What can SLA learn from contrastive corpus linguistics?

The case of passive constructions in Chinese learner English

RICHARD XIAO

Abstract: This article seeks to demonstrate the predictive and diagnostic power of the integrated approach that combines contrastive corpus linguistics with interlanguage analysis in second language acquisition research, via a case study of passive constructions in Chinese learner English. The type of corpora used in contrastive corpus linguistics is first discussed, which is followed by a summary of the findings from a published contrastive study of passive constructions in English and Chinese based on comparable corpora of the two languages. These findings are in turn used to predict and diagnose the performance of Chinese learners of English in their use of English passives as mirrored in a sizeable Chinese learner English corpus in comparison with a comparable native English corpus.

Keywords: contrastive analysis, corpus, learner English, passive construction, Chinese

1. Introduction

Over the past three decades, the corpus methodology has revolutionised nearly all branches of linguistics so that corpora have been increasingly accepted as essential resources in linguistic investigation. Two kinds of corpora that emerged in the 1990s have not only greatly contributed to the vitality of corpus linguistics but have also revived contrastive analysis and interlanguage research. They are learner corpora and multilingual corpora.

A learner corpus comprises written or spoken data produced by language learners who are acquiring a second or foreign language.[1] Data of this type has particularly been useful in language pedagogy and second language acquisition (SLA) research, as demonstrated by the

fruitful learner corpus studies published over the past decade (see Pravec 2002; Keck 2004; and Myles 2005 for recent reviews). SLA research is primarily concerned with 'the mental representations and developmental processes which shape and constrain second language (L2) productions' (Myles 2005: 374). Language acquisition occurs in the mind of the learner, which cannot be observed directly and must be studied from a psychological perspective. Nevertheless, if learner performance data is shaped and constrained by such a mental process, it at least provides indirect, observable, and empirical evidence for the language acquisition process. Note that using product as evidence for process may not be less reliable; sometimes this is the only practical way of finding about process. Stubbs (2001) draws a parallel between corpora in corpus linguistics and rocks in geology, 'which both assume a relation between process and product. By and large, the processes are invisible, and must be inferred from the products.' Like geologists who study rocks because they are interested in geological processes to which they do not have direct access, SLA researchers can analyze learner performance data to infer the inaccessible mental process of second language acquisition. Learner corpora can also be used as an empirical basis that tests hypotheses generated using the psycholinguistic approach, and to enable the findings previously made on the basis of limited data of a small number of informants to be generalised. Additionally, learner corpora have widened the scope of SLA research so that, for example, interlanguage research nowadays treats learner performance data in its own right rather than as decontextualised errors in traditional error analysis (cf. Granger 1998: 6).

A multilingual corpus involves two or more languages. Data contained in this kind of corpora can be either source texts in one language plus their translations in another language or other languages, or texts collected from different native languages using comparable sampling techniques to achieve similar coverage and balance. The two types of multilingual corpora are usually referred to as *parallel corpora* and *comparable corpora* respectively and

used in translation and contrastive studies (see section 2 for further discussion). Contrastive studies can be theoretically oriented or geared towards applied research. Theoretic contrastive studies are language independent and primarily concerned with how a universal category is realised in two or more different languages, whilst applied contrastive studies are preoccupied with how a common category in one language is realised in another language. In its early stage, contrastive linguistics was predominantly theoretic, though the applied aspect was not totally neglected. Theoretically oriented contrastive studies were continued from the late 1920s all the way into the 1960s by the Prague School. On the other hand, WWII aroused great interest in foreign language teaching in the United States, and contrastive studies were recognised as an important part of foreign language teaching methodology (cf. Fries 1945; Lado 1957). As a means of 'predicting and/or explaining difficulties of second language learners with a particular mother tongue in learning a particular target language' (Johansson 2003), applied contrastive studies were dominant throughout the 1960s. However, it was soon realised that language learning could not be accounted for by cross-linguistic contrast alone,[2] and as a result contrastive studies lost ground to more learner-oriented approaches such as error analysis, performance analysis and interlanguage analysis (cf. Johansson 2003). The revival of contrastive studies in the 1990s has largely been attributed to the corpus methodology and the availability of multilingual corpora (cf. Granger 1996: 37; Salkie 1999; Johansson 2003).

Both learner corpora and multilingual corpora have been important areas of corpus research since the 1990s. The introduction in the preceding paragraphs might have given an impression that the two areas have developed in parallel and are totally unrelated to each other. But in fact they are not. Recently, there has been a convergence between the two research areas, as reflected in the 'integrated contrastive model' which was initially proposed by Granger (1996). This article discusses how contrastive corpus linguistics and learner

corpus analysis can be combined to bring insights into SLA research via a case study of passive constructions in Chinese learner English.

2. Contrastive corpus linguistics

While multilingual corpora, and especially comparable corpora, are designed and created with the explicit aim of cross-linguistic contrast, all corpora have 'always been pre-eminently suited for comparative studies' (Aarts 1998: i). For example, the four English corpora of the Brown family (i.e. Brown, LOB, Frown, FLOB) were created for synchronic and diachronic comparisons of English as used in Britain and the US in the early 1960s and the early 1990s,[3] while the Lancaster Corpus of Mandarin Chinese (LCMC) was designed as a Chinese match for FLOB and Frown to facilitate cross-linguistic contrasts of English and Chinese (McEnery et al 2003). The International Corpus of English (ICE) project has used a common corpus design and the same sampling criteria for each of its components to ensure their comparability; similarly, the International Corpus of Learner English (ICLE) is designed in such a way that the subcorpora for learners of different L1 backgrounds are comparable (Granger 1998). Even a corpus like the British National Corpus (BNC), which was designed to be representative of modern British English, also provides a useful basis for various intra-lingual comparisons (e.g. genre-based variations and variations caused by sociolinguistic variables), though corpora that have adopted the BNC model such as PELCRA Reference Corpus of Polish and the American National Corpus (ANC) are undoubtedly suitable for contrastive studies of different languages or different varieties of the same language. Clearly, corpora are intrinsically comparative, and so is the corpus linguistics methodology. For example, collocations are extracted using statistic measures that compare the probabilities of co-occurring words within a specified window span of the node word; keywords are identified by comparing the target corpus with a reference corpus; what Granger (1998: 12) referred to as Contrastive Interlanguage Analysis (CIA) is also mainly concerned with

comparison, e.g. comparing interlanguage with target native language, and comparing different interlanguages (in terms of L1 background, age, proficiency level, task type, learning setting, and medium etc). In short, it can be said that the whole corpus research enterprise is based on comparison, for example, by comparing the same linguistic feature in different corpora, comparing different linguistic features in the same corpus, and comparing what is observed and what is expected.

While corpus linguistics is clearly comparative in nature, the technical terms for corpora used in linguistic comparison are somewhat confusing, with the controversy revolving around the issue of whether a parallel corpus should be a corpus composed of source texts plus translations, or a corpus containing native language data collected using comparable sampling criteria. As we have argued elsewhere (McEnery et al 2006: 47), a parallel corpus is composed of source texts and their translations, whilst a comparable corpus contains L1 texts sampled from different languages which are comparable in sampling criteria. A *translation corpus*, instead of referring to what is actually a parallel corpus as suggested in the literature, comprises translated texts for us in studies of translational language (e.g. the Translational English Corpus). Corpora which are designed primarily for intra-lingual comparison or for comparing different varieties of the same language (e.g. the ICE) are *comparative corpora*.

Having clarified the terminologies, it is appropriate to discuss what types of corpora are to be used in cross-linguistic contrasts. This is in fact an issue which is as debatable as the terminological issue. It has been argued that parallel corpora provide a sound basis for contrastive analysis, as demonstrated in the claims that ‘translation equivalence is the best available basis of comparison’ (James 1980: 178), and that ‘studies based on real translations are the only sound method for contrastive analysis’ (Santos 1996: i). However, as has been widely observed (Baker 1993: 243-5; Hartmann 1995; Gellerstam 1996; Teubert 1996: 247; Laviosa 1997: 315; McEnery and Wilson 2001: 71-72; McEnery and Xiao 2002),

translational language is ‘an unrepresentative special variant of the target language’ which is perceptibly influenced by the source language (McEnery et al 2006: 93). The source texts and translations in a parallel corpus are certainly comparable in terms of sampling criteria such as genres – in fact sampling only applies in selecting source texts but does not apply twice to translations, but this comparability is immediately undermined by so-called ‘translationese’ in translated texts. For example, Laviosa (1998) finds that translational language has four core patterns of lexical use: a relatively lower proportion of lexical words over function words, a relatively higher proportion of high-frequency words over low-frequency words, a relatively greater repetition of the most frequent words, and less variety in the words that are most frequently used. Beyond the lexical level, translational language is characterised by normalization, simplification (Baker 1993), explicitation (i.e. increased cohesion, Øverås 1998), and sanitization (i.e. reduced connotational meanings, Kenny 1998). In addition to these common features of translational language, Granger (1996) has noted some similarity between translationese and what she calls ‘learnerese’: ‘Both are situated somewhere between L1 and L2 and are likely to contain examples of transfer’, and both ‘give evidence of what Gellerstam (1986: 94) calls “syntactic fingerprints”’ (Granger 1996: 48).

As observations resulting from parallel corpus analysis usually invite ‘further research with monolingual corpora in both languages’ (Mauranen 2002: 182), parallel corpora can be a useful starting point of contrastive analysis. Nevertheless, it is also clear from the discussion above that while they are ideal resources for translation studies (see McEnery and Xiao 2007 for further discussion), parallel corpora provide a poor basis for cross-linguistic contrasts if relied upon alone. In the section that follows, we will present the findings of a corpus-based contrastive study of passive constructions in English and Chinese, which will be used to predict and diagnose what is observed in Chinese learner English.

3. Passive constructions in English and Chinese

This section summarises the results of a corpus-based study of passive constructions in English and Chinese, which was published in Xiao et al (2006). The primary corpus resources used in that study included FLOB for written English and LCMC for written Chinese, together with spoken corpora composed of transcripts for casual conversations in the two languages.[4] The corpus-based contrastive study yields a number of interesting findings. Below we will only give a summary of the results that are most relevant to our discussion of the performance of Chinese learners of English in the following section.

Firstly, passive constructions are nearly ten times as frequent in English as in Chinese, with normalised frequencies of 1,026 and 110 instances per million words for the two languages respectively. There are a number of reasons for this contrast. First, *be*-passives can be used for both stative and dynamic situations whereas Chinese passives can only occur in dynamic events; second, Chinese passives usually have a negative pragmatic meaning while English passives (especially *be*-passives) do not; third, English has a tendency to overuse passives, especially in formal writing whereas Chinese tends to avoid syntactic passives wherever possible; Chinese has a number of linguistic devices other than the syntactically marked passive constructions to express a passive meaning, e.g. notional passives, lexical passives, topic sentences, subjectless sentences, sentences with vague subjects (e.g. *youren* 'someone', *renmen* 'people', *dajia* 'all'), and special structures such as the disposal *ba* construction and the predicative *shi…de* structure. Finally, syntactically unmarked notional passives are more common in Chinese than in English because English is a subject-oriented language whereas Chinese is topic oriented. Given that Chinese passives are much more restricted in scope of use, their low frequency in relation to their English counterparts is unsurprising. It can be predicted from this sharp contrast in frequency of use that Chinese learners of English are very likely to underuse passives in their interlanguage.

Secondly, passives are formed by an auxiliary (*be, get*) followed by a past participial verb in English whilst in Chinese they can be marked syntactically by passive markers such as *bei*, indicated lexically by verbs with an inherent passive meaning (e.g. *zao* 'suffer'), or simply expressed by unmarked notional passives or special sentence structures. Unlike English, which inflects the passivised verb morphologically, Chinese is non-inflectional, which means that the same verb form is used for both active and passive voices in Chinese. Also because of the non-inflectional Chinese morphology, the concept of auxiliary is less salient or useful in Chinese. These cross-linguistic differences seem to suggest that the choice of correct auxiliaries as well as proper inflectional forms for passivised verbs can constitute a difficult area for Chinese learners to acquire English passives.

Thirdly, short passives (i.e. passives without a *by*-phrase introducing an agent) are typical of English, accounting for over 90% of total occurrences in both speech and writing. Short passives are predominant in English simply because passives are often used in English as a strategy that allows one to avoid mentioning the agent when it cannot or must not be mentioned.[5] In contrast, three out of five syntactic passive markers in Chinese (*wei…suo*, *jiao* and *rang*) only occur in long passives (i.e. passives with an explicit agent). For the two remaining passive markers *bei* and *gei*, which allow both long and short passives, the proportions of short passives (60.7% and 57.5% respectively) are significantly lower than that for English passives. Early Chinese grammarians (e.g. Wang 1984; Lü and Zhu 1979) noted that an agent must normally be spelt out in passive constructions, though this constraint has become more relaxed nowadays. When it is difficult to spell out the agent, passives are used in English, but an alternative device mentioned in the preceding paragraph is often used in Chinese instead of using passives. This finding can lead one to expect more long passives in the interlanguage of Chinese learners of English.

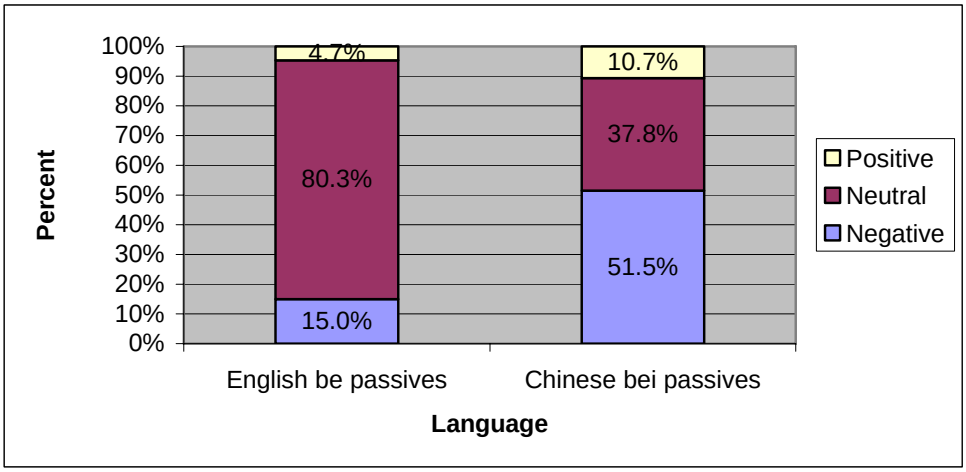

Figure 1. Pragmatic meanings of *be* and *bei* passives

Finally, a major distinction between passives in English and Chinese is that Chinese passives are more frequently used with an inflictive meaning than their English counterparts. With the exception of the archaic passive form *wei…suo*, over half of syntactically marked passives in Chinese occur in adversative situations, a proportion considerably higher than that for English passives (see Figure 1). As the prototypical passive marker *bei* was derived from a verb with an inflictive meaning (i.e. 'suffer'), Chinese passives were used at early stages primarily for unpleasant or undesirable events. While this semantic constraint on the use of passives has become more relaxed, especially in written Chinese, under the influence of western languages, disyllabic words made up of *bei* and a single character verb as used in modern Chinese typically refer to something undesirable, as in *beibu* 'be arrested', *beifu* 'be captured', *beigao* 'the accused', *beihai* 'be a victim' and *beipo* 'be forced'. In contrast, marking negative pragmatic meanings is not a basic feature of English passives, though *get*-passives often refer to undesirable events. An essential difference between English and Chinese passives lies in how much negativity is coded in them, which predicts that Chinese learners of English will use passives more frequently for undesirable situations.

In the next section, we will analyze the use of passives in a Chinese learner English corpus to ascertain how reliably the findings of our contrastive study as summarized in this section can predict and diagnose learner behaviour in interlanguage.

4. Passive constructions in Chinese learner English

This section examines *be* passives in Chinese learner English. The corpus used is the Chinese Learner English Corpus (CLEC), which contains one million words of essays written by Chinese learners at five proficiency levels: high school students (ST2), junior and senior non-English majors (ST3 and ST4), and junior and senior English majors (ST5 and ST6). The five types of learners are equally represented in the corpus. The corpus is fully annotated with learner errors using an error tagset that consists of 61 error types clustered in 11 categories (see Gui and Yang 2002). In order to compare Chinese learners' interlanguage with native English, the Louvain Corpus of Native English Essays (LOCNESS) is used as the control data, which is composed of argumentative essays written by native British and American students on a great variety of topics, totalling approximately 300,000 words (cf. Granger and Tyson 1996).

Table 1. Passives in CLEC and LOCNESS

Corpus	Words	Passives	Per million words	LL score
CLEC	1,070,602	9,711	907	1235.6
LOCNESS	324,304	5,465	1,685	($p<0.001$)

A comparison of CLEC and LOCNESS shows that in relation to native English writing, Chinese learners of English significantly underuse passives in their interlanguage. Table 1 gives the raw frequencies of passive constructions in the two corpora as well as the frequencies normalised to a common base of one million words. As can be seen, passives are nearly twice as frequent in native English as in Chinese learner English. The log-likelihood

test (LL) indicates that this difference is statistically significant (LL=1235.6 for 1 degree of freedom, $p<0.001$). The significant underuse of passives in Chinese learner English is hardly surprising in light of the marked contrast in frequencies for passives in English and Chinese as noted in section 3. Granger (1996: 46) also expected French learners of English to underuse passives in their writing as it was noted that passives were twice as frequent in English as in French (see Granger 1976), but she did not verify this prediction against French learner English data. While Chinese learners' underuse of passives as mirrored in the CLEC corpus is very likely to be caused by the influence of their native language, more cross-linguistic contrasts and interlanguage studies involving learners from other L1 backgrounds are required before we can be more confident that underuse of passives is the result of L1 transfer rather than a common feature of interlanguages, irrespective of the learner's mother tongue, which would mean that learners underuse passives for developmental reasons. As Granger (2007) observes, while native English speakers mainly use the verb *discuss* in the passive, 'learners show a predilection for active structures with first person subjects.'

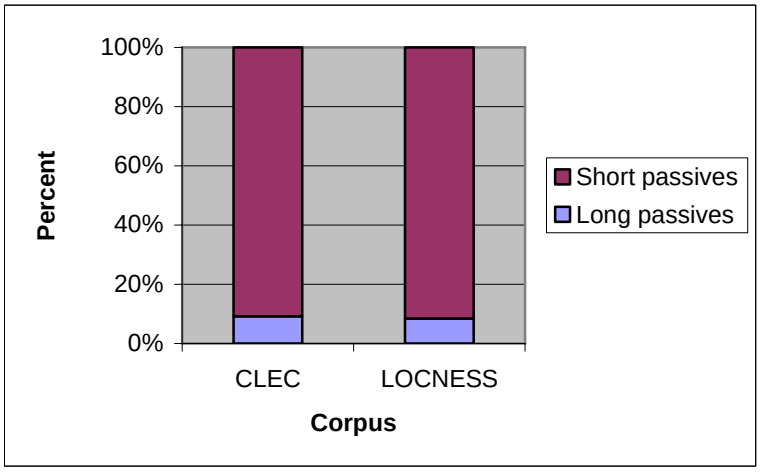

Figure 2. Long and short passive in CLEC and LOCNESS

The results of the contrastive analysis in section 3 predicted that Chinese learners would use long passives more frequently than native English speakers. Figure 2 shows the proportions of long and short passives in CLEC and LOCNESS. It can be seen that in

comparison with native English writings, long passives are indeed slightly more frequent in Chinese learner English (9.14% and 8.44% for CLEC and LOCNESS respectively), though this difference is marginal and not statistically significant (LL=2.18 for 1 degree of freedom, p=0.139).

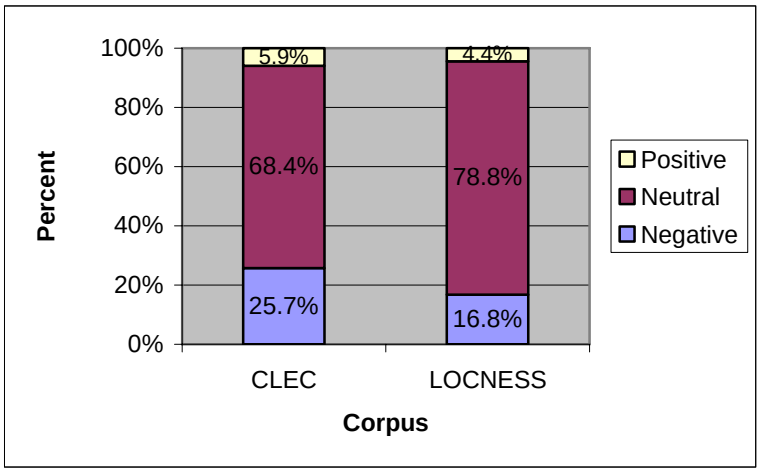

Figure 3. Pragmatic meanings of passives in CLEC and LOCNESS

It was noted in earlier that over 50% of passives in Chinese express an inflictive meaning whereas the corresponding figure for *be* passives in English is merely 15%. Such a contrast would reasonably lead one to expect more negative cases in Chinese learner English than in native English. This expectation is in fact supported by evidence from CLEC and LOCNESS. Figure 3 shows that 25.7% of passives in the Chinese learner English data are negative whilst negative cases account for 16.8% in native English writings. The log-likelihood test indicates that the differences between CLEC and LOCNESS in the three meaning categories are statistically significant (LL=7.4 for 2 degrees of freedom, p=0.025). A comparison of Figures 1 and 3 suggests that the proportions for the three meaning categories for the two types of native English data (i.e. general English and students' essays) are very close to each other. In contrast, the proportions in Chinese learner English shift away from those for L1 Chinese and move closer to the proportions for L2 English. Given that interlanguage is 'situated

somewhere between L1 and L2' (Granger 1996: 48), this movement is only reasonable and as expected.

An inspection of the specific errors related to the use of passive constructions in CLEC also demonstrates the value of contrastive corpus linguistics in SLA research. There are mainly four types of passive-related learner errors: underuse, misuse, misformation, and auxiliary errors. It can be considered as an advantage of the corpus-based approach to be able to view underuse or overuse of a linguistic feature in interlanguage as a type of learner error, as this was not possible in traditional error analysis without corpus data. Misuse of passives means that learners use passive constructions where they are not supposed to use them. Misformation errors are associated with morphological inflections, while auxiliary errors relate to omission and misuse of auxiliaries in passive constructions.

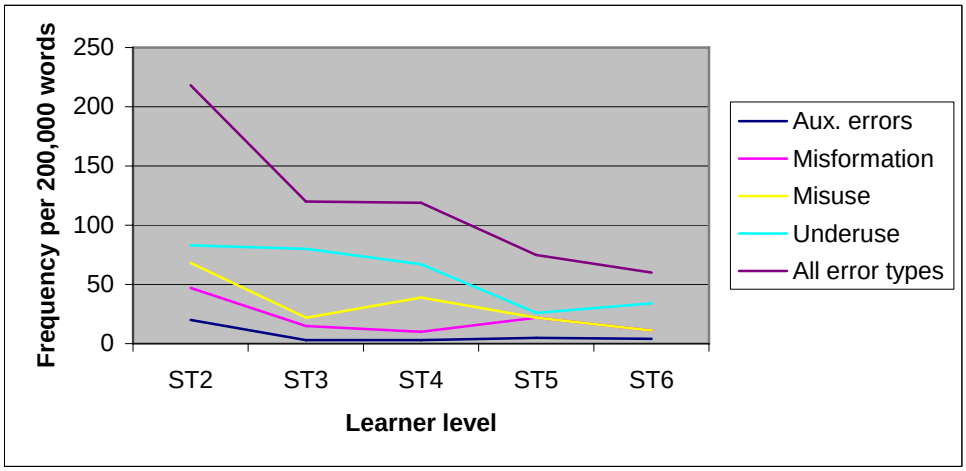

Figure 4. Passive-related errors in Chinese learner English

Figure 4 charts the distribution of four types of errors, as well as all error types as a whole, across learner proficiency levels. Unsurprisingly, when all error types are taken together, learners at higher levels generally make fewer errors related to passives. Of the four types of learner errors, underuse is the most important type, followed by misuse and misformation errors. Auxiliary errors are uncommon for learner groups other than the lowest level ST2 (i.e. high school students). It is also clear from the figure that learning curves are

not straight lines. There can be relapses in the language acquisition process, especially for difficult items.

Table 2. Association between error types and learner levels

From	To	LL score (3 d.f.)	P value
ST2	**ST3**	**27.303**	**<0.001**
ST3	ST4	6.955	0.073
ST4	**ST5**	**18.563**	**<0.001**
ST5	ST6	6.987	0.072

It is of interest to note that while error types are associated with learner levels when the dataset is taken as a whole (LL=51.77 for 12 degrees of freedom, $p<0.001$), similar leaner groups show similar error types. This means that the differences between the two non-English-major learner groups (i.e. ST3 and ST4), and between the two English-majors learner groups (i.e. ST5 and ST6) are not statistically significant, as indicated in Table 2. The table gives the log-likelihood test scores and probability values (3 degrees of freedom for all pairs of data), with significant differences highlighted. Hence, Chinese learners can be divided into three broad groups in terms of their acquisition of English passives: ST2 – ST3/ST4 – ST5/ST6.

While we cannot be conclusive of whether the underuse of passives by Chinese learners of English is a result of L1 transfer or a stage of the developmental path, errors of this type in our learner data typically occur with verbs whose Chinese equivalents are not normally used in passive constructions, as shown in (1).

(1) a. A birthday party *will hold* in Lily's house. (ST2)

b. The woman in white *called* Anne Catherick. (ST5)

(2) a. The supper *had done*. (ST2)

b. *wanfan zuo-hao le*

supper cook-ready ASP

The supper is ready.

Underuse errors also occur under the influence of topic sentences in Chinese, as exemplified in (2a), which is expressed in Chinese as (2b). The Chinese example in (2b) is an instance of topic sentence, which is very common in this language. Here *wanfan* 'supper' in the subject position is the topic and *zuo-hao le* 'cook-ready ASP' is the comment. Sentences like this cannot be used in the passive felicitously (e.g. **wanfan* ***bei*** *zuo-hao le*).

Misuse errors are mostly found in three contexts. Firstly, they occur when intransitive verbs are passivised (e.g. 3); secondly, errors of this type are related to the misuse of ergative verbs (e.g. 4); and finally, misuse errors can be a result of training transfer, i.e. excessive passive training in classroom instructions, as shown in (5). In sentences like these, the passivised verb is followed by an object, yet Chinese learners have been taught that passive transformation involves moving the object to the subject position. This can be taken as a symptom of the overdone passive training in English classrooms in China.

(3) a. A very unhappy thing *was happened* in this week. (ST2)

b. I *was graduated* from Zhongshan University. (ST5)

(4) a. the secince <sic science> *is developed* quickly (ST4)

b. infant mortality *was declined* (ST4)

(5) a. Because they *have been mastered* everything of this job (ST4)

b. many machine and appliance *are used* electricity as power (ST5)

Misformation errors are a result of L1 interference. As noted in section 3, passivised verbs do not inflect in Chinese. Consequently, Chinese learners of English tend to use uninflected verbs or misspelled past participles in passive constructions, as exemplified as (6).

(6) a. His relatives can not stop him, because his choice is *protect* by the laws. (ST6)

b. Since the People's Republic of china <sic China> was *found* on October 1949, great changes <…> (ST2)

(7) a. In China, since the new China *established*, people's life has goten <sic gotten> better and better. (ST3)

b. I am not a smoker, but why *do we forced* to be a second-hand smoker? (ST5)

Auxiliary errors, the final type of passive errors in our annotation scheme, are also the result of L1 interference. We noted earlier that while passives in Chinese can be marked syntactically, lexical passives, unmarked notional passives and topic sentences that express a passive meaning are abundant. As such, it is hardly surprising that Chinese learners of English tend to omit or misuse auxiliaries, as shown in (7).

The discussion in this section suggests that the performance of Chinese learners of English in their use of English passives is closely linked to their native language; and most of the passive-related errors in their interlanguage can be accounted for from the perspective of contrastive corpus linguistics. In the concluding section, we will discuss the implications of this study in SLA research.

5. Discussion and conclusions

This article first discussed the type of corpus data used in contrastive corpus linguistics, on the basis of which comparable corpora were used to contrast passive constructions in English and Chinese. The results of the contrastive analysis were used in turn to predict and diagnose the acquisition of passives by Chinese learners of English. The case study has clearly demonstrated the predictive and explanatory power of contrastive corpus linguistics in SLA research.

Combining contrastive analysis (CA) and contrastive interlanguage analysis (CIA) is undoubtedly a fruitful direction to pursue in SLA research. This is not a new idea. As early as a decade ago, Granger (1996: 46) proposed an 'integrated contrastive model':

> The model involves constant to-ing and fro-ing between CA and CIA. CA data helps analysts to formulate predictions about interlanguage which can be checked against CA data. […] Conversely, CIA results can only be reliably interpreted as being evidence of transfer if supported by clear CA descriptions.

Just as CIA has contributed significantly to SLA research by enabling and foregrounding many areas of investigation which have traditionally been impossible or marginalised (e.g. quantitatively distinctive features of interlanguage such as overuse and underuse, the potential effects of learner parameters on interlanguage), the integrated approach that combines CA and CIA will be an indispensable tool in SLA research, because 'if we want to be able to make firm pronouncements about transfer-related phenomena, it is essential to combine CA and CIA approaches' (Granger 1998: 14).

This emerging and promising area of research has recently become popular. For example, Gilquin (2001) demonstrates, on the basis of a case study of causative constructions in English and French, how the integrated contrastive model can help explain some of the characteristics of learners' interlanguage and thus throw new light on the key notion of transfer, which turns out to be a more complex phenomenon than has traditionally been assumed. Similarly, Borin and Prütz (2004) use the integrated contrastive approach to explore L1 syntactic interference in advanced Swedish learner English by investigating part-of-speech sequences. The increasing interest in the integrated approach is also demonstrated by the specialised workshop 'Linking up Contrastive and Learner Corpus Research', which was affiliated to the 4th International Contrastive Linguistics Conference.

We entirely agree with Granger (1996, 1998) that a combination of corpus-based contrastive study and interlanguage analysis can provide insights into language acquisition research, but we have different opinions of the role of parallel corpora (or 'translation corpora' in her words) in cross-linguistic contrasts, for the reasons outlined earlier in section

2. While Granger (1996: 38, 48) is fully aware of the drawback of using translated texts in contrastive analysis, her examples are largely based on data of this kind. In my revised CIA model, therefore, contrastive corpus linguistics interacts with interlanguage analysis on the basis of comparable native language corpora as illustrated in Figure 5.

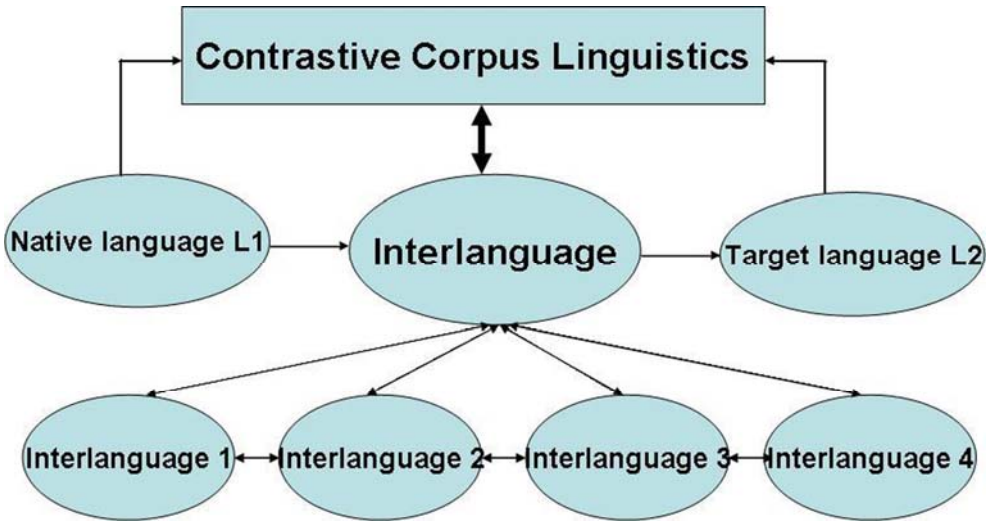

Figure 5. A revised model of contrastive interlanguage analysis

It is true that using a bidirectional parallel corpus can average out, to some extent at least, the undesirable effects of translationese on contrastive analysis. To achieve this aim, however, the same sampling criteria must apply to the selection of source texts in both languages, because any mismatch of proportion, genre, or domain, for example, may invalidate the findings derived from such a corpus (McEnery et al 2006: 93). A well-matched bidirectional parallel corpus is in fact a mixture of parallel corpus and comparable corpus, which can become a bridge that brings translation and contrastive studies together. Yet the ideal bidirectional parallel-comparable corpus will often not be easy, or even possible, to build because of the heterogeneous pattern of translation between languages and genres. This is especially true if the corpus aims to achieve sufficient coverage and balance to produce convincing findings (McEnery and Xiao 2007). Hence, in our approach, comparable native language data is preferred in contrastive corpus linguistics. Other kinds of corpora for comparative studies such as parallel corpora, translational corpora, and comparative corpora

are best suited for their own different purposes. Nevertheless, in spite of some difference in data type used, there has been increasing consensus that contrastive corpus linguistics has something to deliver in second language acquisition research.

Acknowledgements

I am grateful to the UK ESRC for supporting our project *Contrasting English and Chinese* (RES-000-23-0553), on which the work presented in this article was undertaken.

Notes

1. We maintain a distinction between a learner corpus and a developmental corpus, the latter of which is composed of data produced by children acquiring their mother tongue (L1).
2. See Sajavaara (1996) for a discussion of some problems with contrastive linguistics.
3. See Xiao (2007) for a review of the corpora mentioned in this article.
4. The two spoken corpora are the demographically sampled component of the BNC for English and the Callhome Mandarin Transcripts corpus released by the Linguistic Data Consortium (LDC).
5. Passives are also used in English for stylistic and coherence purposes. See Granger (1976, 1983) for more discussion of why passives are used in English and French.

References

Aarts, J. (1998) 'Introduction'. In S. Johansson and S. Oksefjell (eds.) *Corpora and Cross-linguistic Research*. Amsterdam: Rodopi. ix-xiv.

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*. Amsterdam: Benjamins. 233-352.

Borin, L. and Prütz, K. (2004) ‘New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language’. In G. Aston, S. Bernardini and D. Stewart (eds.) *Corpora and Language Learners*. Amsterdam: John Benjamins. 67–87.

Fries, C. (1945) *Teaching and Learning English as a Foreign Language*. Ann Arbor: University of Michigan Press.

Gellerstam, M. (1986) ‘Translationese in Swedish novels translated from English’. In L. Wollin and H. Lindquist (eds.) *Translation Studies in Scandinavia*. Lund: CWK Gleerup. 88-95.

Gellerstam, M. (1996) ‘Translations as a source fro cross-linguistic studies’. In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in Contrast*. Lund: Lund University Press. 53-62.

Gilquin, G. (2001) ‘The integrated contrastive model. Spicing up your data’. *Languages in Contrast* 3(1): 95–123.

Granger S. (1976) ‘Why the passive?’. In J. Van Roey (ed.) *English-French Contrastive Analyses*. Leuven: Acco. 23-57.

Granger, S. (1983) *The Be + Past Participle Construction in Spoken English with Special Emphasis on the Passive*. Amsterdam: North-Holland.

Granger, S. (1996) ‘From CA to CIA and back: An integrated approach to computerised bilingual and learner corpora’. In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in Contrast*. Lund: Lund University Press. 37-51.

Granger, S. (1998) ‘The computer learner corpus: a versatile new source of data for SLA research’. In S. Granger (ed.) *Learner English on Computer*. London: Longman. 3-18.

Granger, S. (2007) ‘Sylviane Granger: Interview’. *Mindbite* 1.

Granger, S. and Tyson, S. (1996) ‘Connector usage in the English essay writing of native and non-native speakers of English’. *World Englishes* 15: 19-29.

Gui, S. and Yang, H. (2002) *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.

Hartmann, R. (1995) 'Contrastive textology'. *Language and Communication* 5: 25-37.

James, C. (1980) *Contrastive Analysis*. London: Longman.

Johansson S. (2003) 'Contrastive linguistics and corpora'. In S. Granger, J. Lerot and S. Petch-Tyson (eds.) *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi. 31-44.

Keck, C. (2004) 'Corpus linguistics and language teaching research: bridging the gap'. *Language Teaching Research* 8(1): 83-109.

Kenny, D. (1998) 'Creatures of habit? What translators usually do with words?'. *Meta* 43(4).

Lado, R. (1957) *Linguistics across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan Press.

Laviosa, S. (1997) 'How comparable can "comparable corpora" be?'. *Target* 9: 289-319.

Laviosa, S. (1998) 'Core patterns of lexical use in a comparable corpus of English narrative prose'. *Meta* 43(4).

Lü, S. and Zhu, D. (1979) *Yufa Xiuci Jianghua* (Talks on Grammar and Rhetoric). Beijing: Chinese Youth Press.

Mauranen, A. (2002) 'Will "translationese" ruin a contrastive study?'. *Languages in Contrast* 2(2): 161-186.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T. and Xiao, R. (2002) 'Domains, text types, aspect marking and English-Chinese translation'. *Languages in Contrast* 2(2): 211-229.

McEnery, T. 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*. Clevedon: Multilingual Matters.

McEnery, T., Xiao, R. and Mo, L. (2003) ‘Aspect marking in English and Chinese: using the Lancaster Corpus of Mandarin Chinese for contrastive language study’. *Literary and Linguistic Computing* 18(4): 361-378.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Myles, F. (2005) ‘Interlanguage corpora and second language acquisition research’. *Second Language Research* 21(4): 373-391.

Øverås, S. (1998) ‘In search of the third code: an investigation of norms in literary translation’. *Meta* 43(4).

Pravec, N. (2002) ‘Survey of learner corpora’. *ICAME Journal* 26: 81-114.

Sajavaara, K. (1996) ‘New challenges for contrastive linguistics’. In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in Contrast*. Lund: Lund University Press. 17-36.

Salkie, R. (1999) ‘How can linguists profit from parallel corpora?’. Paper given at the Symposium on Parallel Corpora. 22-23 April 1999, University of Uppsala.

Santos, D. (1996) *Tense and Aspect in English and Portuguese: A Contrastive Semantical Study*. PhD thesis. Universidade Tecnica de Lisboa.

Stubbs, M. (2001) ‘Texts, corpora, and problems of interpretation: a response to Widdowson’. *Applied Linguistics* 22(2): 149-172.

Teubert, W. (1996) ‘Comparable or parallel corpora?’. *International Journal of Lexicography* 9(3): 238-264.

Wang, L. (1984) *Zhongguo Jufa Lilun* (Syntactic Theories in China). Qingdao: Shandong Education Press.

Xiao, R. (2007) ‘Well-known and influential corpora’. In A. Lüdeling and M. Kyto (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

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

Author’s correspondence address:

Department of Linguistics and English Language

Lancaster University

Lancaster, Lancashire

LA1 4YT

United Kingdom

E-mail: z.xiao@lancaster.ac.uk