The Role of Frequency in ELT: New Corpus Evidence Brings a Re-appraisal

Geoffrey Leech, Lancaster University

1. Why is frequency important?

My subject in this paper is the role of **frequency** in helping to determine teaching priorities in English language teaching. On the one hand, it seems to be a matter of common sense to teach words or forms which are frequent before those which are infrequent or rare. On the other hand, I feel that over the past generation the topic of frequency has been neglected in the teaching of languages, although it has started to reclaim attention in the last few years. There are also problems, both of theory and practice, relating to frequency.

First, what is the point of frequency? Why is it valuable, in particular, for the language teacher? I claim that it is valuable to build frequency considerations into one's curriculum, one's syllabus, one's teaching materials, and one's classroom teaching. If an item naturally occurs frequently in the language being taught, it is likely to be important also for the target behaviour of the learner: the learner will later often come across that item in reading and listening, and will often need to use it in communicating with others. And yet, frequency has been largely ignored, for three reasons.

The first reason is that until recently, knowledge of the frequency of items in a language has been very limited. To consider *why*, we need to ask: How do we find out about frequency? Information about frequencies of words, expressions, and grammatical structures can be gained from a large sample of texts, i.e. a **corpus**, of the language concerned, and of course the computer is indispensable to this work, which may involve sifting through tens or hundreds of millions of words. Such corpora of language data having been increasingly compiled over the past 30 years, but are only now becoming seriously applied to pedagogical purposes. But the breakthrough *is* being made, particularly in dictionaries. The major English-language dictionaries for advanced learners, such as the *Oxford Advanced Learners' Dictionary*, the *Collins Cobuild Dictionary,* and especially the *Longman Dictionary of Contemporary English* (*LDOCE*), now take account of frequency information about items of vocabulary. For example, the senses of words are placed in order of frequency, and the American English edition of *LDOCE* (*Longman Advanced American Dictionary,* 2000) provides little ‘frequency boxes’ alongside important words, giving their frequency rating in spoken and in written English.

return (verb)			**return (noun)**	
S	**W**		**S**	**W**
	1			**1**
2				
			3	

Figure 1

As an example, the boxes in Figure 1 inform us that *return* and a verb and *return* as a noun are both very frequent in written English ('1' means that they are in the top one thousand words), but are not quite so frequent in speech ('2' = in the top two thousand words, and '3' = in the top three thousand words). The same dictionary provides occasional bar charts, contrasting (for example) the different frequencies in American English and British English of the near-synonyms *rubbish*, *garbage* and *trash*. This kind of information is now making an impact in lexicography because publishers have invested a great deal of time, effort and money in building and using such large electronic text corpora of both spoken and written language. So useful knowledge about frequency is now at last becoming available. To give some recently available frequency data on general English, I will make reference in this paper to two books:

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E., *Longman Grammar of Spoken and Written English.* London: Longman 1999. (henceforth *LGSWE*)
Leech, G., Rayson, P. and Wilson, A. (2001), *Word Frequencies in Written and Spoken English, based on the British National Corpus*. London: Longman 2001. (henceforth *WFWSE*)

(The former of these books gives information on grammatical frequency, and the latter gives information on word or lexical frequency.)

The second reason for the neglect of frequency is that specialists in applied linguistics have not given much attention to it since the 1950s. Fifty years ago, frequency was quite a popular topic with leaders of opinion in ELT. People like Michael West, who compiled the *General Service List of English Words* (Longman, 1953), spent years, with teams of helpers, counting the frequency of words in many texts. That was before the age of computers: so, the work of obtaining frequency information *by hand* was extremely time-consuming and boring, and moreover, since there were no tape recorders in those days, it was restricted to written language. So this work was of limited application, and applied linguists have since then given more attention to more interesting topics, like how do people learn languages. The focus turned to the processes and techniques of learning and teaching, rather than course content. It is now instructive to look at the most influential textbooks on applied linguistics over the past 30 years, such as Rod Ellis's *The Study of Second Language Acquisition* (1994), and to notice how little attention is given to frequency, and how little enthusiasm is shown for it. Ellis wrote:

> Overall, there is little evidence to support the claim that input frequency affects L2 acquisition, but there is also little evidence to refute it. Perhaps the safest conclusion is that input frequency serves as *one* of the factors influencing development, often combining with other factors such as L1 transfer and communicative need. (ibid. 272-3)

This is one of the very few passages in that long and highly informative book where Ellis discusses frequency. But looking closely, we see that Ellis is discussing *input frequency* - the frequency with which learners are exposed to language items in the classroom - rather than frequency in the language in general use. He is attending to frequency as an *input* to learning, whereas I want to focus on frequency as a factor steering the *outcome*, assuming that the ultimate goal of learning is to obtain a communicative competence in the language.

I found one other general textbook which gives more attention to this subject: van Els et al *Applied Linguistics and the Learning and Teaching of Foreign Languages[[.* In discussing the selection and gradation of course content, these authors mention frequency in language use as the first consideration, in determining what should be taught and when, for example in selecting vocabulary. But in addition they mentioned other criteria, such as:

1. Range or dispersion
2. Coverage
3. Learnability
4. Communicative need

2. Difficulties and competing factors

Here we come to the third reason for the neglect of frequency: it is actually not such a straightforward idea, because there are difficulties in applying it, both in principle and in practice. This will emerge during the discussion of the above four criteria. In what follows, I will discuss these criteria concentrating initially on vocabulary selection, as the easiest case, and will later give more attention to frequency of grammatical phenomena.

a. Range or dispersion (from now on I will use the single term ‘dispersion’) means how well the item is distributed throughout the use of the language, for example in different texts and text types. To study this objectively, we have to return to the idea of a sample corpus of texts – and bear in mind that a ‘texts’ in this sense include both written texts and transcriptions of speech. Thus in the British National Corpus (BNC), one of the major corpora that can be used for frequency studies on English and the corpus on which *WFWSE* is based, the noun *influence* occurs with the same frequency as the noun *software*, but *software* has a lower dispersion, i.e. is less well distributed throughout the corpus. So by that criterion, *software* is a less useful word for learners in general, although it may be particularly useful for learners of English for computing and technology.[1] (For those interested to know how the distributional spread of a word in a corpus is measured, the easiest measure to use is **range**, which simply means that the corpus is randomly divided into (say) 100 equal parts, to find out how many of them contain the word in question. **Dispersion** is a more sensitive measure, based on a statistical formula known as Juilland’s D.[2]) Hence frequency and dispersion can be judiciously combined to give

[1] This example incidentally provides a warning that corpora may go out of date rather quickly. The BNC was compiled in the early 1990s. It is quite possible that this result would not be found in a corpus collected today, when *software* has become more of an everyday word.

[2] The formula is given in *WFWSE*, p.18. See also Lyne (1985).

a measure of what vocabulary is more central to the language (‘core vocabulary’) and what vocabulary is more peripheral or specialized.

b. Coverage : This is another measure of what might be called 'coreness' of vocabulary: words with wide **coverage** are more useful to the learner than words with narrow coverage. We can distinguish two types of coverage: **coverage of meaning** and **coverage of register** or style. Coverage of meaning can be illustrated by the two verbs *give* and *donate*. *Give* is a word of wider semantic coverage than its partial synonym *donate*, and we can check on this by looking up these words in the dictionary, and noting how many different senses *give* has, compared with *donate*. **Coverage of register or style** overlaps with dispersion, and refers to the extent to which a word is likely to occur in different varieties of the language. For example, the adjective *nice* is over 8 times as frequent in speech as in writing. This measure suggests that the word *nice,* although extremely useful in speech, is far less useful in writing - a factor we might want to take into account in designing core vocabularies for teaching purposes. The opposite is true of *thus*, which is more than 20 times more common in writing than in speech. We can contrast these words, which we can consider colloquial and formal words respectively, with a word like *came*, which is approximately equally common in both speech and writing, and in that sense has a more balanced stylistic coverage than *nice* and *thus*.

In addition to these more or less objective factors, van Els *et al* mention psychological and didactic criteria, especially the criterion of **learnability**, which we now briefly consider.

c. Learnability. No doubt some words are more 'learnable' than others, i.e. for one reason or another students will find them easier to learn. One reason may be that the word has irregular forms: e.g. the noun *corpus* I have used here often occurs with the rare Latin plural *corpora*, which makes is more difficult to learn in this respect than most English nouns. Other factors of difficulty for the learner include cognitive complexity, which can be more easily illustrated in the grammatical sphere. For example, psycholinguistic studies have shown that passive constructions are more difficult to process than active constructions, and that negative constructions are more difficult to process than positive ones (see Clark and Clark 1977: 105, 240-1; also Wason 1962). This is not surprising, and no teacher would dream of teaching passive sentences before active ones, or negative sentences before positive ones.

d. Communicative need. For many teachers, this will be considered the overriding criterion of selection, although it is somewhat difficult to determine. Whereas in the earlier stages of learning, communicative need will be governed by the developing requirements of the curriculum, in a longer perspective it will be determined by the general goals of language learning for speaking and listening, writing and reading, with needs analysis yielding different priorities for different categories of students, such as those learning English for Academic Purposes or English for Specific Purposes.

From this list of factors - frequency, range, coverage and learnability - it appears that high frequency is just one of the variables that lead to the prioritization of an item in the language learning process. But an important thing to notice is that all of the other factors are strongly associated or correlated with frequency. Consider dispersion: my work with *WFWSE* has shown me that it is in fact quite difficult to find items in a frequency list where greater frequency is not significantly associated with a greater dispersion. But there are counterexamples. One counterexample I found is the pair of nouns *answer* and *animal*:

	Frequency per million	*Dispersion index*
answer (n.)	124	0.93
animal (n.)	153	0.90

The explanation of this case seems to be that *answer* is a more generally employed abstract noun, whereas *animal*, as a concrete noun, although more common, is topic-related, and therefore more unevenly distributed. In general, nouns are more topic-related than other parts of speech, and accordingly have a lower dispersion than their frequency might lead one to expect.

Next, consider coverage: here is a small list of verbs of more general coverage (in register and/or meaning) matched with partial synonyms of more restricted coverage. It is obvious that the general-coverage verbs are very much more frequent than the more restricted (and more formal) verbs.

give	1284 per million	*donate*	10 per million
want	945 per million	*desire*	14 per million
build	230 per million	*erect*	15 per million
hide	64 per million	*conceal*	17 per million

As for learnability, if we associate one kind of learning difficulty with morphological complexity of words, there is a well-known law, Zipf's law or principle of least effort (Zipf 1935, 1949) which states among other things that the more complex a word, the less frequent it will be. This intuitively obvious point is confirmed by the following BNC data on complexity (in number of syllables) and frequency from *WFWSE*:

Most common 1-syllable word:	*the*	(61847 per million)
Most common 2-syllable word:	*into*	(1634 per million)
Most common 3-syllable word:	*government*	(622 per million)
Most common 4-syllable word:	*information*	(386 per million)
Most common 5-syllable word:	*international*	(221 per million)
Most common 6-syllable word:	*responsibility*	(93 per million)

It is clear that in this purely formal sense frequency and learnability correlate. On the level of syntax, consider again passives. Passive verb phrases are far less frequent than active ones: the highest percentage of passives is found in academic writing, where they amount to over 20% of all verbs. The figure in conversation is as low as 2%. If we pursue the theme of passives one step further, it is interesting to observe that passives with an agent, so called ‘long passives’, as in *He was surrounded by a ring of men*, are six times less frequent than passives without an agent, or ‘short passives’, as in *Smith was jailed.* The ‘short passive’ without agent is structurally simpler, and this would lead one to suppose that it is easier to learn how to use than the ‘long passive’ with agent. This suggests that, contrary to what is often assumed, the short passive should be given teaching priority over the long passive, on grounds of both frequency and learnability.

We have been looking at only the form of words and structures, but it is clearly not the whole story. We need to consider the meaning and use of words: a frequent word like *give* is not easy to learn in all its senses, but at least *give* in its basic sense should be introduced early.

Leaving this issue aside, I have tried to justify a *prima facie* important role for frequency in determining teaching priorities. One practical point is that with the growing availability of large and varied corpora, frequency has an advantage of convenience over other yardsticks of usefulness: it is easily measurable. Moreover, I would argue that there is an essential link between serving the future communicative needs of the learner - presumably the objective the learner has in learning the language - and the frequency of items in a well-chosen sample corpus.

But here I want to add some points which urge caution in taking the argument for frequency too far.

First, there is a considerable practical difficulty - how can we find a corpus which is ‘well-chosen’ in that it matches the learner's communicative needs? Nowadays, very large and varied corpora, such as the Bank of English and the BNC, are available and can be used all over the world. However, size and variety are not everything. Optimally we also need *targeted* corpora - corpora targeted to represent as closely as possible the learner's future communicative needs. But here there are a number of impediments which I will merely mention briefly:
(a) A representative corpus needs to be large and should contain balanced samples of a wide range of texts and transcribed speech. Apart from this, though, the concept of a 'representative corpus' is not well defined, and has been a subject of controversy in corpus linguistics.[3]
(b) Not only does a representative corpus need to cover a wide range of language varieties, but it needs to be useful for different kinds of learners. Hence frequencies have to be extracted for different varieties - e.g spoken and written registers or genres such as conversation, academic lectures, scientific language and business language.
(c) There is also a need for corpora for learners of different levels of maturity and attainment. For advanced learners, a corpus, such as the BNC, of native speaker speech and professionally-competent writing may be adequate. But for intermediate or lower students, it is important to have a corpus which represents a reasonable target performance for those students at their existing level – for example, *The*

[3] The best-known treatment of this issue of corpus ‘representativeness’ is that of Biber (1993).

American Heritage Frequency Book (Carroll et al 1971) was based on a large corpus of reading materials for American primary school and high school pupils. (Unfortunately this corpus was not computerized and cannot be used; however, we may look forward to other corpora of this kind becoming available[4]). Or perhaps we need a corpus combining both unedited native speaker (NS) texts and pedagogical materials.

(d) Even a well-chosen corpus of NS English may be less than ideal. It will increasingly be argued that a corpus of international English, including English of competent non-native speakers, is required to meet the needs of students in the twenty-first century, when English, as a ‘global language’, can no longer be regarded as under the exclusive influence of native speakers, since more and more communication takes place between competent non-native speakers of different language backgrounds.

So there is, and can be, no ‘ideal corpus’ for ELT. In fact, there is a good case for arguing that for selection and grading of course materials, we need to consult a range of different **reference corpora**, including NS corpora, international corpora, and corpora tailored to different learning conditions. Meanwhile, the corpora that we have already provide a good starting point, with much useful frequency information.

In spite of these cautionary observations, I hope I have succeeded in making the case for a new and positive role for frequency in ELT. Although we still have a long way to go, good progress has now been made along the path to obtaining the frequency information we need.

3. Examples of frequency in vocabulary and grammar

In this next section of this paper, I want to present a range of examples of how knowledge of frequency can be helpful to the world of ELT. In this section I will concentrate on grammatical frequency, drawing illustrative findings from *LGSWE* to show how knowledge of grammatical frequencies may cause some reconsideration of the priorities and assumptions often found in ELT materials for the teaching of grammar. (Previous grammars, including some of which I myself have been an author, will also need reconsideration in the light of these findings.) At this juncture I will point out that the *LGSWE* was based on a special corpus of 40 million words of American and British English, including a core corpus of 20 million words taken from four registers of the language: conversation (Conv), fiction writing (Fict), newspaper writing (News) and academic prose (Acad). Much of the analysis was devoted to a comparison of differences in frequency, often striking, between these four varieties of spoken and written English.

[4] For example, an M.A. student I supervised at Lancaster recently, Kaori Shinohara, was able to obtain and make use of a large electronic corpus of authorized English language learning materials published for high school use in Japan.

a. Multiword verbs

Treatments of grammar usually distinguish between **phrasal verbs, prepositional verbs,** and **phrasal prepositional verbs**. Priority is typically given to phrasal verbs (e.g. *pick up*) which are assumed to be the most important type, perhaps because of the word order difficulties they can present:

> ***Pick** the phone **up** – **pick up** the phone – **pick** it **up** - *pick up it.*

However, in fact prepositional verbs (e.g. *look at*) are much more frequent than phrasal verbs, which in turn are much more frequent than phrasal prepositional verbs (e.g. *look forward to*). Moreover, this order of frequency is the same for all four registers, which is surprising, since phrasal verbs are often considered to be more characteristic of informal than formal English. (Actually, all three types are more common in fiction writing than in conversation.) This finding may cause some rethinking: perhaps prepositional verbs should be introduced to the student before phrasal verbs, being both more frequent and easier to handle.

b. Modal auxiliaries

This small but very important class of auxiliary verbs is often treated as a class, with modals such as *must* and *may* being introduced alongside *can* and *will*. The bar chart in = 2 shows the marked differences in frequency of the modals, with the two pairs of modals *will* and *would* and *can* and *could* being clearly more common than the others.

[[Figure 2: (=Figure 6.8, p.486 of *LGSWE*)]][5]

(Marginal modals such as *ought to* and *dare* are too infrequent to show up on the chart.) The lower-frequency modals *may, must, shall, ought to, need* (+ bare infinitive) and *dare* (+ bare infinitive) are in fact growing more infrequent, especially in American English. Certain of their meanings, such as *may* in the sense of permission and *must* in the sense of obligation, are becoming particularly infrequent. If we examine conversation alone, modals are overall particularly frequent in that variety, and roughly the same order of frequency is found, except that *will* and *can* (including contracted forms such as *'ll* and *can't*) are at the top of the frequency list, and the decline in frequency separating the frequent and less frequent members of the class is more marked – a common pattern in conversational frequencies. These findings may again lead to a reconsideration of priorities. In particular, little time should be spent on rare modals *ought to, shall* and *need,* and on rare usages such as *may* in the sense of 'permission' – a usage now almost absent from conversation.

c. Frequencies in conversation and written language

The following two charts illustrate a common scale-like trend in the frequency differences between conversation, fiction, news and academic prose.

[5] I am grateful to Foreign Language Teaching and Research Press, Beijing, and to Pearson Educational, the publishers of Longman books, for permission to reprint this and subsequent figures from *LGSWE*.

[[Figure 3 (Fig. 4.8, p. 333 of *LGSWE*) and
Figure 4 (Fig. 4.5, p.291 of *LGSWE*) side by side]]

We see from Figure 3 that pronouns, of which by far the most frequent category is personal pronouns, are much more frequent in conversation than in the written registers, and that there is an overall 'stepping down' pattern leading from conversation at one extreme to academic writing at the other. Of the intermediate categories, fiction writing is closer to conversation and news writing is closer to academic prose. The opposite pattern, a 'stepping up' pattern, is seen in the frequency of nouns in Figure 4. A major reason for this highly distinctive pattern is that conversation takes place in situations where speaker and hearer have a shared knowledge of context, whereas the writers of the expository styles of news and academic prose use full noun phrases, often of considerable complexity, to express situation-free informative content. This complementarity of pronouns and nouns is not greatly surprising in itself, but I am using it to illustrate the immense differences of frequency which distinguish prototypical types of spoken and written language. (Incidentally, the 'stepping up' and 'stepping down' profiles illustrated in Figures 3 and 4 are a characteristic pattern of frequency between the four registers, but other patterns also occur.) Although it is sometimes suggested that spoken and written grammar are different systems, my position is that the grammatical system is largely the same for both, but that differences show up markedly in frequency. Now that more attention is being given to learning the spoken language, it is important to bear in mind that the teaching of grammar, which has traditionally be based on the written language, needs to be considerably adapted to accommodate the different learning priorities of spoken language.

d. Verb constructions

The 'stepping down' patterns in Figures 5 and 6 show that finite verbs as a whole are more frequent in conversation than in the written registers. However, I want to use these bar charts to illustrate the overwhelming dominance, in terms of frequency, of the simple aspect, that is, the present simple, the past simple, and the simple modal construction, over forms containing the progressive (or continuous), perfect, and passive constructions.

[[Figure 5 (Fig. 6.2, p.461, *LGSWE*) and
Figure 6 (Fig. 6.7, p.476, *LGSWE*) side by side]]

The perfect aspect and the progressive aspect generally receive a great deal of attention in teaching English grammar, and no doubt rightly, because of their unfamiliarity and difficulty for speakers of many other languages. But Figure 5 below puts things in an unexpected perspective. It shows that perfect and the progressive forms are a surprisingly small proportion – only a few per cent – of all finite verbs.[6] Perhaps the perfect and the progressive should not be allowed to loom too large in the syllabus after all.[7]

Figure 6 shows the frequency of the passive voice compared with the non-passive (or active) voice – including active verbs which have no passive counterpart. While the pattern of Figure 6 looks similar to that of Figure 5, the trend for the passive is the opposite of the trend for the progressive aspect: whereas the progressive is most frequent in conversation, and gradually diminishes in frequency as we move towards academic prose, the passive is most frequent in academic prose, and progressively decreases in frequency as we move towards conversation. In fact the frequency of the passive is about 10 times greater in academic writing than in conversation. As we have seen, the short passive is much more common than the long passive in English, so in academic writing this means that the agent (often the human investigator) is omitted, and the thematic focus is instead placed on the object of investigation. This suits the objective purpose of scientific investigation and explanation. Conversation, on the other hand, shows a human-centred concern for people's actions, thoughts, and feelings, and the human actor (often the speaker) typically takes a thematic position as subject. Here the passive voice is rarely needed. The message from an ELT perspective is that the passive needs to be given very different priorities in different registers of English. This may not be a new insight for many teachers, but the dramatic contrasts in frequency present it with stark clarity.

e. The ordering of adverbials

My last illustration of grammatical frequency concerns the ordering of different classes of adverbials at the end of a clause, focusing on the most common classes of adverbials, those of manner, place and time. In the teaching of grammar, it has often been stated that the normal ordering of these adverbials is M – P – T, that is, manner before place before time. Although this pedagogical rule of thumb has been stated many times, as far as I know it has not been tested again the real evidence of language use until recently. One of the authors of *LGSWE*, Susan Conrad, tested it against the evidence of the Longman corpus, and the results are shown in Figures 7-9:

[[Figure 7 (Fig. 10.16, p. 811 of *LGSWE*),
Figure 8 (Fig. 10.14, p.811 of *LGSWE*) and
Figure 9 (Fig. 10.15, p.811 of *LGSWE*) – two of these can be side by side]]

[6] It might be wondered what has happened to the combination of perfect and progressive aspects (e.g. *What have you been doing?*) in Figure 5. Perhaps surprisingly, the perfect progressive construction is too rare in the corpus to show up on the bar chart.

[7] It is clear that frequency in the target language and difficulty for the learner are independent factors. Yet some evidence suggesting the need for more attention to the simple present and past forms of the verb in teaching is provided by the work of Granger and her associates on learner corpora – see Note 7 below. Granger (1999) analyses tense and aspect errors of French-speaking undergraduate learners of English, and shows that the greatest number of these errors is found with the misuse of the present and past simple. These outnumber errors involving progressive and perfect forms by 61.5% to 38.5%.

The first thing to note is that these three clause-final adverbials rarely occur all together in a single sequence. Consequently, each pair of co-occurring adverbials was examined separately: manner and place, manner and time, and place and time. In each figure, the lower and darker part of each bar represents instances which conform to the M – P – T rule. The upper and lighter part represents instances which break the rule. As the Figures show, the rule is at best a probabilistic tendency. It is true that the rule is upheld more frequently than it is broken, but overall there are more than 30 per cent of cases which go against the rule. Many exceptions to the rule can be explained by supplementary rules which apply generally to the ordering of elements in English grammar: the principle of **end-weight**, and the principle of **information** (or end-focus). The end-weight principle favours the placement of shorter adverbials before longer and more complex ones. The end-focus principle favours the placement of adverbials which express given information before those which express new or focused information. These principles are sometimes powerful enough to outweigh the M – P – T rule, which in retrospect should not be called a 'rule' at all, but a tendency or constraint. It is sometimes useful to give learners 90% rules, or even 70% rules, in the earlier stages of teaching grammar. But the evidence suggests that the M – P – T 'rule' is of limited value.

4. Conclusion

I hope to have shown that frequency information, in the fields of both grammar and lexis, can bring a realistic re-appraisal of what English language content is taught to different kinds and levels of learners in the interests of their communicative needs.

I want to conclude by acknowledging that my focus on NS frequency patterns of the target language, English, suffers from an important limitation. In addition, there is need to investigate frequency patterns in the L1, Chinese, and also in the language performance of Chinese learners themselves. This involves another kind of corpus-based investigation involving what Sylviane Granger has called 'Contrastive Interlanguage Analysis' and which can best be done in the L1 country, China.[8] An essential tool for this is a corpus of learner English, which is the topic of the next paper at this Conference, that of Gui Shichun.

What I earlier called 'learnability' is of course partly determined by L1 influence, which can be studied through Chinese-English contrastive analysis. Its influence on learners' English can be further studied by building a Chinese learner English corpus, and observing patterns of frequency in that corpus. These include patterns of error derived from an error-tagging of the corpus, but they also include patterns of frequency in the learners' English productions generally - which can then be compared with frequencies in appropriate NS corpora. The resulting comparison identifies patterns of **overuse** – where

[8] Granger (1998) provides a general picture of research on learner corpora, with particular reference to the International Corpus of Learner English (ICLE), of which she is the founder and co-ordinator. Granger and her colleagues have also pioneered the error-tagging of learner corpora, which enables teachers and researchers to pinpoint areas of difficulty and (also importantly) areas of non-difficulty for the students in the target language – see, for example, Granger (1999). This is a fast-growing area of research worldwide.

the learners use a feature of English more frequently than the native speakers – and patterns of **underuse** – where they use a feature less frequently than native speakers. These kinds of evidence help to give a more rounded picture of the interlanguage than the purely negative evidence of error analysis. For example, overuse may be due to a tendency to prefer features of English which are closer to the L1 (L1 transfer), and underuse may be due to avoidance strategies, where the learner steers clear of features which are unfamiliar or difficult in the target language. I actually find the terms ‘overuse’ and ‘underuse’ rather unhelpful, because of their implication that the NS’s English is the sole standard for judging the English of Chinese learners, but it is difficult to find suitable alternatives. The important point is that corpus-based interlanguage analysis enables us to identify areas of difficulty which are not derivable from NS corpora alone, and which can often be attributed to particular causes, especially L1 transfer. I therefore regard Gui Shichun’s paper as an important, and indeed necessary complement to my own.

References

Biber, D. (1993), ‘Representativeness in corpus design’, *Literary and Linguistic Computing* 8: 243-257.

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

Carroll, J. B., Davies, P., and Richman, B. (1971), *The American Heritage Word Frequency Book.* Boston MA: Houghton Mifflin.

Clark, H. H., and Clark, E.V. (1977), *Psychology and Language: An Introduction to Psycholinguistics*. New York: Harcourt Brace Jovanovich

Ellis, R. (1994), *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

van Els, T., Bongaerts, T., Extra, G., van Os, C. and Jansen-van Dieten, A-M. (1984), *Applied Linguistics and the Learning and Teaching of Foreign Languages*. London: Edward Arnold.

Granger, S. (ed.) (1998), *Learner English on Computer*. London: Longman.

Granger, S. (1999), ‘Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus.’ In H. Hasselgård and S. Oksefjell, *Out of Corpora: Studies in Honour of Stig Johansson,* Amsterdam: Rodopi, 191-202.

Leech, G., Rayson, P. and Wilson, A. (2001), *Word Frequencies in Written and Spoken English, based on the British National Corpus*. London: Longman.

Lyne, A. (1985), *The Vocabulary of French Business Correspondence*. Geneva: Slatkine.

Wason, P. C. (1962), *Psychological Aspects of Negation*. Communication Research Centre Papers 2. London: University College London.

West, M. (1953), *A General Service List of English Words.* London: Longman

Zipf, G. (1935), *The Psycho-biology of Language*. New York: Houghton Mifflin.

Zipf, G. (1949), *Human Behavior and the Principle of Least Effort*. Reading: Addison-Wesley.