An analysis of formal errors in a corpus of l2 English produced by Chinese students

Chuang, F-Y. and Nesi, H.

Original citation:
Chuang, F-Y. and Nesi, H. (2008) An analysis of formal errors in a corpus of l2 English produced by Chinese students. *Corpora* 1 (2), 251-271. DOI: 10.3366/cor.2006.1.2.251

http://dx.doi.org/10.3366/cor.2006.1.2.251

AN ANALYSIS OF FORMAL ERRORS IN A CORPUS OF L2 ENGLISH PRODUCED BY CHINESE STUDENTS

Chuang, F-Y & Nesi, H. (2006) An analysis of formal errors in a corpus of l2 English produced by Chinese students. *Corpora*, 1 (2) 251-271

ABSTRACT

This paper describes the investigation of a small corpus of writing in English for academic purposes produced by L1 speakers of Mandarin. The investigation involved the development of a tagset for the identification of formal errors in the corpus, and the subsequent analysis of these errors with a view to creating remedial grammar materials for Chinese students studying in the medium of English. Some prior approaches to error analysis are discussed, the process of developing the tagging system is described, and error types are identified, categorised, quantified, described and (as far as possible) explained.

1. INTRODUCTION

This paper reports on the initial stage of a project that aims to develop electronic self-access grammar materials for Chinese students studying in the medium of English. The project was initiated in response to a survey of the wants and needs of Chinese students attending a full-time foundation course in preparation for undergraduate entry to British universities (Wei, 2003). Although the Chinese students were fairly proficient users of spoken and written English, the survey indicated that they wanted grammar teaching to feature more prominently in the English language component of the course. Their English language tutors found it difficult to correct their recurring

grammar errors in class time, as lessons were primarily designed to develop academic literacy skills and did not focus on the accurate use of grammatical forms. To resolve this mismatch between student wants and course provision, we decided to provide extra grammar input for the Chinese students to access outside class. This provision would target their particular grammar problems and would not interfere with the kind of language teaching and study skills training that was taking place in their English classes.

To develop grammar-learning materials of this kind, it is useful to examine the formal errors Chinese students typically make. We decided to carry out a corpus-based error analysis of the students' academic writing, initially working with a small pilot corpus of 50 essays. The contributors were Chinese participants enrolled on the Business Studies strand of the foundation programme. All of them were aged between 18 and 21, with Mandarin Chinese as their L1. They had completed their middle school education in Mainland China and had been studying on the foundation programme for eight or nine months, with intermediate or upper intermediate English language proficiency (6.0 in IELTS or equivalent). The essays attempted to follow the conventions of undergraduate academic writing and dealt with serious topics in the Social Sciences, such as the ethics of genetic engineering, the European Monetary Union, methods of restricting car use, and the advantages and disadvantages of identity cards. Each essay was 1500 – 2000 words in length and the total corpus consisted of about 88,000 running words.

The process of error analysis involved four stages: error identification, classification, quantification and explanation. After the 50 typewritten essays were collected and

converted to electronic form, we first identified errors in the essays and provided corrections for them. We then followed our own specially devised tagging scheme to code the errors, and used WordSmith Tools version 3.0 (Scott, 1999) to search the tagged corpus and quantify error types. Instances of each type were then retrieved for further examination. The errors were systematically scrutinized and described. Tentative explanations for the causes of the errors were generated where possible at this stage.

2. THE DEVELOPMENT OF OUR ERROR TAGGING SYSTEM

Our own error categorisation system was developed in accordance with the three tagging principles proposed by Granger, Meunier and Tyson (1994): error categories should not overlap, should have precise definitions, and should describe, not explain. Further work by Dagneaux, Denness and Granger (1998) was also influential for our scheme. Dagneaux et al. emphasize the importance of using a purely descriptive system, criticizing EA categories because they often "rest on hybrid criteria" and "mix two levels of analysis: description and explanation" (ibid: 164). According to Dagneaux et al., categories such as *spelling error, grammatical error, vocabulary error* and *L1 induced error* are faulty because they are not mutually exclusive; the first three categories of error may also be classed as belonging to the fourth category, because they can all be L1 induced.

The principles of exclusivity and precision are followed fairly well by linguistic category taxonomies such as the one devised by Dulay, Burt and Krashen (1982). Linguistic category taxonomies describe errors in terms of the linguistic units they

belong to, for example a morpho-syntactic grammar error is described in terms of its word class (e.g. noun, verb) and the grammatical system it affects (e.g. tense, number, voice). Such systems are popular with teachers because the units included in the taxonomy are similar to those used in standard grammar reference books. They are reasonably straightforward to apply, provided that those involved in the tagging of errors work to the same specifications (for example by following a manual expressly created for their purposes). For this reason the linguistic category approach was also adopted by Granger et al. (1994) and Dagneaux et al. (1998) when tagging the ICLE corpus. ICLE error tags consist of one major category code to indicate the language level of the error, and a series of subcodes to show the linguistic unit it belongs to.

The problem with using a solely linguistic descriptive system, however, is that the descriptive detail is limited to the level of the word class and the grammatical system an error affects. The tag does not describe errors in terms of the placement or the choice of word**.** In contrast, the additional surface strategy taxonomy devised by Dulay, Burt and Krashen (1982) was designed to examine the ways surface structures are altered in terms of four kinds of deviations: omission, addition, misformation and misordering (ibid: 150). In this taxonomy omission errors were characterised by 'the absence of an item that must appear in a well-formed utterance' (either a content word or a grammatical morpheme) (ibid: 155), addition errors by 'the presence of an item which must not appear in a well-formed utterance' (ibid: 156), misformation errors by 'the use of the wrong form of the morpheme or structure' (ibid: 158), and misordering errors by 'the incorrect placement of a morpheme or group of morphemes' (ibid: 162).

Dulay, Burt and Krashen's taxonomy was helpful in that it identified four possible deviances in surface structures, but the error categories were not well defined and were not mutually exclusive. For example, both addition and misformation included a subtype called 'regularization' which was defined in terms of the same feature. Similarly an inflection error such as the verb error in '**I walk to school yesterday*' could be classified as either an omission error (omission of –ed) or a misformation error (the use of the incorrect form 'walk' for 'walked'). In response to these weaknesses James (1998) proposed a new taxonomy with five error types instead of four: omission, overinclusion, misselection, misordering and blend, but these five categories were still not clearly defined, and failed to provide a solution to the problem of overlap.

As both the linguistic category taxonomy and surface strategy taxonomy have their complementary strengths and weaknesses there is a strong argument for combining them, as James (1998) himself suggests. A combined-taxonomy approach can generate a bi-dimensional or even multi-dimensional error profile which can facilitate a more thorough understanding of learner errors. The description of both linguistic categories and surface structural deviances should facilitate more efficient and sophisticated error searches, and, as Dulay, Burt and Krashen (1982) claim, the combination should not only enrich the information the tag can provide, but also enable errors to be examined from different analytical perspectives.

Both kinds of taxonomies are descriptive in nature, and a combined system is by no means a 'descriptive-explanatory hybrid system' of the kind Dagneaux et al. deplored (1998: 166).

3. A TENTATIVE ERROR TAGGING SYSTEM

We therefore decided on an error tagging system consisting of two kinds of taxonomy: a linguistic category taxonomy (to describe errors in terms of the linguistic units they belong to) and a surface strategy taxonomy (to describe errors in terms of their surface structural deviances). For reliability each error category needed a clear definition, distinguishing it from other categories.

Following Granger and her fellow researchers, our linguistic category taxonomy used a hierarchical code structure consisting of one major category code and a series of subcodes. The major code indicated the targeted language level (grammatical {G}, lexical-grammatical {X} or lexical {L}) and the subcode indicated the linguistic unit of the error. For example, {dtar **G**} signified a grammatical determiner-article error, {n **X**nu} signified a lexical-grammatical noncount noun error, {av **L**ms} involved a lexical misspelt adverb, and so on. Grammatical errors were defined as morpho-syntactic errors mainly at the sentence level (i.e. concerning a specified word class). Word classes were identified in accordance with the Collins COBUILD English Language Dictionary (1994), with some modifications based on insights drawn from pedagogical grammars. Lexical-grammatical errors were defined as cases 'where the morpho-syntactic properties of words have been violated' (Granger et al., 1994: 107). The error subcategories were based on Granger et al.'s scheme and identified with reference to five features: the countability of nouns, the transitivity pattern of verbs, the attributive/predicative function of adjectives, the special syntactic pattern of a

word and the association of a preposition with a verb, a noun or an adjective. Our description of lexical errors also drew on the lexical classification devised by Granger et al. (1994), and included misspellings, non-existent L2 words (i.e. incorrect word coinage and borrowing), lexical misconceptions (i.e. misconceptions concerning the denotative or referential meaning of words), and collocational errors.

Our additional surface strategy taxonomy was an improved version of the one created by Dulay, Burt and Krashen (1982), incorporating notions from James' five categories (1998) where appropriate. For example, James' 'overinclusion' was adopted to replace Dulay et al.'s 'addition' because the term was more self-explanatory, and Dulay et al.'s 'misformation' and James' 'misselection' were redefined to create the contrastive categories of misformation and misselection, so that the relative frequency of mechanical (misformation) errors and conceptual (misselection) errors could be compared. We were interested in testing our hypothesis that the contributors to our corpus, as advanced learners of English, would make fewer misformation errors (relating to morphology and agreement) than misselection errors (relating to the expression of meaning).

Our system consisted of five categories: omission{-}, overinclusion {+}, misformation {#}, misselection {|}and misordering {[]}. An omission error was defined as a missing item (e.g. a word or a group of words) which would have appeared in a well-formed sentence. The missing item had to be a whole word; missing inflected morphemes (e.g. –s, -ed) were not tagged as omission errors. An overinclusion error was defined as a redundant item (e.g. a word or a group of words) which would not have appeared in a well-formed sentence. The overincluded item

had to be a whole word; redundant inflected morphemes (e.g. +s, +ed) were not tagged as overinclusion errors. Misformation was used to refer to a mechanical error that involved the use of the incorrect form of morpheme (e.g. an incorrect past tense form of a verb), whereas the term misselection was used when the selection of the incorrect item entailed a more complex conceptual judgement (e.g. the incorrect choice of tense/aspect). A misordering error involved the incorrect placement of an item in a sentence. With these specifications we aimed to overcome some of the major defects of Dulay et al.'s system and improve on the mutual exclusiveness of error categories.

A typical tag structure thus consisted of three parts, indicating the linguistic unit, surface alteration and language level of the error (e.g. '*She lives in {dtar – the G} UK', where 'dtar' signifies a determiner article error, '- the' signifies a missing 'the', and G signifies a grammatical error). Although detailed, the system did not attempt to describe the errors completely by means of their codes, but was rather designed to enhance the efficiency of error retrievals and facilitate the analysis of errors.

Though great efforts were made to create mutually exclusive error categories, some ambiguities remained. For example, establishing a clear-cut boundary between misformation errors and misselection errors was not straightfoward (e.g. Is '*He is interesting in reading' a misformation or misselection of 'interested'?). Dagneaux et al. (1998) have suggested that for the purposes of achieving consistent tagging it is more effective to exemplify error categories than to define them. Our tentative tagging system did not provide a complete set of predetermined subcategories for each error category, and we therefore needed to carry out pilot studies (i.e. apply the

tagging system to student essays) with a view to uncovering further possible subcategories and identifying suitable examples for inclusion in a tagging manual.

4. VALIDATING THE SYSTEM AND COMPILING A TAGGING MANUAL

To develop the tentative tagging system, we carried out a series of pilot studies to enlarge the tagset and revise the scheme. During the process, patterns emerged and new codes evolved. The whole development consisted of three stages, illustrated in Figure 1.

(Figure 1 here)

In the first stage, we tagged one essay, revised the system and applied it to ten essays. All the error types that emerged were added to the taxonomies; erroneous instances taken from those essays were fed into the system, so that gradually an exemplified system (tagging manual) could be compiled. The system then went through the second stage of development - checking inter-rater agreement. The two authors tagged the same essay individually, and then compared and discussed the differences and reached agreement about the tags and tagging rules. The system was modified accordingly, and the revised system was applied to 20 essays. In the process new error types and error instances emerged. The third stage involved further checking for inter-rater reliability. The two authors and one additional rater (all with PhDs in Applied Linguistics) tagged another essay independently following the newly compiled tagging manual. No training was given to the additional rater because we assumed that the manual was sufficiently self-explanatory. The three versions were compared. Because each tag contained three parts, each part (i.e. linguistic unit,

surface structural deviance and language level) was compared separately, and the reliability for each part was calculated. In the case of the first author versus the rater, the results indicated 81% reliability for language level, 81% reliability for linguistic category and 84% reliability for surface strategy. In the case of the two authors the results indicated 84% reliability for language level, 83% reliability for linguistic category and 81% reliability for surface strategy. These results suggest that the system was fairly stable, even though there were some discrepancies in rater judgement.

This validated system was applied to the whole corpus.

5. ERROR ANALYSIS RESULTS

Having tagged the entire corpus, we counted each error type and retrieved error instances for examination. As indicated above, the tagging system was designed to help us examine errors from a variety of perspectives. In this section, errors are first described from the perspective of linguistic categories at three language levels (grammatical, lexical-grammatical and lexical). They are then described according to their surface structure alternations.

5.1 The analysis of linguistic error categories

Table 1 shows the breakdown of errors in the three major categories (the three language levels). The results indicate that the relative frequency of grammatical,

lexical-grammatical and lexical errors was 85.9%, 5.0% and 9.1% respectively. A total of 5,232 errors were identified in the 50 essays, so each essay averaged just over 100 errors.

TABLE ONE HERE

Errors at each level were further analysed and examined. Tables 2, 3 and 4 show the breakdown of grammatical, lexical-grammatical and lexical errors respectively. Each table lists the error categories and their salient errors, together with some statistics and examples. The error in each example is underlined, followed by its correction marked in brackets. The term '*sic*.' is used to mark any other type of error which is not the focus of the designated category.

Table 2 shows the breakdown of grammatical errors. The results show that the ten most problematic linguistic features were determiners (27.6%), nouns (17.8%), verbs (8.9%), prepositions (8.1%), punctuation (5.9%), sentence parts (4.7%), tense/aspect (4.4%), modals (4.1%), conjunctions (3.9%) and pronouns (3.9%).

TABLE TWO HERE

Table 3 shows the breakdown of lexical-grammatical errors, together with salient error features and examples in each category. The results show that the incorrect association of a preposition with a verb, a noun, or an adjective was the most frequent cause of error (52.3%). The second most frequent error involved the countability of the noun (19.8%), usually the result of using a/an or the plural morpheme –s with a

noncount noun. The third most frequent error involved using the incorrect syntactic pattern of a word (14.9%), and the fourth involved the transitive verb (13.0%).

TABLE THREE HERE

Table 4 shows the distribution of lexical errors, together with the salient errors in each category. The results indicate that lexical misconception was the most frequent error type (63.1%), followed by collocational errors (23.1%), misspelling (13.4%) and non-existent words (0.4%).

TABLE FOUR HERE

5.2 The analysis of surface structure deviances

Table 5 shows the distribution of the surface strategy error categories, together with salient error features and statistics. Examples are not included because they have already been provided in Tables 2, 3 and 4. The results show that misselection was the most frequent error type (48.1%), followed by omission (24.7%), overinclusion (17.8%) and misformation (8.2%). Misordering errors (1.2%) were much less frequent than the other error types.

TABLE FIVE HERE

6. DISCUSSION

The data reveal some salient and systematic error features of the Chinese foundation students' interlanguage grammar. For example, grammatical errors were considerably more common than lexical and lexical-grammatical errors, suggesting that the students had more problems with morpho-syntactic features than with lexis. Lexical-grammatical and lexical errors together, however, made up 14.1% of the total number, implying that lexis and the morpho-syntactic properties of lexical items should not be ignored in the preparation of materials to help improve the students' accuracy when writing. In terms of the most frequent error type in each language level, mismanagement of the article system accounted for about 27.6% of grammatical errors, the incorrect association of a preposition with a noun, a verb or an adjective accounted for 52.3% of lexical-grammatical errors, and the incorrect choice of a lexical item (lexical misconception) accounted for 63.1% of lexical errors.

An examination of all three levels of errors shows that the students' formal errors fell into ten broad categories. These were, in order of frequency, determiners (23.7%), nouns (15.3%), verbs (7.6%), grammatical prepositions (6.9%), lexical misconceptions (5.8%), punctuation (5.1%), sentence parts (4.1%), tenses and aspects (3.8%), modals (3.5%) and conjunctions (3.3%). The top ten most frequent error features were:

Error type	**No. of errors**	**% out of all errors**
1) Missing definite article	529	10.1%
2) Bare count noun for plural	458	8.8%
3) Redundant definite article	446	8.5%
4) Misselection of preposition	321	6.1%
5) Lexical misconception	301	5.8%
6) Incorrect tense and aspect	198	3.8%
7) S-V non-agreement	125	2.4%
8) Incorrect collocation	110	2.1%
9) Missing 'a'/'an'	104	2.0%
10) Comma splice	103	2.0%

The top three most frequent error features were ‘missing definite article’, ‘bare count noun for plural’ and ‘redundant definite article’. Two of these involved the definite article, whilst the remaining one, ‘bare count noun for plural’, involved the omission of the plural morpheme (e.g. *Employer should provide free parking for their employees [Employers]) and was therefore related to the Ø article. As Palmer (1939, cited in Master, 1997: 221) suggested, there may be two forms of the zero article, one that occurs with non-count and plural nouns and the other that occurs with certain singular count and proper nouns. This would mean that the top three errors, together with the ninth most common error feature (missing ‘a/an’), all concerned the article system. Mismanagement of the article system was thus the most frequent cause of error in the corpus. Similar findings have been reported in other studies. For example, Milton (2001) examined Hong Kong university students’ writing and found four kinds of article errors among the top ten most frequent errors (‘singular noun for plural / Ø for indefinite article’ (1st), ‘indefinite article for Ø’ (3rd), ‘definite article for Ø’ (6th) and ‘definite article for indefinite article’ (8th)). Papp (2004) also found many errors concerning the article system and ‘number marking on nouns’ in a 200,000-word corpus of writing by Chinese ESL university students.

It is probable that several factors contribute to the high frequency of article errors in Chinese students’ writing. First, the articles (*a, an* and *the*) are used extremely frequently in writing. As the COBUILD frequency count (Sinclair, 1991) indicates, in a corpus of 20 million English words, *the* is the most common word, with a frequency rate of 35.0%, and *a* is the fifth most common (14.7%). Master (1997) compared the

frequencies of *the*, *a* and Ø in a corpus of 200,000 words, and found that the Ø article was more frequent (48.0%) than *the* (36.3%) and *a* (15.7%). Since *the* and *a/an* make up 8.5% of all text (Sinclair, 1993) and the Ø article is even more frequent than *the* or *a/an*, any difficulties with the article system are bound to make themselves apparent in learners' language production. Secondly, it appears that the complex concepts expressed within the English article system (e.g. specificity/non-specificity, genericness, definiteness/indefiniteness) make it particularly difficult for learners of English to master (Whitman, 1974; Master, 1990, 1997, 2002; Berry, 1993; Swan, 1995). This is especially true for those whose L1s have no articles (Swan, 1995). The L1 of the contributors to our corpus was Mandarin Chinese, which does not have an article system. Thirdly, the use of the article is closely related to the features of nouns (countability and number) (Celce-Murcia and Larsen-Freeman, 1983; Master, 2002). The Chinese language does not distinguish between count and noncount nouns, and does not have a rigid formal distinction between singular and plural (plural markers are not required), which make the concept of countability problematic for Chinese learners. Milton (2001) has reported that Hong Kong students find it very difficult to determine the countability of the noun and decide whether to assign plural forms. To make matters worse, bare singular noun forms are the normal form used to refer to something in general in the Chinese language.

We decided to use the term 'bare count noun error' to describe the erroneous use of a singular count noun without a determiner (e.g. The scientist is inserting a gene into a crop's DNA to get new *crop [a new crop]), or the erroneous use of a plural count noun without a plural marker (e.g. GE tomatoes are sold in *supermarket worldwide [supermarkets]). The very high frequency of bare count noun errors in our corpus

seems to suggest L1 interference, indicating that when treating bare count noun errors, L1 transfer should be taken into account. Chinese learners need to be reminded that a count noun needs a determiner when it is singular and a plural marker when it is plural. Also they need to learn to distinguish if a noun is countable or uncountable in different contexts - a great challenge for learners, teachers and materials writers in view of the complexity of the English noun.

Preposition errors were the fourth most frequent grammatical error category and the most problematic lexical-grammatical feature in our corpus. Milton (2001) also recorded the frequent occurrence of incorrect and redundant prepositions (his 2nd and 5th most frequent error types). Grammatical preposition errors identified in our corpus mainly involved incorrect prepositions, missing prepositions and redundant prepositions. For example,

- *Since a member of a stronger economic union, individuals could have more benefits. [as]
- People could create an animal just *getting the gene from the original animal…[by getting]
- People want to get a better quality * life. [a better quality of life]
- The United Kingdom still remain outside *of the European Union. [redundant *of*]

Lexical-grammatical prepositional errors mainly involved the incorrect association of a preposition with a noun, a verb or an adjective. For example,

- The car is different *with public transport. [different from]

- People are *<u>suffering</u> poverty. [suffering from]
- People lost confidence *<u>towards</u> the euro. [lost confidence in]

Grammatical preposition errors suggest that learners have problems with the roles/functions of prepositions in sentences, while lexical-grammatical preposition errors suggest that they do not know the proper association of a preposition with a lexical item. The preposition is particularly difficult because both global (syntactic) and local (lexical) features need to be taken into account when choosing a preposition in a particular context. In some cases, more than one preposition is acceptable. Chinese learners' L1 backgrounds are unlikely to help them because the Chinese preposition system is not as rigid and complicated as the English system. For example, the Chinese language only uses one preposition, "在" (Tsai), in association with various time references (year, month, week, day, time) whilst the English preposition system uses different prepositions (*in*, *on* and *at*). Moreover, learners may resort to word-for-word translation when trying to reproduce L2 phrases they are still unsure of, a strategy that could explain the following errors in the corpus:

- There is a huge increase *<u>of</u> part-time workers… [in]
- There is a radical reduction *<u>of</u> car use. [in]
- There is also a decline *<u>of</u> social integration,… [in]

We concluded that, apart from introducing the students to syntactic rules and lexical features, we also needed to provide them with more L2 exposure to increase their familiarity with English prepositional phrases.

The data show that tense and aspect errors occurred less frequently than many other error types in the corpus. The foundation course tutors surveyed for Wei's study (2003), however, claimed that this type of error was persistent in their students' writing. It is possible that tense and aspect errors were particularly salient to tutors because they perceived this area of grammar to be particularly important, and to cause particularly serious problems in communication.

In terms of surface deviances, the results show that the misselection errors (48.1%) identified in our corpus greatly outnumbered the misformation errors (8.2%). Misselection errors involve the selection of incorrect items and entail conceptual misjudgements, for example in the choice of tense/aspect, word class, voice or lexis, while misformation errors are those which involve incorrect forms of morphemes in cases such as S-V agreement, incorrectly formed irregular verbs, incorrectly formed noncount nouns, and misspelt or non-existent words. These two categories were used contrastively to test whether the students had more problems with conceptual or mechanical features. The fact that misselection errors were more frequent than misformation errors suggests that the students had more problems with conceptual judgement than with the mechanical application of rules. It confirmed our initial assumption that the Chinese foundation students, being intermediate or upper intermediate learners, would be able to correct many of their mechanical errors when editing their own writing. The data show that they tended to misselect noun forms (bare count nouns for plural forms), prepositions, lexical items, tenses/aspects and word classes. Remedial materials should thus use more consciousness raising activities to make students aware of problematic features in these areas and help them to understand correct form-function mapping. Although misformation errors were

much less frequent than misselection errors, the high frequency of S-V non-agreement errors (31.3% of verb errors) indicates that the students were not always capable of avoiding mechanical errors when they had to deal with the organisation of ideas and linguistic features at the same time. This lends support to VanPatten's (1990, 1996) input processing theory, according to which the L2 learner tends to prioritise meaning processing at the expense of formal accuracy when required to simultaneously attend to both meaning and form.

Another salient feature is that missing definite article errors accounted for 40.9% of omission errors and redundant definite article errors accounted for 47.9% of overinclusion errors. This indicates that the students had great difficulties in using the definite article correctly in their writing. As discussed above, concepts associated with the definite article are new and potentially difficult for Chinese learners; remedial materials focusing on this particular area are urgently needed.

7. CONCLUSION

This paper has reported the processes and findings of our investigation into the formal errors Chinese foundation students typically make in their written English production. Frequent features have been identified and possible causes of errors have been suggested. A tagging system was specially devised for this study; throughout the three stages, patterns emerged and influenced the development of the tagset, and an exemplified system (tagging manual) was gradually compiled. We found that this way of developing an error tagging system, though time-consuming, was useful in

two respects: 1) the tagset can now cater for most errors identified in the corpus, and 2) the manual has reduced tagging ambiguities to the minimum and aids consistent categorisation. Moreover, the use of a combined taxonomy and a hierarchical code structure in our tagging system has greatly facilitated the quantification and retrieval of errors, whilst enabling us to analyse errors from more than one perspective.

In view of the fact that mismanagement of the article system accounted for the largest number of errors identified in our corpus, we have prioritised article errors for treatment and are scrutinizing them and developing online self-study materials which focus on article use. These materials will form the first unit of *GrammarTalk*, an English for Academic Purposes grammar resource for Chinese students studying in the medium of English. The materials contain sample texts from both the Chinese foundation student corpus and the pilot version of the British Academic Written English (BAWE) corpus of proficient university student writing. Materials tackling other frequent errors are also needed, of course, and we hope in time to draw on the research findings reported here to develop further units in the *GrammarTalk* series.

ACKNOWLEDGEMENTS

GrammarTalk is currently under development with the aid of Teaching Quality Enhancement funding from the University of Warwick.

The British Academic Written English (BAWE) corpus is part of the project 'An investigation of genres of assessed writing in British Higher Education', funded by the Economic and Social Research Council (project number RES-000-23-0800). www.warwick.ac.uk/go/bawe

REFERENCES

Berry, R. 1993. Collins COBUILD English Guides (3): Articles. London: HarperCollins Publishers.

Celce-Murcia, M. and Larsen-Freeman, D. 1983. The Grammar Book. Cambridge, MA: Newbury House.

Dagneaux, E., Denness, S. and Granger, S. 1998. Computer-aided error analysis, System 26, pp 163-174.

Dulay, H., Burt, M. and Krashen, S. 1982. Language Two. New York, Oxford: Oxford University Press.

Granger, S., Meunier, F. and Tyson, S. 1994. New insights into the learner lexicon: a preliminary report from the international corpus of learner English in L. Flowerdew & A. K. K. Tong (eds.) Entering Text, pp102-113. Hong Kong: Language Centre, the Hong Kong University of Science and Technology.

James, C. 1998. Errors in Language Learning and Use: Exploring Error Analysis. Harlow: Addison Wesley Longman Limited.

Master, P. 1990. 'Teaching the English articles as a binary system', TESOL Quarterly 24(3), pp 461-478.

Master, P. 1997. 'The English article system: Acquisition, function, and pedagogy', System 25(2), pp 215-232.

Master, P. 2002. 'Information structure and English article pedagogy', System 30, pp 331-348.

Milton, J. 2001. Elements of a Written Interlanguage: A Computational and Corpus-Based Study of Institutional Influences on the Acquisition of English by Hong Kong

Chinese Students. Research Reports, Volume Two. Hong Kong: Language Centre, the Hong Kong university of Science and Technology.

Papp, S. 2004. 'The use of learner and reference corpora to foster inductive learning and self-correction in Chinese learners of English', paper given at the Meeting the Needs of the Chinese Learner in Higher Education Conference, 17-18 July 2004.

Scott, M. 1999. Wordsmith Tools version 3. Oxford: Oxford University Press.

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

Sinclair, J. 1993. Foreword in R. Berry, Collins COBUILD English Guides (3): Articles. London: HarperCollins Publishers.

Sinclair, J., Hanks, P., Fox, G., Moon, R. and Stock, P. (eds.) 1994. Collins COBUILD English Language Dictionary. London: HarperCollins Publishers.

Swan, M. 1995. Practical English Usage (Second edition). Oxford: Oxford University Press.

VanPatten, B. 1990. 'Attending to content and form in the input: An experiment in cousciousness', Studies in Second Language Acquisition 12, pp 287-301.

VanPatten, B. 1996. Input Processing and Grammar Instruction in Second Language Acquisition. Norwood, NJ: Ablex.

Wei, Y. 2003. Investigating Chinese HEFP Students' Target Needs: How to Help Chinese Students both in China and in the UK Prior to Entering British Tertiary Education. Unpublished MA dissertation, CELTE, University of Warwick, UK.

Whitman, R. L. 1974. 'Teaching the article in English', TESOL Quarterly 8(3), pp 253-262.

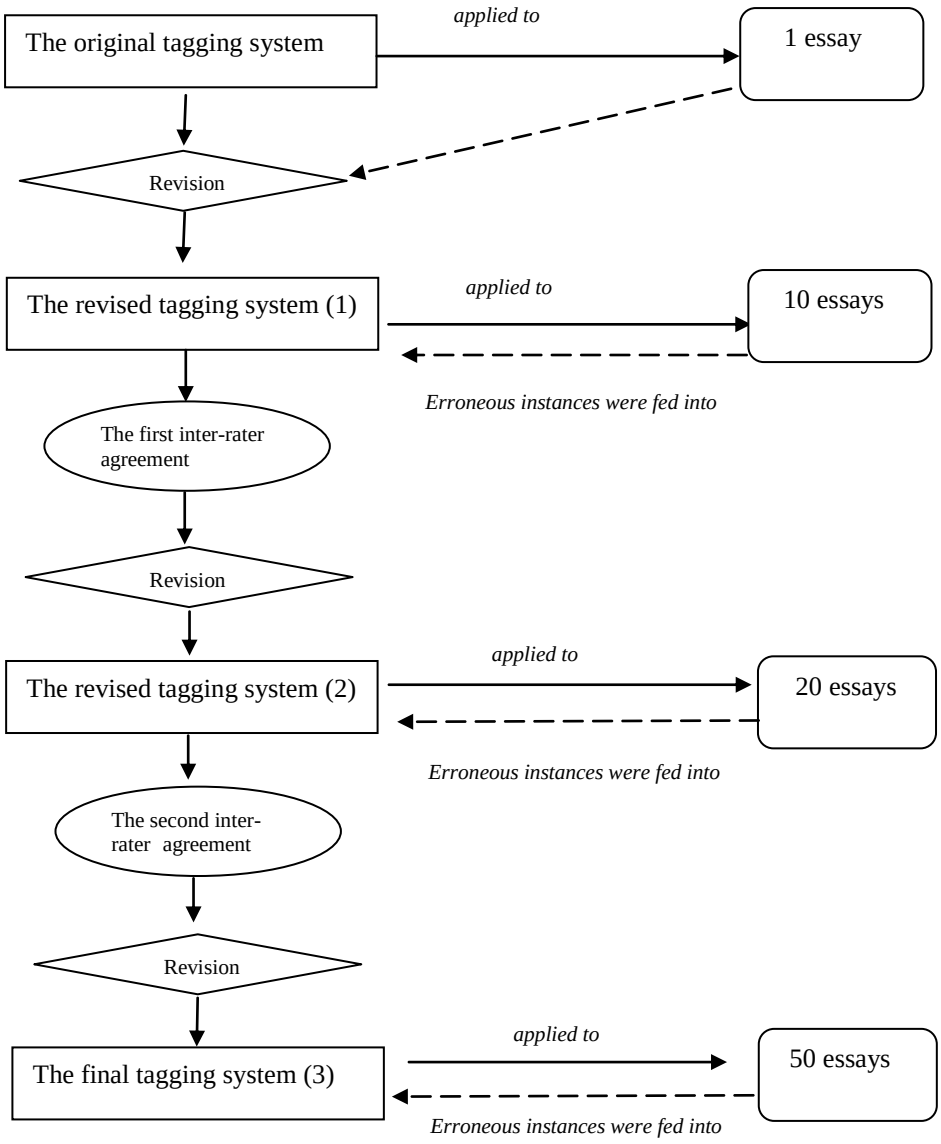

Figure 1: The three stages of development of the tagging system

Language level	Number of occurrences
Grammatical errors	4493 (85.9%)
Lexical-grammatical errors	262 (5.0%)
Lexical errors	477 (9.1%)
Total	5232

Table 1: The relative frequency of errors for each language level

Grammatical Category/ No. of errors (Frequency %)	**Salient errors/ No. of errors**	**Erroneous instances**
Determiner 1242 (27.6%)	Missing definite article (529)	From then on, racism has gone deep into human mind. [the human mind]
	Redundant definite article (446)	In the free [free] societies, people freely choose how to travel.
	Missing 'a'/'an' (104)	Although restricting the use of the car is [a] very complicated issue and is related to many problems,…
	Misselection between 'the' and 'a/an' (64)	Dr. Arpad Pusztai, a world renowned geneticist working at a [the] government-funded Rowett Institute in Aberdeen, Scotland, showed that…
Noun 800 (17.8%)	Bare singular noun for plural (458)	They provide parking area [areas] on the outskirts of cities.
	Noun for adjective (83)	Pollen can pass herbicide resistance [resistant] genes from genetically engineered crops to cultivated and wild relatives over a mile away.
	Quantifier/determiner-noun non-agreement in number (81)	There are seven different value [values] of the euro banknote: c5, c10, c20,c50, c100, c200 and c500.
Verb 399 (8.9%)	S-V non-agreement (125)	Genetic engineering also increases genetic diversity, and produce [produces] more variant alleles.
	Misselection between verb form, past participle and present participle (74)	Up to now, there is not any team claimed [claiming] that they have managed to clone a human being,
	Missing or redundant copula (58)	The European Communities Act 1972 enacted that relevant common law should [be] applied in the UK and should override English law. A single currency will be an important complement to the single European Market, which will make the EU be a more powerful factor…
	Verb for gerund (35)	*(Sic.)* Government may promote this by increase [increasing] the quality of the service.
Preposition 363 (8.1%)	Misselection between prepositions (226)	The policy of [on] road tax and petrol tax needs to be modified.
	Missing preposition (73)	[By] Simply manipulating the genes inside the food, scientists produce various kinds of GM food.
	Redundant preposition (63)	The new specie of plant could resist pest *(Sic.)* by itself without using of [using] artificial pesticide.
Punctuation 266 (5.9%)	Comma splice (103)	British taxpayers has *(Sic.)* properly funded private pensions, the [. The] euro will end up this funding.
	Redundant comma (87)	Some have argued that, people who are keen on identity cards see them as a way of getting at groups of people they dislike.
Sentence part 212 (4.7%)	Incorrect use of groups of words (79)	Germany cannot use its own fiscal and monetary policy to rescue the economy from the recession as suffering [because of] the EMU's fixed exchange and interest rates.

	Redundant groups of words (62)	The Serbs say that Kosovo lay at the heart of its medieval kingdoms and that during the middle Ages, so they will not leave here whatever happens to them.
	Missing out necessary groups of words (38)	GE can be used to increase the crop yield so that [we can solve] solving some of the world hunger problems.
Tense + aspect 198 (4.4%)	Misselection between present and past tenses (83)	Over the centuries and particularly during the decades of the past century, plant breeding is [was] used more precisely.
	Misselection between simple past and present perfect (59)	Last year, several of the largest airline companies have laid off [laid off] 127,000 employees.
	Misselection between simple and perfect aspects (46)	Since the successful research on crops and animals, more and more scientists change [have changed] their study *(Sic.)* to the research of human genetic engineering.
Modal 185 (4.1%)	Missing 'will' (67)	Joining the euro will enable businesses to sell more products and gain greater economies of scale. Also, it [will] enable families and businesses to buy…
	Misselection between 'will' and 'would' (42)	…we will find that the developing *(Sic.)* of genetic engineering would [will] bring us a great future. There are various reasons you might want to clone a human. It would allow an infertile couple to have a child… To replicate the talents of exceptional human beings seems to be a controversial issue. It will [would] be an *(Sic.)* amazing to listen to Einstein explain his "principle of relativity" personally.
	Missing 'would' (28)	David Blunkett, the Home Secretary, suggested that national compulsory identity cards, which is [would be] called "citizen entitlement cards", could be introduced as part of the anti-terrorism measures.
Conjunction 175 (3.9%)	Missing 'and' (55)	Some serious diseases like asthmas *(Sic.)*, cancer [and] even AIDS could be possibly cured before we are even born.
	Sentence fragment (27)	Because it seems illogical to allow the creation of a cloned human embryo and produce a cloned baby. They think cloning will create a new human being to destroy our society.
Pronoun 174 (3.9%)	Relative pronoun error (36)	, but poor people's diets often lack fat and other key nutrients and so the GE rice [that/which] contains pro-Vitamin A will not benefit them.
	Misselection of 'it' for 'this' (35)	Some people argue that the welfare state has become the supporter of new family relationship *(Sic.)*. I intend to discuss it [this] by presenting *(Sic.)* evolution of Britain.
	Non-agreement between Pronoun and referred noun (26)	They think people here need to carry no papers and do not have to inform the authorities of what you [they] are doing or get permission for anything.
Adjective 108 (2.4%)	Adjective for adverb (36)	'Golden Rice" is one of these genetic [genetically] engineering products.
	Adjective for noun (29)	Unlike any chemical treatment, it is harmless to *(Sic.)* environmental [environment] and people's health.
Gerund 104 (2.4%)	Gerund for noun (45)	The last one is the wild using [use] of insecticides and chemical fertilizer,
	Gerund for infinitive (33)	The police felt that there was an urgent need of introducing [to introduce] *(Sic.)* identity card system in respect of immigration control.

Auxiliary 77 (1.7%)	Missing auxiliary 'be' (34)	The euro currency will incorporate specific characteristics that allow blind people or the visually impaired to distinguish between the different euro note *(Sic.)* and coins. These will [be] <u>designed</u> to help disabled persons to adjust to the single currency.
	S-AUX Verb non-agreement (18)	They think that this technology <u>have</u> {auha # ag G} been released without adequate knowledge about their *(Sic.)* effects…
Adverb 77 (1.7%)	Misordering of adverb (25)	…some of them <u>even do not</u> [do not even] know how much information the government has.
	Adverb for adjective (12)	Her concern is that the question "should we have ID cards" is deceptively <u>simply</u> [simple]. (S44)
Infinitive 42 (0.9%)	Missing 'to' (24)	The UK Company does not <u>need [to] worry</u> about the devaluing *(Sic.)* of the other trading country.
Voice 36 (0.8%)	Active voice for passive voice (27)	In the 1970s new forms of communication mediated by computers began to <u>use</u> [be used] as well.
Extential 27 (0.6%)	Redundant 'there+be' (18)	But if <u>there were</u> no one tries to experiment there would *(Sic.)* be no development of this science.
Negative 7 (0.2%)		It is therefore <u>no</u> [not] possible for them to change the real exchange rate by changing the nominal rate.
Order 1 (0.0%)		…three concepts of *(Sic.)* genetic engineering: first GM Food, second Clone *(Sic.)* and [third] Medical research.
Total 4493 (100%)		

Table 2: The analysis of grammatical errors

Linguistic feature No of errors (Frequency %)	**Salient errors/ (No. of errors)**	**Erroneous instances**
Incorrect association of a preposition with a noun, a verb or an adjective 137 (52.3%)	Noun-related preposition error (66)	It seems that xenotransplantation is the best solution <u>of</u> [to] this problem,…
	Verb-related preposition error (59)	Different government *(Sic.)* has to <u>think</u> [think of] different solution *(Sic.).*
	Adjective-related preposition error (12)	…that would make Europe itself more compatible <u>to</u> [with] the world's two powers- the U.S.A and Japan.
Countability of the noun 52 (19.8%)	Incorrect form of noncount noun (adding –s or 'a/an') (48)	People have to waste *(Sic.)* enormous amount of times [time] on congestion.
Incorrect syntactic pattern of a word (a noun, a verb, an adjective, etc.) 39 (14.9%)	Verb-related syntactic pattern (25)	<u>It is accused that the United States</u> [The United States is accused of] interfered the world's economies.
Transitivity pattern of the verb 34 (13.0%)	Redundant preposition (29)	We have <u>considered about</u> [consider] the second situation.
Total 262 (100%)		

Table 3: The analysis of lexical-grammatical errors

Linguistic category/ No. of errors (Frequency %)	**Salient errors/ No. of errors**	**Erroneous instances**
Lexical misconception 301 (63.1%)	Misuse of a lexical item for another (301)	If we use GM animal *(Sic.)* to produce human apparatus [organs] and blood, … By the early 196Os, most British colonies had acquired independence, but the *(Sic.)* racism has not decayed [decreased].
Collocation 110 (23.1%)	Incorrect word in a collocational unit (66)	Because the operation for nuclear transfer is extremely hard, so *(Sic.)* it could become wrong [go wrong] for several reasons. …people have more money to spend on goods and services, *(Sic.)* as a return [result] there will be a high standard of living.
	Missing word in a collocational unit (32)	[A] Large number of modified animals are suffering as laboratory tools etc. On [the] one hand, doctors and scientists warn that these foods are not safe in the human diet.
Misspelling 64 (13.4%)	Words with similar sounds or shapes (e.g. alone/along, serious/series, there/their) (54)	They do not make *(Sic.)* contribution to the state and even course [cause] the lack of labor resource. People argue cloning human *(Sic.)* would bring *(Sic.)* ethnical [ethical] problem *(Sic.)*.
Non-existent words 2 (0.4%)		…in the mid-1990s the government of Britain devaluated [devalued] the pound successfully…
Total 477 (100%)		

Table 4: The analysis of the lexical errors

Surface deviance/ No. of errors/ Frequency (%)	**Salient errors**
Misselection 2514 (48.1%)	• Bare singular noun for plural (458) • Misselection between prepositions (321) • Lexical misconception (301) • Comma splice (103) • Misselection between tenses (83) • Noun for adjective (83) • Incorrect choice of groups of words (76) • Misselection between verb base form, past participle and present participle (74) • Incorrect word in a collocational unit (66) • Misselection between modals (64) • Misselection between 'the' and 'a/an' (64) • Misselection between simple past and present perfect (59) • Misselection between aspects (46) • Gerund for noun (45) • Misselection between 'will' and 'would' (42) • Adjective for adverb (36) • Misselection of 'it' for 'this' (35) • Verb for gerund (35) • Gerund for infinitive (33) • Adjective for noun (29) • Active voice for passive voice (27)
Omission 1294 (24.7%)	• Missing definite article (529) • Missing preposition (115) • Missing 'a'/'an'(104) • Missing modal (101) • Missing conjunction (86) • Missing auxiliary 'be' (34) • Missing necessary groups of words (34) • Missing word in collocational unit (32) • Missing copula (31) • Sentence fragment (27) • Missing 'to' (24)
Overinclusion 931 (17.8%)	• Redundant definite article (446) • Redundant comma (87) • Redundant preposition (63) • Redundant groups of words (62) • Redundant preposition in transitive verb (29) • Redundant copula (27)
Misformation 431 (8.2%)	• S-V non-agreement (125) • Quantifier/Determiner-noun non-agreement in number (81) • Misspelling (64) • Incorrect form of noncount noun (e.g. adding –s or 'a/an') (48) • Pronoun-referred noun non-agreement (26)
Misordering 62 (1.2%)	• Misordering of adverb (25)
Total 5232 (100%)	

Table 5: The analysis of surface structural deviances