The Applications of Keywords and Collocation to Translation-Studies and Teaching Based on the Translation Corpora
---A Tentative Research on the Parallel Corpus of 17th NCCPC Report

Dai Guangrong

(English Department Fujian University of Technology Fuzhou 350007)

Abstract: This paper tends to discuss the corpora applications to translation studies and teaching. Using the parallel corpus of 17th NCCPC (17th National Congress of the Communist Party of China) as the investigation corpus, this paper disagrees with the opinion of the keywords and collocations which based on an abrupt conclusion without scientific statistic strategy. It is very important for the translator(s) to identifying the keywords and key collocations in translation studies and teaching. Keywords provide a useful way to characterize a text or a genre which can offer a clear and comprehensive understanding of the original texts. Collocation is a mode of meaning which can help the translator(s) to decide the field and style of word-usage in translation.

    With the help of parallel corpora and some famous native language corpora such as BNC, BROWN etc, we can learn some useful strategies to deal with the collocation of the keywords and their translations.

Keywords: corpus-based translation, concordance, collocation, translation teaching and research.

1 Introduction: Corpora Applications to Translation Studies and Teaching

Translation Studies became a scholarly discipline in China these years, and translation teaching has become the foci of many translation researchers and teachers. This paper tries to apply translation corpora to translation studies and teaching, showing the tentative research methods in

daily translation studies and teaching.

Different approaches have been taken to Translation Studies, from the earlier workshop approach, the philosophical and linguistic approach, the functionalist approaches, to Descriptive Translation Studies, the post-structuralist and post-modernist approaches, and the cultural studies approach (Munday 2001; Richard Xiao 2008: forthcoming). Translation theories develop quickly, but the practice, especially the translation teaching, are lagging far behind theories.

With the rapid development of corpus linguistics in the mid-1980s, corpus linguistics started to be interested in translations. Firth is convinced that linguistic description can be used as a basis for translation: linguistic analysis at the grammatical, lexical, collocational and situational levels can be used as a basis for total translation (Firth 1968: 76-78). The Firthian approach, and its latter-day incarnation in corpus linguistics, might bear fruit in studies of translation (Kenny 2001: 21).

Corpora are useful for revealing "relations between frequency and typicality, and instance and norm" (Stubbs 2001:151). According to Baker (1993:243), "the most important task that awaits the application of corpus techniques in translation studies is the elucidation of the nature of translated text as a mediated communicative event". Some hypotheses such as translated texts tend to be more explicit, unambiguous, and grammatically conventional than their source texts have already been investigated using translation corpora (Baker & Malmkjaer 1998: 52). And some scholars point out the usefulness of corpora and corpus linguistic techniques in translation, such as providing a powerful tool to identify the characteristic features of translational language, and helping translators understand what translation is and how it works (Baker 1993:243).

One of the advantages of the corpus-based approach to translation studies and teaching is that it can reveal the "regularities of actual behavior" (Toury 1995:265). Toury stress the need to observe regularities—and provide explanatory hypotheses for those regularities—on the basis of

ever-expanding corpora of texts, claiming that explanations of features observed even in a single translation must rely on the study of bigger corpora (Kenny 2001: 57). Tymoczko (1998:652) predicates that "Corpus Translation Studies is central to the way that Translation Studies as a discipline will remain vital and move forward." This predication has been realized by an even-growing number of corpus-based translation studies (cf. Øverås 1998; Baker 1999; Baker 2000; Granger et al 2003; Wang Kefei 2004; McEnery et al 2006, 2007; Richard Xiao 2008, forthcoming etc).

In the remainder of this paper, we will present a tentative study that seeks to uncover some features of collocation and keywords in translation with the parallel corpora including 17th NCCPC—National Congress of the Communist Party of China report.

2 Keywords used in Translation Studies and Teaching

In translation studies and practices, we may encounter all kinds of texts among which some are complex and difficult to analyze the contents and plots, especially for the literary texts. It's great challenge for the beginners to translate the texts word by word, sentence by sentence without a holistic analysis and understand of the source texts.

In this research, we put forward the corpus-based translation studies and teaching. With the help of corpus software, we can retrieve some useful information such as word frequency, collocation, pattern, clusters, dispersion of the search words and other statistic data. In this section, we apply "Keywords" to our research. "Keywords" is program within the corpus software (i.e. Wordsmith) for identifying the key words in one or more texts.

Key words are those whose frequency is unusually high in comparison with some norm. Key words provide a useful way to characterize a text or a genre. "Keywords" offers an important measure for the content and language characteristics of the texts (Wang 2007: 27). It also can be

used as analyzing the stylistics and retrieving texts. The program compares two pre-existing word-lists, which must have been created using the program of WordList in Wordsmith. One of these is assumed to be a large word-list which will act as a reference file. The other is the word-list based on one text which you want to study. The aim is to find out which words characterize the text you're most interested in, which is automatically assumed to be the smaller of the two texts chosen. The larger will provide background data for reference comparison. Using keywords, we can understand the translation materials (Scott 2004: 95). We can analyze the texts which should be translated and obtain the important hints such as the theme, main characters and plot. Translator can form a schema for the content of the source texts and find the proper words for the equivalents.

We once asked students to do some translation practices including the materials of government documents (in this research, we deal with 17th NCCPC report text). Before the discussion of 17th NCCPC, some students found a research paper about the keywords of it. The paper[1] shows the keywords of the Chinese version of the report. Figure 1 is the result of the research:


```
<!--[if !supportLists]-->(1) <!--[endif]-->有 "党" 的句子数：125
<!--[if !supportLists]-->(2) <!--[endif]-->有 "人民" 的句子数：102
<!--[if !supportLists]-->(3) <!--[endif]-->有 "经济" 的句子数：77
<!--[if !supportLists]-->(4) <!--[endif]-->有 "民主" 的句子数：48
<!--[if !supportLists]-->(5) <!--[endif]-->有 "科学" 的句子数：46
<!--[if !supportLists]-->(6) <!--[endif]-->有 "教育" 的句子数：31
<!--[if !supportLists]-->(7) <!--[endif]-->有 "干部" 的句子数：27
<!--[if !supportLists]-->(8) <!--[endif]-->有 "军" 的句子数：19
<!--[if !supportLists]-->(9) <!--[endif]-->有 "收入" 的句子数：13
<!--[if !supportLists]-->(10) <!--[endif]-->有 "监督" 的句子数：13
<!--[if !supportLists]-->(11) <!--[endif]-->有 "法制" 的句子数：9
<!--[if !supportLists]-->(12) <!--[endif]-->有 "腐败" 的句子数：8
<!--[if !supportLists]-->(13) <!--[endif]-->有 "分配" 的句子数：8
<!--[if !supportLists]-->(14) <!--[endif]-->有 "医疗" 的句子数：7
<!--[if !supportLists]-->(15) <!--[endif]-->有 "农民" 的句子数：6
<!--[if !supportLists]-->(16) <!--[endif]-->有 "税" 的句子数：5
<!--[if !supportLists]-->(17) <!--[endif]-->有 "工人" 的句子数：4
<!--[if !supportLists]-->(18) <!--[endif]-->有 "房" 的句子数：2
<!--[if !supportLists]-->(19) <!--[endif]-->有 "民主党派" 的句子数：2
<!--[if !supportLists]-->(20) <!--[endif]-->有 "中医" 的句子数：0
```

Figure 1 Keywords of 17th NCCPC

It reports that the number of sentence with keyword of "party" is 125, other figures are: "people" 102; "economics" 77; "democracy" 48; "science" 46; "education" 31; "cadre" 27; "army" 19; "income" 13; "supervision" 13; "law" 9; and "herbalist doctor" zero, etc. Judging from the results, we can say, the research is based on the author's imagination or his own purpose. It wants to prove the researcher's imagination by offering the data but it breaks the rule of objectivity in the research.

Students were puzzled by the result for its subjectivism. In order to help the students to understand the long passage clearly, we use some corpus software to retrieve some data. The text of 17th NCCPC is a kind of government documents and has its specific language characteristics. We make a "Keywords" (17th NCCPC English text as the observed corpus, and "Selected works of Mao Zedong" as the reference corpus) using Wordsmith (Version 4.0). Figure 2 shows part of the result of the Keywords.

	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	DEVELOPMENT	226	1.08	102	0.05	691.73	0.0000000000
2	IMPROVE	123	0.59	25	0.01	456.97	0.0000000000
3	AND	1,536	7.35	8,357	4.07	416.40	0.0000000000
4	SYSTEM	129	0.62	135	0.07	275.41	0.0000000000
5	SOCIAL	86	0.41	61	0.03	222.26	0.0000000000
6	WILL	318	1.52	1,135	0.55	209.68	0.0000000000
7	CHARACTERISTICS	57	0.27	14		203.87	0.0000000000
8	OPENING	53	0.25	10		199.39	0.0000000000
9	CHINESE	121	0.58	189	0.09	198.77	0.0000000000
10	SCIENTIFIC	58	0.28	22	0.01	186.58	0.0000000000
11	REFORM	87	0.42	96	0.05	180.05	0.0000000000
12	PROMOTE	55	0.26	24	0.01	169.75	0.0000000000
13	INNOVATION	38	0.18	4		155.44	0.0000000000
14	CULTURAL	54	0.26	37	0.02	141.56	0.0000000000
15	OVERALL	29	0.14	0		138.18	0.0000000000
16	ENSURE	44	0.21	19		136.22	0.0000000000
17	ETHNIC	28	0.13	0		133.41	0.0000000000
18	WE	392	1.88	1,941	0.95	132.81	0.0000000000
19	PUBLIC	63	0.30	76	0.04	123.50	0.0000000000
20	DEVELOP	57	0.27	61	0.03	120.02	0.0000000000

Figure 2: Keywords of 17th NCCPC and Selected works of Mao Zedong

From figure 2, we can see the frequency in observed corpus and reference corpus of the key words and know the Keyness of the keywords in the observed corpus. So, we can draw a conclusion that: "development" is the most important key word in 17th NCCPC and this can shed light on theme and help the students to understand the translation materials.

We make parallel word lists of the bilingual versions of 17th NCCPC, just retrieving the content words (not the grammatical or functional words), and retrieve some data from the lists. It can shed some light on the contrastive study of the language and translations. Figure 3 and Figure 4 represent the first twenty content words of English and Chinese version respectively.

N	Word	Freq.	%
1	DEVELOPMENT	226	1.08
2	PARTY	169	0.81
3	PEOPLE	146	0.70
4	SYSTEM	129	0.62
5	REFORM	87	0.42
6	SOCIALIST	77	0.37
7	WORK	67	0.32
8	SOCIALISM	64	0.31
9	CHINA	60	0.29
10	CHARACTERISTICS	57	0.28
11	AREAS	54	0.26
12	SOCIETY	54	0.26
13	OPENING	53	0.25
14	EFFORTS	47	0.22
15	LAW	47	0.22
16	WORLD	43	0.21
17	INTERESTS	41	0.20
18	COUNTRY	39	0.19
19	INNOVATION	38	0.18
20	DEMOCRACY	36	0.17

Figure 3: Key words of English version

N	Word	Freq.	%
1	发展	281	0.96
2	建设	187	0.64
3	社会	152	0.52
4	社会主义	147	0.50
5	人民	140	0.48
6	中国	130	0.44
7	党	115	0.39
8	改革	101	0.34
9	制度	83	0.28
10	经济	81	0.28
11	文化	76	0.26
12	特色	66	0.22
13	体系	66	0.22
14	民主	58	0.20
15	科学	56	0.19
16	国家	53	0.18
17	创新	52	0.18
18	体制	52	0.18
19	开放	49	0.17
20	管理	46	0.16

Figure 4: Key words of Chinese version

From figure 3 and figure 4, the two word lists show some difference in the first 20 keywords. In English version, "development" appears 226 times, and other keywords not included in Figure 3

such as "economy" (36), "government" (36), "education" (35), "force" (34), "management" (34), "progress" (34), "support" (34), "market" (33), "cadre" (31), "need" (31). In Chinese version, "发展（fazhan）" appears 281 times. So, we can see the figures show some difference in the two lists. The results tell some information: the translation encounter a lot of changes such as word order, frequency etc. As translation teachers, we should ask the questions: why, where, when and how to change the translation strategies?

Judging from the analysis of the passage, we know the most important keywords is "发展(development)" in the report. With the data, we can direct the translation practice in many factors.

The students once translated the sentence in Chinese "推动科学发展，促进社会和谐" (tuidong kexue fazhan, cujin shehui hexie) into "Promote the development of science, accelerate the social harmony". In the Chinese sentence, "科学 (kexue)" used as an adverbial, modifying "发展 (fazhan)", because "发展(development)" is the theme of the report. This sentence is very important in the whole passage. With the help of this information, we can translate it into "pursue development in a scientific way, promote social harmony".

3 Collocation used in Translation Studies and Teaching

Peter Newmark once pointed out: "He (one who writes or speaks in a foreign language) will be 'caught' out every time, not by grammar, which is probably suspiciously 'better' than an educated natives, not by his vocabulary, which may well be wider, but by his unacceptable or improbable collocations"(1981:180).

In the translating process, when the translators transfer the native language into foreign language, they have to choose the proper collocations which is difficult to decide in many situations, such as the idiomatic usage, restricted collocations etc.

Collocation is not only connected to lexis, but also relate with cohesion of the whole text. The cohesive effect of collocation is a key factor to avoid the translationese. Halliday and Hasan (1976:286) said, "The cohesive effect of such pairs depends not so much on any systematic relationship as on their tendency to share the same lexical environment, to occur in COLLOCATION with one another. In general, any two lexical items having similar patterns of collocation—that is, tending to appear in similar contexts—will generate a cohesive force if they occur in adjacent sentences." These remarks remind the translators of paying attention to the collocations in the translating process; otherwise, it will encounter the problem of translationese.

Sinclair once said, "There are virtually no impossible collocations, but some are more likely than others" (1966:411). In this research, we follow Wei Naixing's definition of collocation: A collocation is a conventional syntagmatic association of a string of lexical items which co-occur in a grammatical construct with mutual expectancy greater than chance as realization of non-idiomatic meaning in texts (Wei 2002: 100).

Firth (1957:12) put forward that "You shall know a word by the company it keep." So, collocation is a mode of meaning: Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night* (ibid: 196).

Bowker (1998: 631) observes that "corpus-assisted translations are of a higher quality with respect to subject field understanding, correct term choice and idiomatic expressions."

Here we can retrospect the research which puzzled our students. It says, "This kind of statistic cannot count the collocability, especially the collocation of adjectives and adverbs with the keywords which can explain the Keyness of the problem."[1] There are many factors which affect

the collocability, such as semantic, grammatical and conventional ones. All these factors should be considered in our daily translation practices.

Using the parallel corpora, we can decide the proper collocations in the target language. For example, when we encounter the word "结果(jieguo)"，it is very limited by using the Chinese-English dictionaries, for the dictionary just tell us some limited results and collocations. But the larger corpora will avoid these kinds of limitations. The parallel corpora will show us the plenty collocations and it will give the translators a lot of choices to decide the proper collocations. The translations of "结果" are plenty, such as "result/ outcome/ aftermath/ consequence/ effect/ conclusion/ sequel/ finding/ end" etc. with the parallel corpora, we can know the proper collocations of this keyword.

Other example:

In translating "前任（qianren）" into English, students face the difficulties of deciding the collocation even with the help of dictionaries.

With the parallel corpus, we can solve this kind of problem. "前任（qianren）"as the search item, we can draw some examples like the followings:

[1]"Especially with the growth of the amateur program, it's here to stay," says former heavy-weight champion Evander Holyfield.

[2]"You'll recognize the ornaments," said my former daughter-in-law.

[3] Her arrogance has disenchanted many of her former admirers.

[4]He is working hard to excel his predecessors.

[5]John Smith is a past president of our club.

[6] Mr. Heath is the former Prime Minister of Britain.

[7]The saddest were the eight ex-cadres who lost their executive jobs.

[8]The late Prime Minister attended the ceremony.

From these examples, we can say, "前任（qianren）" is not simply equal to "former", the collocations show that "predecessor/past/ex-/late" etc can used the equivalences (Dai, 2008: forthcoming).

Just as the same issue, when the students encounter the phrases, such as "台湾问题 (Taiwan wenti)", maybe we cannot find the translation from the dictionary, and it is difficult for the beginners to choose the collocations, "Taiwan problem/Taiwan issue/Taiwan question" etc. We can consult the parallel corpora and can obtain abundant examples. Figure 5 shows us the result of input collocation "台湾问题 (Taiwan wenti)" from the parallel corpus of 17th NCCPC.
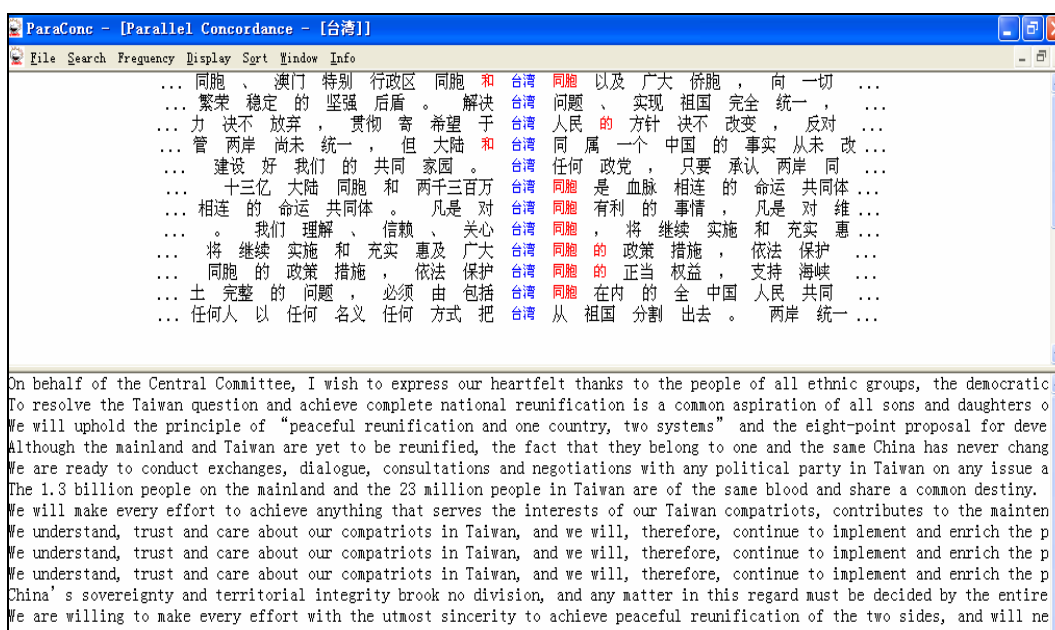

Figure 5

In the practice of translating the 17th NCCPC Chinese version into English, students encounter a lot of collocation problems. We can find a lot of translationese in their translation practice. How to improve this kind of situation? We still depend on the parallel corpus. First, we suggest students to make some KWICs practices from the corpus and learn to find the restricted collocations, such as

the Cluster showed in Figure 6:

```
     Cluster                      Freq.
1    THE SCIENTIFIC OUTLOOK        19
2    SCIENTIFIC OUTLOOK ON         19
3    OUTLOOK ON DEVELOPMENT        19
4    THE DEVELOPMENT OF            15
5    AND SOCIAL DEVELOPMENT        14
6    DEVELOPMENT OF THE            12
7    ECONOMIC AND SOCIAL           11
8    DEVELOPMENT WE MUST           10
9    DEVELOPMENT WE WILL           8
10   APPLY THE SCIENTIFIC          8
11   AND DEVELOPMENT OF            8
12   SCIENTIFIC DEVELOPMENT AND    6
13   DEVELOPMENT AND PROGRESS      6
14   DEVELOPMENT AMONG REGIONS     6
15   TO SCIENTIFIC DEVELOPMENT     5
```

Figure 6

Maybe, it's difficult for the beginners to decide the collocation whether it's a strong collocation or has a high mutual information of the collocation. With the help of corpus software, such as Wordsmith, A Corpus Worker's Toolkit etc, we can obtain the MI value of the collocation. The students are not sure the collocation "scientific development" are proper, we can get the confidence from the result of Figure 7:

```
                     Results:

        Frequency of node word [scientific] = 299
        Frequency of collocate [development] = 242
Frequency of node and collocate [scientific development] = 4
                   Window span = 4
              No of words in corpus = 1015530

       Mutual Information (Church et al 1991)= 5.81
                    MI (BNC) = 3.81
                   MI3 (BNC) = 7.81
```

Figure 7

From the figure, we can have a strong confidence about the collocation of "scientific development" for its mutual information value is 5.81.

We can also compare the results of "scientific development" and "science development" showed in Figure 8. The inter-pair T-score can strengthen the confidence of the translators. It can judge the proper collocation of "scientific development" in the target language.

```
Results:

* 1st word pair is 'scientific development', N=175
* 2nd word pair is 'science development', N=2
* Corpus size = 44300000 words

Inter-pair T-Score = 13.00

Hints: 'scientific development' is 13.00 standard deviations more likely
than 'science development', or,
 'science development' is 13.00 standard deviations less likely than
'scientific development'.

FYI, the usual confidence threshold is 1.65.
```

Figure 8

4, Conclusions

Corpus-based translation studies and teaching encounter a lot of challenges and opportunities in the information era. The present paper has explored a case study about the keywords and collocation in corpus-based translation studies and teaching. The keywords can provide some useful information about the materials for translation and help the translators to decide the translation strategies. The collocation strategy can help the translators to decide the proper collocations in target language and check the the naturalness in the translation with the help of translation corpora. The two aspects give us some confidence in translation and teaching. All these can stimulate the improvement and innovations of the translation studies and teaching.

Notes:

[1] http://sns.chinavalue.net/Blog/BlogThread.aspx?EntryId=25430 2008-2-16

References

Baker, M. (1993), 'Corpus linguistics and Translation Studies: Implications and applications', in

   M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honour of John

   Sinclair*, 233-250. Amsterdam: John Benjamins.

Baker, M. and Kirsten Malmkjaer. (1998), *Routledge Encyclopedia of Translation Studies.*

 London: Routledge.

Baker, M. (1999), 'The role of corpora in investigating the linguistic behaviour of professional translators.'

   International Journal of Corpus Linguistics 4: 281-298.

Baker, M. (2000), 'Towards a methodology for investigating the style of a literary translator.' *Target* 12(2):

   241-266.

Barlow, M. (2000), 'Parallel texts and language teaching', in S. Botley, A. McEnery and A. Wilson (eds)

   Multilingual Corpora in Teaching and Research, 106-115. Amsterdam: Rodopi.

Bowker, L. (1998), 'Using specialized native-language corpora as a translation resource: A pilot study'.

 Meta 43(4): 631-651.

Dai Guangrong. (2008 forthcoming), A Corpus-based Research on Translation between Chinese and

English Words/Phrases, *Xiamen University of Technology Journal.*

Granger S., Lerot, J., and Petch-Tyson, S. (eds) (2003), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Amsterdam: Rodopi.

Firth, J.R (1957) *Papers in Linguistics 1934-1951.* London: Oxford University Press.

Firth, J.R. (1968), *Selected Papers of J.R.Firth* 1952-59, edited by F.R.Palmer, London and Harlow: Longmans, Green and Co.Ltd.

Halliday, M. A. K. and Hasan, R. 1976. *Cohesion in English* [M]. London: Longman.
Kenny, D. (2001), *Lexis and Creativity in Translation. A Corpus-based Study.* Manchester: St. Jerome Publishing.

McEnery, A., Xiao, R. and Tono, Y. (2006), *Corpus-based Language Studies: An advanced resource book.* London: Routledge.

McEnery, A. and Xiao, R. (2007), 'Parallel and comparable corpora: What is happening?', in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 18-31. Clevedon: Multilingual Matters.

Munday, J. (2001), *Introducing Translation Studies: Theories and applications.* London: Routledge.

Newmark, P. (1981), *Approaches to Translation.* Oxford: Pergamon Press.

Øverås, L. (1998), 'In search of the third code: An investigation of norms in literary translation.' *Meta* 43(4): 557-570.

Scott, M. (2004), *The WordSmith Tools* (v. 4.0). Oxford: Oxford University Press.

Sinclair, J. (1966), Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M.A.K. Halliday, and R. H. Robins, eds., *In Memory of J. R. Firth*. 410-430. London: Longrman.

Stubbs, M. (2001), 'Text, corpora and problems of interpretation: A response to Widdowson.' *Applied Linguistics* 22(2): 149–172.

Toury, G. (1995), *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Tymoczko, M. (1998), 'Computerized corpora and the future of translation studies.' *Meta* 43(4): 652-660.

Wang Kefei. (2004), *Parallel Corpora: Construction and Applications*. Beijing: Foreign Language Teaching and Research Press

Wang Lifei. (2007) *Computer-Aided Second Language Research Methods and Their Applications*. Beijing: Foreign Language Teaching and Research Press.

Wei Naixing (2002) *The Definition and Research System of the Word Collocation.* Shanghai: Shanghai Jiaotong University Press.

Xiao, Z. and Yue, M. (forthcoming) Using corpora in Translation Studies: The state of the art. In P. Baker (ed.) *Contemporary Approaches to Corpus Linguistics*. London: Continuum.