particular attention to how the grammar of speech can vary by context and can be influenced by the relationship that exists between the speaker and hearer (McCarthy 1998). This observation informed much of the design and development of the CANCODE corpus (Carter 2004) and other corpora developed at Nottingham, notably the Nottingham Multi-Modal Corpus (Carter and Adolphs 2008, Knight *et al.* 2009). This latter corpus has grown directly out of Nottingham's focus on speech, which has led the Nottingham team to look at how speech, gesture and prosody combine to create meaning, following on in particular from the work reported in Schmitt (2004).

The position developed at Nottingham in turn links clearly to the development of later grammars produced in the UCL tradition: if the work of linguists like Brazil, and more importantly McCarthy and Carter, can be viewed as having a distinct influence, it has been on grammars of English. While large grammars of English in the 1970s and 1980s rarely engaged with spoken language, grammars produced since then have done so routinely, and moreover have done so in a way that closely resembles what McCarthy and Carter called for – an acknowledgement of the differences between grammar in speech and grammar in writing. The *Longman Grammar of Spoken and Written English* (LGSWE, Biber *et al.* 1999) is a perfect case in point. This grammar might reasonably be viewed as part of the UCL tradition discussed earlier in this section, yet it engages fully with the differences between speech and writing. That said, the LGSWE moves more firmly to the middle ground than Carter and McCarthy did, by arguing that

these grammatical differences are largely a matter of degree rather than absolute distinctions. This point is developed by Leech (1998: 11), who explains how he decided to approach the question when he worked on the chapter in the LGWSE dealing with spoken English:

"The first task I set myself was to read through the drafts of all of the preceding chapters, noting grammatical phenomena which were strongly biased in frequency towards the spoken medium. The result of this was a rich profile of conversational grammar as it distinguishes itself from written grammar in all its variety. The profile included some features, such as disjunctive prefaces and tags ... which had found their way into the 'mainstream' presentation of Chapters 2-4. But in frequency terms, I noted a scale of conversational features going from those which are well represented also in the written medium to those which are virtually absent from it – such as dysfluency phenomena, which in written language are restricted to writing modelled on speech, as in fictional dialogue."

There were on the other hand features which appeared not to find a place in conversation – such as *for* as a conjunction although even here there was the occasional exception. It never seemed realistic, on reflection, to argue that certain features would *never* occur in speech, or would *never* occur in writing, because even if they were not detectable in several million words of conversation or written language (as the case might be), they could crop up if more data were considered. I therefore found myself adopting the 'same grammar' point of view, seeing both speech and writing as making use of the same overall grammatical repertoire, but allowing always for cases where the feature in question might be overwhelmingly commoner in one than the other.

In sum, the work of researchers pushing forward and redefining what a grammar is in the light of developments in corpus linguistics has led not simply to better grammars of English. It has also led to the notion of what a grammar is to be challenged and to a deeper awareness of the very real differences between the grammar of speech and writing. There has been a shift from a polarised debate about whether or not speech has a distinct grammar, towards the view now predominant in most schools of English descriptive grammar, where the grammatical system is both flexible and dynamic – where, as Leech (1998: 13) argues, 'English grammar is common to both written and spoken language – but its shape can be moulded to the constraints and freedoms of each.'

VI Chinese Corpus Linguistics and Descriptive Grammars

If, as discussed earlier, corpus-based lexicography for Chinese can be said to have developed at a comparable rate as for English, it is fair to say that the development of

CCL in descriptive grammars is still lagging far behind ECL. There is presently no corpus-based comprehensive reference grammar of the Chinese language.[8] Research in corpus-based descriptive grammars in Chinese is rather sporadic and fragmentary. What there is has focused on specific linguistic features of interest to individual researchers.

For example, Huang and Ahrens (2003) study the relationship between nouns and nominal classifiers in Mandarin Chinese on the basis of the data from *Mandarin Chinese Classifier Dictionary* (Huang *et al.* 1995), which is in turn based on the data from the Academia Sinica Balanced Corpus of Modern Mandarin Chinese, a grammatically tagged and balanced corpus containing a total of five million words of Mandarin texts as used in Taiwan. The authors challenge the traditional view that nominal classifiers classify individuals, arguing that 'classifiers do not simply agree with a noun, but instead coerce a particular meaning from it' (2003: 353).

Zhao (2010) studies the grammatical meanings and usage contexts of one function word in Chinese, 來著 *laizhe*, on the basis of 610 valid instances of the word in a sizeable online corpus of modern Chinese developed by the Centre for Chinese Linguistics of Peking University (hence the PKU Corpus). It is found that the word typically (99.8 % of the time) occurs at the end of a sentence or clause. The two main grammatical meanings the word expresses include indicating that something took place in the past (48.4%), and that the speaker seeks to be reminded of something that took place in the past (45.2%). While the first meaning is normally used in declaratives, the second meaning usually occurs in questions. The two types of sentences account for 69.4% and 29.3% respectively of total occurrences of *laizhe* in the corpus. It is also of interest to note that sentences that include this function word tend to avoid verb reduplication, verbs denoting termination, as well as perfective aspect markers *-le* and *-guo*. Another syntactic condition is that *laizhe* is predominantly used in affirmative (99.7%) rather than negative sentences.

Zhang (2010) is concerned with a syntactic and pragmatic analysis of a commonly used degree complement structure in Chinese, 'X 得很', on the basis of the PKU corpus. It is found that, semantically, this complement structure primarily expresses the speaker's psychological feeling, including evaluation of people and attitude toward a particular object or event. Syntactically, the words that can be used in the X slot of the structure include adjectives and verbs. Both qualifying and evaluating adjectives are found to occur in the structure, which can be positive, neutral or negative in evaluative meaning, while the verbs used in the structure are largely psychological verbs. The corpus analysis also suggests the words in the X slot of the structure are mostly disyllabic words, though monosyllabic words are not uncommon. Pragmatically, this complement structure can not only intensify the degree of the quality denoted by the word in the X slot to strengthen the expressive effect, but it can also be used as a parenthesis to express the speaker's attitude.

Siewierska, Xu and Xiao (2010) undertake a corpus-based study of splittable compounds in interaction with morphology, syntax, and pragmatics, with the aim to produce a systematic and realistic account of splittable compounds as attested in two million words of authentic spoken and written Chinese data as represented in the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao 2004) and the Lancaster Los Angeles Spoken Chinese Corpus (Xiao and Tao 2007). It is found that the typical grammatical pattern of splittable compounds is constitutive of an aspect marker (e.g. *-le, -zhe, -guo*) or resultative verb complements as post-verbal adjacent elements (54% of all instances), and a quantifier, a classifier, a modifier or a combination of two or more of them which precede the nominal components of splittable compounds. Drawing on morpho-syntactic and phonological criteria, the split uses, together with their combined uses, of splittable compounds with one inserted aspectual morpheme are viewed as words, while the others are regarded as phrases. From a discourse-pragmatic perspective, the split use of splittable compounds is more often found in the spoken genres of Chinese. Insertions of splittable compounds tend to function as mitigation or modification to the verbal heads or final nominal/complement elements.

Similarly, Wang and Wang (2009) approach splittable compounds from a pedagogical perspective by discussing the implication of their research findings based on the PKU corpus for teaching Chinese as a foreign language. Their investigation of 207 splittable compounds in the *Syllabus of Graded Words and Characters for Chinese Proficiency* indicates that for most of these splittable compounds, their continuous use is regular whereas their discontinuous use is idiosyncratic in attested language. The discontinuous use of splittable compounds is a colloquial feature which tends to occur in informal spoken genres. The authors suggest that the teaching of splittable compounds should be informed by the core patterns observed in corpus analysis by focusing on the most commonly used splittable compounds and their major forms of insertions in discontinuous use individually on the basis of their frequency bands while also taking account of genre variation and key sentence types so that learners can use splittable compounds appropriately in context.

Xiao, McEnery and Qian (2006) provide a systematic account of passive constructions in Chinese in contrast with English, covering a range of characteristics of passives in the two languages including various passive forms, long vs. short passives, semantic, pragmatic and syntactic features as well as genre variations. It is found that while passive constructions in both languages express a basic passive meaning, they also show a range of differences in terms of overall frequencies of use, syntactic features and functions, semantic properties, and distributions across genres. For example, passive constructions are nearly ten times as frequent in English as in Chinese because Chinese passives are much more restricted in use than their English counterparts. Short passives are predominant in English whereas Chinese displays a preference for long passives. English passives are more stylistically oriented to make the discourse sound

more impersonal, objective, formal and technical, while Chinese passives are a pragmatic voice that typically expresses a negative semantic prosody. Chinese passives in the predicate position typically interact with aspect; passive constructions with bare verbs in this position are uncommon, though they are frequent in other sentential positions. In contrast, the interaction between passives and aspect is not as apparent in English as in Chinese because all English sentences and clauses are formally marked by combined tense-aspect markers. Passives in English occur more frequently in informative than imaginative genres; reports/official documents and academic prose, in particular, show very high proportions of passives. But these two genres have the lowest proportions of passives in Chinese, where mystery/detective fiction and religious writing show exceptionally high proportions of passives. These differences are closely associated with the origins and functions of passive constructions in the two languages. The passive is primarily used as a style marker in English whereas it is typically an 'inflictive voice' in Chinese.

Xiao and McEnery (2008) explore negation in Chinese on the basis of spoken and written corpora of Mandarin Chinese. They found that in addition to negating functions, negative adverbs in Chinese also display differences in distribution across genres. The best characterisation of *mei* is that it negates the realisation of a situation, which distinguishes it from the general purpose negator *bu*. Regarding the interaction between negation and aspect marking, the actual aspect marked by *-le* and the experiential aspect marked by *-guo* are negated by mei while the durative aspect marked by *-zhe* and the progressive aspect marked by *zai* rarely occur in negative sentences. Imperfective aspects of the latter two types usually undergo a viewpoint aspect shift when they are negated. While sentences taking *zai* can be negated by either *bu* or *mei*, co-occurrences of *bu* and *zai* are typically in double negation structures or rhetorical questions, which are essentially positive in meaning. Sentences taking *-zhe* are negated more frequently by *mei* than *bu*, and *-zhe* is often omitted in negative sentences unless it appears in the V-*zhe* V structure where V-*zhe* acts as an adverbial. The scope of negation typically extends from the word immediately following the negator to the end of a clause unless the context provides clues that suggest otherwise. Word order is important in determining the focus of negation, which typically falls on some modifying element that usually follows the negator immediately, or on the end of a clause, unless a contrast present in context suggests otherwise. Transferred negation is uncommon in Chinese. When it occurs, the transferred focus of negation suggests a reduced degree of negation. Double negation is common in Chinese, but the negation of negation often means more than two negators cancelling each other, not only in terms of emphatic force, but in meaning as well. Finally, redundant negation typically occurs in sentences with some element which is inherently negative semantically.

Xiao and McEnery (2004) provide the first book-length corpus-based comprehensive account of aspect in Mandarin Chinese, encompassing both situation aspect and

viewpoint aspect. The corpus-based model of aspect developed in this book represents a significant advance in aspect theory, which explores aspect at both the semantic and grammatical levels. The two levels correspond to the two components of aspect, namely, situation aspect and viewpoint aspect. The former is language independent while the latter is language specific. While the two-level approach to modelling situation aspect taken by them gives a better account of the compositional nature of situation aspect by proposing a set of rules mapping verb classes at the lexical level onto situation types at the sentential level, it also provides a more refined classification of situation aspect, most notably by distinguishing between two types of states (i.e. individual-level state and stage-level stage). Their work is a systematic and structured exploration of the linguistic devices which Chinese employs to express aspectual meanings. In addition to situation aspect, which is inherent in linguistic expressions of situations in human languages, the book identifies, on the basis of corpus data, four perfective and four imperfective viewpoints in Chinese. While some of these viewpoints have already been identified in previous studies of aspect in Chinese, this corpus-based study has corrected many intuition-based misconceptions and associated misleading conclusions readily found in the literature. Some viewpoints, e.g. the completive aspect marked by resultative verb complements (RVCs), were considered for the first time as independent viewpoints based on their behaviours in attested language use.

Apart from the corpus studies of specific linguistic features in Chinese that have been reviewed so far, there is hardly any descriptive grammar of Chinese based on or informed by corpora. Xiao and McEnery (2010) might be an exception, which provides the first book-length corpus-based contrastive studies of major aspect-related grammatical categories in Chinese and English, i.e. grammatical categories such as aspect markers, completive and durative temporal adverbials, quantifiers, passives, and negation, which all contribute to aspectual meaning by interacting with situation aspect or viewpoint aspect in one way or another. However, this book is a research monograph more than a reference grammar of Chinese for general or pedagogical use.

In addition to linguistically oriented corpus research of Chinese grammar as exemplified above, there are a number of knowledge bases or electronic dictionaries of Chinese grammar that have been developed for use in automatic Chinese information processing. A key player in this area of research is the Institute of Computational Linguistics at Peking University. The team of computational linguists led by Professor Yu Shiwen have developed numerous standards and resources that have profoundly influenced natural language processing of the Chinese language including for example, *Specifications for Basic Processing of Contemporary Chinese Corpus at Peking University* (Yu and Duan 2002), *Grammatical Knowledge Base of Contemporary Chinese* (Yu 2003), *Grammatical Knowledge Base of Chinese High Frequency Words* (Zhu *et al.* 2004), *Chinese Function Word Usage Knowledge Base* (Liu *et al.* 2005), and *Modern Chinese New Words Information Electronic Dictionary* (Kang 2002). While

these knowledge bases can provide useful grammatical information about the Chinese language in natural language processing, they are nevertheless not descriptive grammars in a sense that a reference grammar is expected to be.

VII Contrasting English and Chinese Corpus Based Descriptive Grammars

The discussions of corpus-based descriptive grammars in English and Chinese in the two earlier sections reveal a huge gap in research in this area between the two languages. On the one hand, from *A Grammar of Contemporary English* (Quirk *et al.* 1972) to *A Comprehensive Grammar of the English Language* (Quirk *et al.* 1985) and further to the *Longman Grammar of Spoken and Written English* (Biber *et al.* 1999), corpus-based descriptive grammars of English have developed through three generations so that the third generation English reference grammar (Biber *et al.* 1999) is not only systematically and substantially corpus-based but also takes account of register variations, explores the differences between spoken and written grammars, and includes lexical information as an integral part of grammatical descriptions. On the other hand, corpus-based studies of descriptive grammar in Chinese are rather fragmentary and sporadic, with the first corpus-based Chinese reference grammar yet to be published.

The sharp contrast in the development of corpus-based descriptive grammars in the two languages may be due to the fact that the development of corpus linguistics started with the English language, which has caused corpus-based research of Chinese grammar to lag behind, perhaps, in some areas. But a more important reason for the significant lagging behind for corpus-based Chinese descriptive grammar, in our view, is the separation between Chinese corpus research and linguistic research in Chinese context. In mainland China, for example, members of the Corpus Linguistics Society of China (CLSC) are almost exclusively university professors and lecturers working in schools and department of foreign studies, who are more interested in ECL than CCL. On the other hand, those working with Chinese corpora are usually computer specialists and computational linguists who are more interested in natural language processing technologies than linguistic theorisation. We believe that cooperation and collaboration between these two groups of scholars, and therefore between their corpus building and analysis expertise and linguistic knowledge, would substantially facilitate the healthy development of Chinese corpus linguistics in the right direction. This is also an important area where CCL can learn from ECL.

In the sections that follow, we will consider an important area of corpus linguistics that is of particular relevance to language education, namely learner corpora and interlanguage research.

VIII English Corpus Linguistics and Learner Corpora

Learner corpus research on the English language has burgeoned in the past twenty years. While work based on what might very broadly be termed learner corpus data was undertaken in the 1970s, for example by Dulay and Burt (1973) and Krashen *et al.* (1978), a major contribution to the development of ECL was made when Sylviane Granger inaugurated the International Corpus of Learner English (ICLE) in 1990 (see Granger 1993a, 1993b, 1994). This is a consortial effort led by Granger dedicated to making comparable corpora composed of the English writings of L2 learners of English with a specific L1 background.[9] In order to make the corpora comparable, the students generally produce a series of essays of roughly similar length on similar topics. These topics require the students either to write argumentative essays in response to questions such as 'Is Fox Hunting Justifiable?' or to produce essays as part of a literature exam. To date ICLE contains over 4.5 million words arranged in 16 subcorpora, each of which contains the writing of students from a distinct L1 background (Granger *et al.* 2009). As well as collecting the text, the corpus also captures some background information on the contributors to the corpus, recording information such as their sex, educational level and their educational experience. By comparison to work from the 1970s, the scope and systematic nature of work with learner corpora allows for a much more wide-ranging and systematic exploration of learner data than the earlier studies could. Crucially, the studies can benefit from the frequency data that corpora are ideally suited to provide.

While early work on learner corpora focused on written language, later work by De Cock (1998) expanded the study of learner English from written to spoken English with the Louvain International Database of Spoken English Interlanguage (LINDSEI). This corpus is more limited than ICLE, as the first release of the corpus covers only 50 speakers with a French L1 background, and is only 100,000 words in size, though a number of teams are working internationally to expand the corpus.[10] The production of LINDSEI was a pioneering attempt to facilitate the study of learner speech, which has been further enabled by the production of a native speaker corpus, the Louvain Corpus of Native English Conversation (LOCNEC), which has been used a control corpus for exploring LINDSEI (see, for example, Aijmer 2009).

In addition to the production of learner corpora produced by L2 learners, Granger's team also had to engage with the issue of how to compare the material gathered to L1 English material. In response to this, a further corpus was collected, LOCNESS (Louvain Corpus of Native English Essays). This consists of argumentative essays by British and American writers and is some 324,000 words in size. The corpus is divided into three subparts – British L1 English writers in pre-university study (17-18 years old), British L1 English writers at university, and American L1 English writers at university.

There is little doubt that work on learner corpora has stimulated a whole new field of ECL. It has generated a great deal of academic output (see Granger *et al.* 2002 and Gilquin *et al.* 2008, for example) and has proven highly influential in the world of English language teaching (ELT) publishing, as well as in studies of second language acquisition (Ellis 2008). The publishers Longman and Cambridge University Press have both developed 'in house' learner corpora that they now regularly use to inform their publications aimed at the ELT market. For example, the Longman Learner Corpus is ten million words in size and, like ICLE, is composed of a series of subcorpora of data from students with specific L1 backgrounds producing L2 English writing. Longman use the corpus to create language learning materials which are designed to counteract the errors made by students with specific L1 backgrounds.

The Cambridge Corpus is some thirty million words in size and is continually growing. It is of particular interest in that annotated within it are the locations and types of the errors made by the learners when writing. The advantages of the inclusion of such annotations are obvious:[11]

"We can see which errors are typical of different learner levels or of particular language groups because all the scripts have information about the first language and English level of the writer. This means that when we produce a book designed for a particular level, e.g. Upper Intermediate, we can look at all the scripts written by Upper Intermediate learners and very easily see exactly what mistakes they make. In this way we can make sure the book contains appropriate help for an Upper Intermediate student."

Error tagging was another development in learner corpus research strongly advocated by Granger (1999, 2003). Yet while Granger has been clear on the potential benefits of error tagging, noting that 'once the corpus is error-tagged, the return on investment is huge' (Granger 2003: 10), she has also sounded a useful note of caution about the process of error tagging. Having noted that it is one of the more subjective forms of corpus annotation, Granger (2003: 11) adds:

"It is also important to bear in mind that error tagging, in spite of its numerous advantages, is only concerned with learner misuse. It fails to uncover other aspects of interlanguage such as the under- and overuse of words and phrases, which together with downright errors contribute to the non-nativeness of learner productions."

In short, while useful, error tagging is not a panacea for the problem of non-nativeness in L2 language use. It is one type of information amongst many that a learner corpus might provide which can be of assistance to the learner or teacher of English.

The expansion of ECL to include learner English was very much pioneered by

Granger; much of the work of commercial publishers such as Longman and CUP using their own corpora is influenced, either directly or indirectly, by her work, as is the research of other scholars using corpora outside of the ICLE project, such as Tono (2009). However, the learner corpus approach to using corpora in language teaching is but one of many ways in which corpora have impacted upon research in ELT. Corpora have penetrated many aspects of ELT, and have contributed to more specialised areas such as English for Academic Purposes (e.g. Alsop and Nesi 2009) and English for Specific Purposes (e.g. Mohamad-Ali 2007). Corpus-based research in ELT in particular, and language teaching in general, has focused upon issues such as syllabus design (Mindt 1996, Shortall 2007), language testing (Alderson 1996, Taylor and Barker 2008), classroom teaching practice (Amador-Moreno *et al.* 2006), reference and classroom material production, and student-led learning through the so-called *data driven learning* approach (Johns 1994, 1997, Boulton 2009; see McEnery and Xiao 2010 for a comprehensive review of corpus-based language education). There is also a regular conferences series, *Teaching and Language Corpora* (TALC), whose remit is the use of corpora in language teaching. Possibly one of the most significant contributions of corpus linguistics to ELT has been corpora developed with the goal of producing advanced learners' dictionaries and frequency dictionaries, such as the COBUILD dictionary, the *Macmillan English Dictionary for Advanced Learners*, and the *Routledge Frequency Dictionaries* series which define a core vocabulary for learners of a range of popular languages. The corpora developed by publishers for these dictionaries have also been employed for other purposes, notably ELT materials such as the *Longman Language Activator* series.

So while Granger's contribution is undoubtedly important, the full impact of corpora upon language learning is now greater than Granger's body of work and encompasses a great mass of diverse research in the general area of ELT, as reported in publications such as Wichmann *et al.* (1997), Burnard and McEnery (2000), Kettemann (2002) and Reppen (2009).

IX Chinese Corpus Linguistics and Learner Corpora

While corpus-based research of native Chinese, especially descriptive Chinese grammar, is under-developed in relation to ECL as noted earlier, interlanguage analysis of learner errors appears to be the focus of corpus-based Chinese teaching and learning at present. With the rapid development of Teaching Chinese as Foreign Language (TCFL) since the mid 1990s, there has been an increasingly pressing demand for Chinese interlanguage corpora to aid Chinese teaching and learning (cf. Ren 2010). As a result, a number of learner corpora of Chinese have been created or planned over the past decade because for teachers of Chinese as a foreign language, learner corpora of Chinese can not only provide more direct and readily available help in teaching and

learning, for example, in terms of real-time error analysis, computer-aided teaching, pertinent exercises for individual learners, learning evaluation, but they also play an increasingly important role in syllabus design, materials development, lexicography and so on (cf. Luo 2008: 48).

The earliest corpus of learner Chinese is probably the Chinese Interlanguage Corpus created at Beijing Language and Culture University in 1993-1995 (Chu *et al.* 1995). The corpus is composed of 1,371 compositions by 740 students, amounting to 1.04 million Chinese characters. It is encoded with 23 metadata features including, for example, the learner's gender, age, nationality, education level as well as task type, topic, length, and writing time, and annotated with part-of-speech information and learner errors at character, word and sentence level. The corpus also comes with an integrated corpus exploration tool.

Another corpus of similar nature developed by Beijing Language and Culture University is the HSK Dynamic Composition Corpus, which contains over 4.24 million Chinese characters of learner data in the form of 11,569 compositions written by learners of Chinese as a second or foreign language when they took the HSK Chinese language proficiency tests in 1992-2005. This corpus is encoded with rich metadata and tagged with learner errors at character, word and sentence level. The corpus is publicly available online.[12] After signing up for a free user account, registered users can search for sentences with a specific string of characters or an error type, or access various statistics as per the encoded metadata. As the compositions included in the corpus are continuous data over time, it is also suitable for longitudinal study.

A number of other learner Chinese corpora have been reported in China. These, however, do not appear to be publicly available, including, for example, the three-million-character Chinese interlanguage corpus created by the College of Chinese Language and Culture of Jinan University (CCLC), the 900,000-character error-tagged corpus of compositions and exercises developed at Nanjing Normal University, and the 750,000-character Chinese interlanguage corpus developed at Zhongshan University. In addition, a country-specific (L1 Korean) Chinese interlanguage corpus, which is funded by China's National Social Science Foundation, is in progress at Ludong University, with a target size of over three million characters; another research project is funded by China's Ministry of Education and undertaken by Shanghai Jiaotong University to develop a corpus of Chinese compositions by foreign learners.

There are also a few learner Chinese corpora that have been reported outside mainland China. For example, the Modern Interlanguage Chinese Corpus is corpus searchable online,[13] which comprises tasks, such as writing compositions or specific sentences, collected twice per semester from years 2-4 Chinese studies students at six Korean universities between 2004 and 2006, totalling 10,135 sentences (Piao 2010). The Mandarin Interlanguage Corpus (MIC) has recently been completed at the University of Hong Kong. A total of 19 participants from two groups of second year students taking a

two-year Certificate Course in Chinese Language course were recruited to participate in the 1.5-year corpus project, whose proficiency level in Chinese was supposed to reach the intermediate level at the HSK proficiency test upon the completion of the course. Their first languages include English (5), Korean (3), Japanese (3), German (2), French, Tamil, Indonesian, Spanish, Dutch, and Thai. Both written and spoken data were collected during the course at the end-of-course exam. The written data is in the form of short compositions ranging from 150 to 700 characters, depending on the genre type, while the spoken data is from their 1-2 minute short presentations during in-class conversation and the examination. A total of 88 compositions (amounting to 50,000 characters) and 120 hours of recordings (60 hours of which were observed to be coherent and therefore usable data) have so far been processed from course-work and examination. The corpus also comes with a user-friendly online interface that allows a number of search options including searching by source, word category, first language and topic of the task (Tsang and Yueng 2010). National Taiwan Normal University created a small interlanguage corpus in 2004-2005, comprising 41,053 sentences by 210 learners of Chinese (mostly L1 English), which is available online for download.[14]More recently, this university has also planned a more ambitious project to develop a learner interlanguage corpus (LIC) for written and spoken texts produced by learners of Chinese as a second language at the university.[15]

Apart from the development of corpora of Chinese interlanguage, a range of corpus-based studies that analyse specific grammatical features of learners' Chinese interlanguage have been published. For example, Yang's (2004) study is based on the Chinese Interlanguage Corpus developed by Beijing Language and Culture University, which compares Japanese learners' acquisition of Chinese directional complements at different stages and analyses the factors affecting their acquisition. He finds that over-generalisation of object is the main cause of the errors while over-generalisation is in turn a result of intra-lingual or inter-lingual transfer.

Yuan (2005a, 2005b) respectively provide i) a detailed analysis of about 100 sentences from a Chinese interlanguage corpus that contain errors with the negator 不 *bu* and ii) 100 sentences containing errors with the negator 沒有 *meiyou*, classified them into various semantic and syntactic types. Explanations of the errors and corrections are also suggested in his papers. Wang (2005) investigates foreign learners' acquisition of the Chinese 比 *bi* comparative structure, finding that their errors mainly relate to i) unnecessary addition; ii) mismatch of adjectives and nouns; iii) misuse of the *bi* structure and iv) sentence order inversion. The author argues that the method for correcting errors and improving language ability is to develop learners' language awareness through intensive practice.

Zheng (2006) investigates the actual use of degree adverbs in the Chinese Interlanguage Corpus and compares her results with the published literature. She finds that the types of errors identified in the literature are all attested in the corpus data,

but some error types in the literature which have been proposed on the basis of researchers' introspection are not common in the corpus; there are also some new error types identified in the corpus which have not been predicted and accounted for in the literature.

Hua (2009) analyses the types of errors made with the Chinese preposition 给 *gei* in Korean students' interlanguage. Hua explores the reasons for the misuse from the perspective of L1 interference, and proposes remedial teaching strategies to deal with the errors. Shen (2009) undertakes a corpus-based study of international students' errors with the double object structure in their Chinese interlanguage, finding that the most common types of errors with the structure relate to the omission of the indirect object or using a preposition to move the indirect object before the verb.

Zhou and Hong (2010) investigate the use of figure of speech in the Chinese interlanguage of foreign learners at intermediate and advanced levels. Their results show that foreign learners use fewer types of figures of speech than native Chinese speakers, and they use figures of speech less frequently in their interlanguage; however, the error rate is not low. The errors are due to learners' insufficient understanding of grammar rules, cultural diversity and improper creative language use. It is found that the development of figures of speech in learners' Chinese interlanguage is positively correlated with the level of Chinese language proficiency.

Zhang (2011) investigates the current situation of and the problems with research on the acquisition of Chinese sentence structures in teaching Chinese as a foreign language. It was found that the central problem, among others, was that that the research scope is too narrow and the empirical studies are insufficient. Consequently, we have no clear idea of the real situation of acquisition of Chinese sentence structures by foreigners, and the teaching strategies proposed in the previous studies also lack practical values.

In addition to grammatical features, there are also numerous studies of lexical features in Chinese interlanguage. For example, Xing (2003) provides an exhaustive analysis of the 520 erroneous compounds of 17 subcategories in five categories found in the Chinese Interlanguage Corpus created by Beijing Language and Culture University. Error analysis shows that foreign learners have strong awareness of the structures of compound words (e.g. words formed with morphemes, structures and semantic relevance), and they may employ two different ways of acquiring compound words in Chinese, namely, compounding from morphemes and recognising compound words as a whole. It is believed that the first of these methods is dominant.

Shi (2008) provides a systematic and comprehensive analysis of the errors with Chinese learners' use of Chinese style idioms called *chengyu* on the basis of the Chinese Interlanguage Corpus, explores the reasons for the errors from the perspective of cognitive linguistics, and suggests some strategies for teaching Chinese *chengyu* to learners at advanced proficiency level.

Zhang (2008) is concerned with confusable words in Chinese interlanguage. It is argued that confusable words should be inspected separately according to learners' L1 backgrounds, and cross-linguistic contrast of Chinese with the learner's L1 could be useful in uncovering the causes of errors they make. In discriminating a group of confusable words, more attention should be paid to the most confusable part, and the meaning characters of the words should be associated with their collocation rules.

It can be seen even from this brief review of the development of Chinese learner corpora and various corpus-based studies of grammatical and lexical features of learner Chinese that Chinese interlanguage research has been a focus of TCFL research. The increasing interest in this research area is well evidenced by the newly launched popular international conferences series 'International Symposium on Chinese Interlanguage Corpus Development and Application' organised by Beijing Language and Culture University and Nanjing Normal University, with the first meeting held in Nanjing on 29-31 July 2010 and the second to take place in Beijing on 23-25 August 2012. On the other hand, the review above also reveals a number of issues with current Chinese interlanguage research (cf. Zhang 2010). For example, in comparison with learner English corpora, there are very few existing Chinese interlanguage corpora that are publicly available and can actually be used in teaching and research. In relation to many reference corpora, Chinese interlanguage corpora are typically rather small in size, with annotated corpora composed of only one million Chinese characters. Existing Chinese learner corpora also suffer from a lack of balance in terms of learners' first language backgrounds and the nature of the data included in the corpora. The existing Chinese interlanguage corpora are also seriously biased towards Asian learners such as Korean, Japanese and Southeast Asian learners while learners from Europe and America are seriously under-represented. The range of genres in the learner material is also limited - such corpora are almost exclusively composed of compositions completed by foreign learners under test conditions. Currently available corpora hardly contain any spoken data. While Yang *et al.* (2006) reported the development of the Chinese Learners' Spoken Corpus, this corpus does not appear to have actually been used in TCFL research (Zhang 2010). Existing Chinese interlanguage corpora also suffer from inaccurate and inconsistent annotation and limited public availability (Cui and Zhang 2011). In addition to these issues with corpus resources, corpus-based research of Chinese interlanguage has also been confined to error analysis while the actual usage patterns in learners' interlanguage, which could be of interest to second language acquisition research in their own right, have largely been overlooked.

X Contrasting English and Chinese Learner Corpora

Learner corpus development and analysis have recently been a focus of corpus linguistic research in China. This is true not only of Chinese interlanguage corpus

research, as reviewed in the previous section, but also of learner English corpus research because of the importance of English in Chinese society. A range of Chinese learner English corpora have been published in China, including for example, Chinese learner English Corpus (CLEC; Gui and Yang 2003), Chinese Learners' Spoken English Corpus (COLSEC; Yang and Wei 2005), Spoken and Written English Corpus of Chinese Learners (SWECCL; Wen *et al.* 2005, 2008), Parallel Corpus of Chinese EFL Learners (PACCEL; Wen and Wang 2008), and Corpus of English Majors (CEM; China High Education Foreign Language Majors Multilingual Corpus Research Project Team 2008).

It appears, then, that learner English corpus research has been undertaken extensively around the world (e.g. in Belgium, Japan, and China, Sweden and Poland, but ironically, not notably in a native English speaking country), whereas Chinese interlanguage research is highly concentrated in China. On the other hand, while learner English corpus research has covered the interlanguages of learners from an extensive range of first language backgrounds including languages both similar to and distinctly different from English, Chinese interlanguage research is unbalanced as regards learners' first languages, which are essentially limited to learners from East and Southeast Asian countries, with learners with a first language of English or an European language markedly under-represented.

As English is the most popular and most widely learned second or foreign language in the world, learner English has unsurprisingly been studied more extensively than Chinese interlanguage. As a result, CCL clearly has a lot to learn from the ECL experience with learner English corpora when seeking to address the numerous issues with Chinese interlanguage corpus research as noted in earlier section. Fortunately, there has been a proposal to create the International Corpus of Learner Chinese as a joint research project between a number of universities in and outside China. According to the proposal (see Cui and Zhang 2011), the final corpus will comprise 50 million Chinese characters, with an annotated written component of 20 million characters and a raw text written component of 25 million words, in addition to five million characters of spoken data, with an annotated component of two million characters and a raw text component of three million characters. Data will be collected from non-native Chinese learners including both Chinese majors and non-Chinese majors at beginner, intermediate, and advanced levels. The corpus will cover narrative, argumentative, and expository genres while the task types to be represented in the corpus will include homework, exam script, HSK test, as well as paragraphs in answer to questions. The corpus will be encoded with rich metadata about learners (e.g. nationality, age, L1; whether they are a Chinese descent) and about the text sample (e.g. topic, genre, grade etc.), and annotated with learners' errors at various levels (e.g. character, word, sentence and discourse levels) as well as basic tagging such as tokenisation, part-of-speech tagging, sentence constituent analysis, and identification of sentence type and sentence pattern. The resulting corpus is expected to be mounted at a

dedicated website to allow registered users to search online in addition to a CD edition to be published for use offline on standoff PCs.

In addition to this major endeavour, there is also a need to construct a native Chinese corpus comparable to the learner Chinese corpus, which mirrors the relationship of ICLE/LINDSEI with LOCNESS/ LOCNEC, to facilitate comparisons of learner Chinese with native Chinese. On the other hand, since international students who study in China are mostly from the neighbouring East and Southeast Asian countries while Chinese is often taught locally in major European and American countries such as the UK and the US where ECL has also developed most rapidly, corpus linguistics in these areas can also contribute to the research of Chinese interlanguage by creating corpora of learner Chinese produced by their local native students to complement the existing interlanguage Chinese corpora created in China, which will facilitate contrastive analysis of interlanguages by learners from Asia and those from Europe and America.

XI Conclusion

In this paper, we have taken a historical approach to the development of corpus linguistics in English and Chinese contexts in contrast, by highlighting the key points and unique challenges in the development for each, and identifying possible fruitful avenues of development where ECL and CCL can inform and learn from each other. It is clearly impossible for a paper of this size to cover all aspects of corpus linguistics in relation to both languages. Hence, we have chosen to focus on the major areas of linguistic research that have all been deeply influenced by the development of corpus methodology, namely lexicography, descriptive grammars, and interlanguage analysis. We hope that the parallel survey of the development of corpus linguistics in English and Chinese contexts in the above three areas will contribute to the further development of both ECL and CCL in future.

Notes

[1] For readers interested in the development of English corpus linguistics, chapter 4 of McEnery and Hardie (2012), which informs this paper, has more detail on this.

[2] This corpus is 650 million words in size at time of writing.

[3] See the official website of the Sketch engine http://www.sketchengine.co.uk/.

[4] http://wordsketch.ling.sinica.edu.tw/

[5] These exceptions include, for example, multiwords (e.g. *so that* and *in spite of*), mergers (*can't* and *gonna*), and variably spelt compounds (e.g. *noticeboad, notice-board* versus *noticeboard*).

[6] See http://www.linguistics.ucsb.edu/research/sbcorpus.html.

[7] See http://www.cambridge.org/elt/corpus/cancode.htm.

[8] An exception to this is the forthcoming Cambridge Chinese Reference Grammar, a joint book project under contract with Cambridge University Press which is launched by the Research Centre on Chinese Linguistics of the Hong Kong Polytechnic University and Peking University. According to the project site (http://p2u2.cbs.polyu.edu.hk/enpublicview.asp?bid=3&sid=4&pid=3), this book aims to be 'a comprehensive and accessible reference grammar of Chinese' covering all important linguistic facts of the language, which 'are presented in a way that does not presuppose knowledge of a particular linguistic theory or grammar of Chinese. One of the key features of the volume is that it is data-driven and the examples used to illustrate the linguistic facts will be derived mostly from corpus data.'

[9] Of course, corpora of the academic writing of L1 speakers' writing in their L1 also exist – see, for example, Ebeling and Heuboeck (2007).

[10] See http://cecl.fltr.ucl.ac.be/Cecl-Projects/Lindsei/lindsei.htm.

[11] Quoted from http://www.cambridge.org/elt/corpus/learner_corpus2.htm.

[12] http://202.112.195.192:8060/hsk/login.asp.

[13] http://jit.jj.ac.kr:8080/corpus/index.jsp.

[14] http://chinese.mtc.ntnu.edu.tw/moodle/mod/forum/discuss.php?d=210.

[15] http://140.122.100.145/topntnu/cltc/e/5.html.

References

Aijmer, K. (Ed.) 2009. *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Alderson, J. C. 1996. 'Do corpora have a role in language assessment?', in J. Thomas & M. Short (Eds.) *Using Corpora in Language Research* (pp. 248–59). London: Longman.

Alsop, S. & Nesi, H. 2009. 'Issues in the development of the British Academic Written English (BAWE) corpus', *Corpora* 4 (1), 71–83.

Amador-Moreno, C. P., O'Riordan, S. & Chambers, A. 2006. 'Integrating a corpus of classroom discourse in language teacher education: the case of discourse markers', *Recall* 18 (1), 83–104.

Beijing Aeronautical University. 1985. *Xiandai Hanyu Yong Zi Pindu Biao (A Frequency Table of Character Usage in Modern Chinese)*. Beijing: Beijing Aeronautical University.

Beijing Language and Culture University. 1986. *Xiandai Hanyu Pinlü Cidian (A Frequency Dictionary of Mandarin Chinese)*. Beijing: Beijing Language and Culture University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Boulton, A. 2009. 'Testing the limits of data driven learning: language proficiency and training', *ReCall* 21 (1), 37–51.

Brazil, D. 1995. *A Grammar of Speech*. Oxford: Oxford University Press.

Burnard, L. & McEnery, T. 2000. *Rethinking Language Pedagogy from a Corpus Perspective*. Berlin: Peter Lang.

Carter, R. 2004. *Language and Creativity: The Art of Common Talk*. London: Routledge.

Carter, R. & Adolphs, S. 2008. 'Linking the verbal and visual: new directions for Corpus Linguistics', In A. Gerbig & O. Mason (Eds.) *Language, People, Numbers: Corpus Linguistics and Society* (pp. 275–91). Amsterdam: Rodopi.

Carter, R. & McCarthy, M. 1997. *Exploring Spoken English*. Cambridge: Cambridge University Press.

Carter, R. & McCarthy, M. 2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: Cambridge University Press.

Chen, H. 1922. 'Yutiwen yingyong zihui (The applied glossary of Modern Chinese)', *Xin Jiaoyu (New Education)* 5(5).

Chen, H. 1928. *Yutiwen Yingyong Zihui*. Beijing: The Commercial Press.

China High Education Foreign Language Majors Multilingual Corpus Research Project Team. 2008. *Yingyu Yuliaoku (Corpus of English Majors)*. Shanghai: Shanghai Foreign Language Education Press.

Chu, C., Chen, X., Zhang, W., Wei, P., Zhang, W. & Zhu, Q. 1995. *Hanyu Zhongjieyu Yuliaoku Xitong Yanzhi Baogao (Research Report of the Corpus of Chinese Interlanguage (CCI 1.0))*. Beijing: Beijing Language and Culture University.

Cui, L. 2011. 'Yuliaoku zai *Duiwai Hanyu Xin Ciyu Cidian* weiguan jiegou zhong de yunyong (Application of the corpus in microstructure of new words dictionary of CFL)', *Chongqing Ligong Daxue Xuebao (Journal of Chongqing University of Technology (Social Science)* 25 (10), 84-89.

Cui, X. & Zhang, B. 2011. '*Quanqiu Hanyu Xuexizhe Yuliaoku* jianshe fang'an (The principles for building the International Corpus of Learner Chinese)', *Yuyan Wenzi Yingyong (Applied Linguistics)*, 2011 (2), 100-108.

De Cock, S. 1998. 'A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English', *International Journal of Corpus Linguistics* 3 (1), 59–80.

Dulay, H. & Burt, M. 1973. 'Should we teach children syntax?', *Language Learning* 23, 95–123.

Ebeling, S. & Heuboeck, A. 2007. 'Encoding document information in a corpus of student writing: the British Academic Written English Corpus', *Corpora* 2 (2), 241-256.

Ellis, R. 2008. *The Study of Second Language Acquisition* (second edition). Oxford: Oxford University Press.

Firth, J. R. 1957. *Papers in Linguistics*. Oxford: Oxford University Press.

Francis, N. & Kučera, H. 1964. *Manual of Information to Accompany a Standard Corpus*

of Present-Day Edited American English for Use with Digital Computers. Providence, Rhode Island: Department of Linguistics, Brown University.

Fries, C. 1940. *American English Grammar.* New York: Appleton Century.

Gilquin G., Papp, S. & Díez-Bedmar, M. B. (Eds.) 2008. *Linking up Contrastive and Learner Corpus Research.* Amsterdam and Atlanta: Rodopi.

Granger, S. 1993a. 'The International Corpus of Learner English', In J. Aarts, P. de Haan & N. Oostdijk (Eds.) *English Language Corpora: Design, Analysis and Exploitation* (pp. 57–69). Amsterdam: Rodopi.

Granger, S. 1993b. 'The International Corpus of Learner English', *The European English Messenger* 2 (1), 34.

Granger, S. 1994. 'The learner corpus: a revolution in applied linguistics', *English Today* 39 (10/3), 25–29.

Granger, S. 1999. 'Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus', In H. Hasselgard & S. Oksefjell (Eds.) *Out of Corpora - Studies in Honour of Stig Johansson* (pp. 191-202). Amsterdam: Rodopi.

Granger, S. 2003. 'Error-tagged learner corpora and CALL: a promising synergy', *CALICO* (special issue on Error Analysis and Error Correction in Computer-Assisted Language Learning) 20 (3), 465–80.

Granger, S., Hung, J. & Petch-Tyson, S. (Eds.) 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam and Philadelphia: John Benjamins.

Granger, S., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English Version 2.* Louvain: Presses Universitaires de Louvain.

Gui, S. & Yang, H. 2003. *Zhongguo Xuexizhe Yingyu Yuliaoku (Chinese Learner English Corpus).* Shanghai: Shanghai Foreign Language Education Press.

Halliday, M. A. K. 1985. *Introduction to Functional Grammar.* London: Edward Arnold.

Hanks, P. 2009. 'The impact of corpora on dictionaries', In P. Baker (Ed.) *Contemporary Corpus Linguistics* (pp. 214–36). London: Continuum.

Hong Kong Polytechnic University. 1991-1997. *Zhongguo Dalu, Taiwan, Xianggang Hanyu Ciku (A Chinese Word Bank from Mainland China, Taiwan, and Hong Kong).* Hong Kong: Hong Kong Polytechnic University.

Hua, X. 2009. 'Hanguo liuxuesheng xide jieci *gei* de pianwu fenxi ji jiaoxue duice (The error analysis of Korean students' acquisition of the preposition *gei* and the corresponding teaching strategies)', *Jinan Daxue Huawen Xueyuan Xuebao (Journal of the College of Chinese Language and Culture of Jinan University)* 2009 (1), 24-29.

Huang, B. 2004. 'Yixiang huafen de yiju yu biaozhun (The basis and criteria for word sense division)'. *Cishu Yanjiu (Lexicographical Studies)* 2005 (5), 31-36.

Huang, C. R. & Ahrens, K. 2003. 'Individuals, kind and events: Classifier coercion of nouns', *Language Sciences* 25 (4), 353-373.

Huang, C. R., Chen, K. J. & Lai, C. X. 1995. *Mandarin Chinese Classifier and Noun-Classifier Collocation Dictionary*. Taipei: Mandarin Daily Press.

Hunston, S. & Francis, G. 1999. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Ide, N. & Reppen, R. 2004. 'The American National Corpus: overall goals and first release', *Journal of English Linguistics* 32 (2), 105–13.

Jiang, Z. 2005. 'Dianzi yuliaoku yu yuwen cishu de bianzuan xiuding (Electronic corpora and compilation and revision of Chinese dictionaries)', *Sichuan Daxue Xuebao (Journal of Sichuan University, Social Science Edition)* 140 (5), 79-86.

Johns, T. 1994. 'From printout to handout: grammar and vocabulary teaching in the context of Data-driven Learning', In T. Odlin (Ed.) *Perspectives on Pedagogical Grammar* (pp. 293–313). Cambridge: Cambridge University Press.

Johns, T. 1997. 'Contexts: the background, development and trialling of a concordance-based CALL program', In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.) *Teaching and Language Corpora* (pp. 100–15). London: Longman.

Kang, S. 2002. '*Xiandai Hanyu Xin Ciyu Zixun Dianzi Cidian* de yanjiu yu shixian (Development and study of the *Modern Chinese New Words Information Electronic Dictionary*)', *Computational Linguistics and Chinese Language Processing* 7 (2), 89-100.

Kang, S. 2003. '*Xin Ciyu Da Cidian* de bianzuan (Compilation of *A dictionary of New Words*)', *Cishu Yanjiu (Lexicographical Studies)* 2003 (2), 12-20.

Kettemann, B. & Mark, G. 2002. *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. 2004. 'The Sketch Engine', In *Proceedings of Euralex*, (pp. 105-116). July 2004. Lorient, France.

Kilgarriff, A., Huang, C. R., Rychly, P., Smith, S. & Tugwell, D. 2005. 'Chinese word sketches', In *Proceedings of Asialex*. June 2005. Singapore.

Knight, D., Evans, D., Carter, R. & Adolphs, S. 2009. 'HeadTalk, HandTalk and the corpus: towards a framework for multi-modal, multi-media corpus development', *Corpora* 4 (1), 1–32.

Krashen, S., Butler, J., Birnbaum, R. & Robertson, J. 1978. 'Two studies in language acquisition and language learning', *ITL Review of Applied Linguistics* 39–40, 73–92.

Leech, G. 1998. 'The special grammar of conversation', *Longman Language Review* 5, 9–14.

Li, D. 2008. *Pingxing Yuliaoku yu Jiji-xing Han Ying Cidian de Yanbian (Parallel Corpora and Development and Compilation of Active Chinese-English Dictionaries)*. Shanghai: Shanghai Translation Publishing House.

Li, E. 2002. *Xiandai Cidianxue Daolun (Introduction to Modern Lexicography)*.

Shanghai: Chinese Dictionary Publishing House.

Li, J. 1922. 'Guoyu jiben yuci de tongji yanjiu (A statistic analysis of basic vocabulary Chinese)', *Guowen Xuehui Congkan (Journal of Chinese Association)* 1 (1).

Li, J. & Zhong, L. 2011. 'Jiyu pingxing yuliaoku de jiji-xing Han Ying cidian peili yuanze (Illustration principles for parallel corpus-based active Chinese-English dictionary compilation)', *Jiangsu Daxue Xuebao (Journal of Jiangsu University, Social Science Edition)* 13 (2), 44-48.

Li, L. 2006. 'Cong shuangyu shangwu cidian kan zhuanye yuliaoku dui cidian bianzuan de zhongyaoxing (Specialised corpus for dictionary compilation seen in a bilingual business dictionary), *Cishu Yanjiu (Lexicographical Studies)* 2006 (3), 93-100.

Liu, E. S. 1973. *Frequency Dictionary of Chinese Words.* Berlin: Mouton.

Liu, Y., Yu, S., Zhu, X. & Duan, H. 2005. 'Xiandai Hanyu xuci zhishiku de jianshe (The building of knowledge base of contemporary Chinese function words)', *Yuyan Wenzi Yingyong (Applied Linguistics)* 2005 (1), 130-136.

Luo, L. 2008. 'Guanyu jianli Hanyu xuexizhe yuliaoku de sikao (Some thoughts on the construction of the learner Chinese corpus)', *Gaodeng Gongcheng Jiaoyu Yanjiu (Higher Engineering Education and Research)* 2008 (Supplementary issue), 47-49.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics.* Cambridge: Cambridge University Press.

McEnery, T. & Xiao, R. 2004. 'The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study', In M. Lino, M. Xavier, F. Ferreire, R. Costa, R. Silva (Eds.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004* (pp. 1175-1178). May 2004, Lisbon.

McEnery, T. & Xiao, R. 2010. 'What corpora can offer in language teaching and learning', In E. Hinkel (Ed.) *Handbook of Research in Second Language Teaching and Learning* (Vol. 2) (pp. 364-380). London and New York: Routledge.

Mindt, D. 1996. 'English corpus linguistics and the foreign language teaching syllabus', In J. Thomas & M. Short (Eds.) *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech* (pp. 232-247). London: Longman.

Mohamad-Ali, A. 2007. 'Semantic fields of problem in business English: Malaysian and British journalistic business texts', *Corpora* 2 (2), 211–39.

Mukherjee, J. 2006. 'Corpus linguistics and English reference grammars', In A. Renouf & A. Kehoe (Eds.) *The Changing Face of Corpus Linguistics* (pp. 337–54). Amsterdam: Rodopi.

National Language Committee. 1988. *Xiandai Hanyu Changyongzi Biao (Commonly Used Characters in Mandarin Chinese).* Beijing: National Language Committee.

Piao, Y. 2010. 'Muyu wei Han yu de xiandai Hanyu zhongjieyu yuliaoku (L1 Korean Learners' Interlanguage Corpus of Modern Chinese)'. Chonbuk National University, Korea.

Project Code 748. 1976. *Xiandai Hanzi Zonghe Shiyong Pindu Biao (A Comprehensive Frequency Table of Character Usage in Modern Chinese.* Beijing.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1972. *A Grammar of Contemporary English.* London: Longman.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Ren, H. 2010. 'Guanyu zhongjieyu yuliaoku jianshe de jian dian sikao (Some thoughts on the construction of the Chinese interlanguage corpus), *Yuyan Jiaoxue yu Yanjiu (Language Teaching and Research)* 2010 (6), 8-15.

Reppen, R. 2009. 'English language teaching and corpus linguistics: lessons from the American National Corpus', In P. Baker (Ed.) *Contemporary Corpus Linguistics* (pp. 204–13). London: Continuum.

Schmitt, N. 2004. *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam: John Benjamins.

Shen, X. 2009. 'Jiyu yuliaoku de waiguo xuesheng shuang binyu ju pianwu fenxi (A corpus-based error analysis of foreign students' double object sentences)', *Yuwen Xuekan (Journal of the Chinese Language)* 2009 (12), 1-4.

Shi, L. 2008. 'Jiyu zhongjieyu yuliaoku de chengyu shiyong pianwu fenxi (An interlanguage corpus based error analysis of Chinese idioms)', *Shehui Kexuejia (Social Scientist)* 130 (2), 158-161.

Shortall, T. 2007. 'The L2 syllabus: corpus or contrivance?', *Corpora* 2 (2), 157-185.

Siewierska, A., Xu, J. & Xiao, R. 2010. '*Bang-le yi ge da mang* (offered a big helping hand): A corpus study of the splittable compounds in spoken and written Chinese', *Language Sciences* 32 (4), 464-487.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. 2004a. *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Sinclair, J. (Ed.) 2004b. *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins.

Sinclair, J., Jones, S. & Daley, R. 1970. *The OSTI Report.* Available in reprint as Sinclair *et al.* (2004).

Sinclair, J., Jones, S., Daley, R. & Krishnamurthy, R. 2004. *English Collocational Studies: The OSTI Report.* London: Continuum.

Svartvik, J. (Ed.) 1990. *The London–Lund Corpus of Spoken English: Description and Research.* Malabar, Florida: Krieger Publishing Company.

Tao, X. & Zhu, J. 1923. *Pingmin Qian Zi Ke (Thousand Character Lessons for Civilians).* Youth Association Book Store.

Taylor, L. & Barker, F. 2008. 'Using corpora in language testing', In N. H. Hornberger (Ed.) *The Encyclopedia of Language and Education,* Volume 7 – Language Testing and Assessment (pp. 241–54). New York: Springer.

Thorndike, E. 1921. *The Teacher's Word Book.* New York: Columbia Teachers College.

Tono, Y. 2009. 'Integrating learner corpus analysis into a probabilistic model of second language acquisition', In P. Baker (Ed.) *Contemporary Corpus Linguistics* (pp. 184–203). London: Continuum.

Tsang, W. L. & Yueng, Y. 2010. 'The construction of a Mandarin interlanguage corpus', In R. Xiao (Ed.) *Proceedings of the 2010 Conference of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2010)*. July 2010, Ormskirk.

Tsou, B. K. & You, R. 2010. *Quanqiu Huayu Xin Ciyu Cidian (A Dictionary of Global Chinese Neologisms)*. Beijing: The Commercial Press.

Tsou, B. K., Chin, A. C. & Kwong, O. Y. 2011. 'From synchronous corpus to monitoring corpus, LIVAC: The Chinese case'. Paper presented at the Third International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2011), St. Maarten.

Wang, H. 2006. 'Hanyu cihui tongji yanjiu (Statistical research of Chinese vocabulary)'. Working paper. Singapore: Singapore National University.

Wang, H. & Wang, T. 2009. 'Jianli zai yuliaoku jichu shang de Hanyu benti yanjiu yu duiwai Hanyu jiaoixue: yi xiandai Hanyu liheci yanjiu wei li (Corpus-based study of Chinese and Teaching Chinese as a foreign language: the case of splittable compounds in modern Chinese)', *Hanzi Wenhua (Chinese Script Culture)* 88 (2), 84-91.

Wang, M. 2005. 'Liuxuesheng *bi* zi ju xide de kaocha (A study of foreign learners' acquisition of *bi* structure)', *Jinan Daxue Huawen Xueyuan Xuebao (Journal of the College of Chinese Language and Culture of Jinan University)* 2005 (3), 28-35.

Wang, Y. 2010. 'Yuliaoku jieru de Hanyu yuwen cidian shiyi tujing (A corpus-based approach to definition in the Chinese dictionary)', *Cishu Yanjiu (Lexicographical Studies)* 2010 (1), 111-118.

Wen, Q., Wang, L. & Liang, M. 2005. *Zhongguo Xuesheng Yingyu Kou Bi Yu Yuliaoku (Spoken and Written English Corpus of Chinese Learners* (SWECCL 1.0). Beijing: Foreign Language Teaching and Research Press.

Wen, Q., Liang, M. & Yan, X. 2008. *Zhongguo Xuesheng Yingyu Kou Bi Yu Yuliaoku (Spoken and Written English Corpus of Chinese Learners (SWECCL 2.0)*. Beijing: Foreign Language Teaching and Research Press.

Wen, Q. & Wang, J. 2008. *Zhongguo Daxuesheng Ying Han Han Ying Kou Bi Yi Yuliaoku (Parallel Corpus of Chinese EFL Learners)*. Beijing: Foreign Language Teaching and Research Press.

Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (Eds.) 1997. *Teaching and Language Corpora*. London: Longman.

Xiao, H. 2010. 'Cidian duoyici yixiang guanxi yu ciyi qufen (The sense relations and sense distinction of polysemes in the dictionary)', *Yunnan Shifan Daxue Xuebao (Journal of Yunnan Normal University, Social Science Edition)* 42 (1), 43-48.

Xiao, R. & McEnery, T. 2004. *Aspect in Mandarin Chinese: A Corpus-Based Study.* Amsterdam: John Benjamins.

Xiao, R. & McEnery, T. 2008. 'Negation in Chinese: A corpus-based study', *Journal of Chinese Linguistics* 36 (2), 274-330.

Xiao, R. & McEnery, T. 2010. *Corpus-Based Contrastive Studies of English and Chinese.* London and New York: Routledge.

Xiao, R., McEnery, T. & Qian, Y. 2006. 'Passive constructions in English and Chinese: A corpus-based contrastive study', *Languages in Contrast* 6 (1), 109-149.

Xiao, R., Rayson, P. & McEnery, T. 2009. *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners.* London and New York: Routledge.

Xiao, R. & Tao, H. 2006. *The Lancaster Los Angeles Spoken Chinese Corpus.* Lancaster: UCREL, Lancaster University.

Xing, H. 2003. 'Liuxuesheng pianwu hechengci de tongji fenxi (A statistical analysis of errors with compound words by foreign learners of Chinese)', *Shijie Hanyu Jiaoxue (Teaching Chinese in the World)* 2003 (4), 67-78.

Yang, D. 2004. 'Riyu muyu xuexizhe quxiang buyu xide qingkuang fenxi (A study of the Japanese-speaking students' acquisition of Chinese directional complements)', *Jinan Daxue Huawen Xueyuan Xuebao (Journal of the College of Chinese Language and Culture of Jinan University)* 2004 (3), 23-35.

Yang, H. & Wei, N. 2005. *Zhongguo Xuexizhe Yingyu Kouyu Yuliaoku (College English Learners' Spoken English Corpus).* Shanghai: Shanghai Foreign Language Education Press.

Yang, Y., Li, S., Guo, Y. & Tian, Q. 2006. 'Jianli Hanyu Xuexizhe Kouyu Yuliaoku de jiben shexiang (Tentative ideas of constructing the Chinese Learners' Spoken Corpus)', *Hanyu Xuexi (Chinese Language Learning)* 2006 (3), 58-64.

Yu, S. 2003. *Xiandai Hanyu Yufa Xinxi Cidian (The Grammatical Knowledge-base of Contemporary Chinese).* Beijing: Tsinghua University Press.

Yu, S. & Duan, H. 2002. 'Beijing Daxue xiandai Hanyu yuliaoku jiben jiagong guifan (Specifications for basic processing of contemporary Chinese corpus at Peking University: Specification)', *Zhongwen Xinxi Xuebao (Journal of Chinese Information Processing)* 16 (5), 49-64.

Yuan, Y. (2005a) 'Shi xi zhongjieyu zhong gen *bu* xiangguan de pianwu (An analysis of interlanguage errors related to *bu*)', *Yuyan Jiaoxue yu Yanjiu (Language Teaching and Research)* 2005 (6), 39-47.

Yuan, Y. 2005b. 'Shi xi zhongjieyu zhong gen *meiyou* xiangguan de pianwu (An analysis of interlanguage errors related to *meiyou*), *Shijie Hanyu Jiaoxue (Chinese Teaching in the World)* 2005 (2), 56-70.

Zhang, B. 2008. 'Di'er yuyan xuexizhe Hanyu zhongjieyu yi hunxiao ci jiqi yanjiu fangfa (On confusable words in Chinese interlanguage and related research methods)', *Yuyan Jiaoxue yu Yanjiu (Language Teaching and Research)* 2008 (6), 37-45.

Zhang, B. 2010. 'Hanyu zhongjieyu yuliaoku jianshe de xianzhuang yu duice (The state of the art, problems, and coping strategies of the Chinese interlanguage corpus development)', *Yuyan Wenzi Yingyong (Applied Linguistics)* 2010 (3), 129-138.

Zhang, B. 2011. 'Waiguoren Hanyu jushi xide yanjiu de fangfalun sikao (On the methodology of the Chinese sentences acquisition by foreigners)', *Huawen Jiaoxue yu Yanjiu (TCSOL Studies)* 42 (2), 23-29.

Zhang, Q. 2010. 'Chengdu buyu *X de hen* de jufa ji yuyong fenxi (A syntactic and pragmatic analysis of the complement of degree *X de hen*)', *Xiandai Yuwen (Modern Chinese)* 2010 (8), 47-49.

Zhao, Z. 2010. 'Zai tan *laizhe*: jiyu yuliaoku de kaocha (A restudy of *laizhe*: a corpus-based investigation)', *Linyi Shifan Xueyuan Xuebao (Journal of Linyi Normal University)* 32 (2), 90-94.

Zheng, Y. 2006. 'Zhongjieyu zhong chengdu fuci de shiyong qingkuang fenxi (An analysis of foreign learners' use of degree adverbs)', *Hanyu Xuexi (Chinese Language Learning)* 2006 (6), 66-72.

Zhou, X. & Hong, W. 2010. 'Zhonggaoji liuxuesheng Hanyu zhongjieyu cige shiyong qingkuang kaocha (A study on the use of figures of speech in the Chinese interlanguage of the intermediate and advanced Chinese learners)', *Shijie Hanyu Jiaoxue (Teaching Chinese in the World)* 2020 (4), 536-547.

Zhu, X., Zhang, H., Duan, H. & Yu, S. 2004. '*Hanyu Gaopinci Yufa Xinxi Cidian* de yanzhi (Development of the *Grammatical Knowledge Base of Chinese High Frequency Words*)', *Yuyan Wenzi Yingyong (Applied Linguistics)* 2004 (3), 98-104.