Corpus Linguistics and Foreign Language Teaching

语料库语言学与外语教学

Conference Proceedings

2003 International Conference on Corpus Linguistics

Shanghai Jiao Tong University
Shanghai, P.R.China
October 25-27, 2003

Organizers:

Professor YANG Huizhong	(Shanghai Jiao Tong University)
Professor ZHUANG Zhixiang	(Shanghai Foreign Language Education Press)
Professor GUI Shichun	(Guangdong University of Foreign Studies)
Professor WANG Tongshun	(Shanghai Jiao Tong University)
Professor WEI Naixing	(Shanghai Jiao Tong University)

Table of Contents

John Sinclair's talk given at the International Conference on Corpus Linguistics, Shanghai Jiao Tong University, Shanghai, China, October 25-27, 2003

Distinguished colleagues:

It's a great pleasure for me to return to Shanghai and to Jiao Tong University. And it's not coincidence that this is the latest of several visits over the last twenty years. And because, as several speakers have said already, the work that was done here twenty years ago and more, was of a highly pioneering nature and it opened up, not only as far as I know the first corpus work in Asia, but also it was a pioneering step in the development of a specialized corpus, the first specialized corpus in the language as far as I know anywhere. And was put together by the efforts of Professor Yang and his colleagues at a time when there was no such thing really is corpus linguistics. It was more or less unknown. And indeed I first time came to Shanghai on their business, and met Yang and his colleagues at that time after since when we've had a long and fruitful association. So I'm very pleased indeed to be back here, and to see what a large gathering this is, and what a very health state corpus linguistics is now in China. And I'm sure that it's the lots efforts of people in this room who have made it possible. So I've been given an impossible task this morning, and, I was just asked to do my best. I did the silly thing you should never do if you're invited to give a talk, which is to say, would you like me to talk about a, b, or c, which was what I said to Yang, and his answer was yes please or three. So I have to talk very quickly. But I'll be able to move some this, this work to a workshop that will take place tomorrow which does relieve me of some of the things that I would otherwise have to say.

But essentially, the three topics I want to touch on are, the development of corpus linguistics and its relation to other types of language work on computers, and a little bit about the theoretical side as very proud to here Yang, introducing the matter of theory as an important need in corpus work in the future. I talked to review that. And I should be, try to outline, the possibilities later on.

And then I do want to give you, just a little example of corpus work in action. So, I'm, you have a handout which includes everything you'll see on the screen. It's simply print out everything on the screen, and so, if I skip or rather fast through some sections, then you will be able to find them on the handout, and I'm going to be here the whole of the conference end up to if you want to raise some questions and discuss it, I will be very pleased to. This has to be a very brief and very brisk round through, 40 years of the development, and I, I hope I might do it injustice to anybody in it.

But essentially, corpus linguistics, I would like to say in the beginning it is not a branch of computational linguistics. It has quite separate and distinct origins. And although in recent years they are tended to come together, their origins are quite different and their underlying philosophies are quite different. Because computational linguistics, originated in the early theories of Chomsky, which at that time looked as if they would work very well. In computer, they look like algorithms and this is a very common perceptions in the early 1960s, some 40 years ago.

And whereas the impetus to corpus linguistics came from the other end of the spectrum of linguistics, the traditions of field work, the traditions of gathering data, and studying them in the America as Yang said the Brown corpus was the first and became the kind of corpus of the record model for other corpora including the Jiao Tong EST corpus. And in, in Britain, I was working in spoken corpora from the beginning, and, because we are also fascinated by the tape recorders being available for the first time, record and transcribe, speech.

And, so, and, there is an example which I, I give on the sheet because at the time in my university at the time, exactly the same time, these two types of study originated and started in 1963, one by my colleague Jame Thorn , and the other one that I was running myself.

And, now, after the some years, computational linguistics loses its theoretical foundation. And, it was realized that the generative grammars of today, would now actually generate the sentences. What is that the little things in your office? (laughter). See something introduced by my hosts. So, (laughter) actually the, the movement of computational linguistics was closely to artificial intelligence, and to try, and, to get a background there and, but not moving towards data, and to somewhere in the late1980s. And, again, this is supported without my position particularly with, if there was a recently published hand book on computational linguistics published by Oxford edited by Mitkov. And, I can recommend the introduction very much. It's quite short. It's available on the Internet from Oxford university press by Martin Kay. And if you want to follow up the development of computational linguistics, you will find it very nice to put out in Martin Kay's work. In the absence of, theory, and , they, I don't say the absence of data. Then, what, what came in, was through traditional approach to natural language processing, and treebanks, part-of-speech tagging, or other types of annotation. But still more theory, and judgment was by results, and this is where, as we say the judgment of our free is very negative. And, because neither then or now can any machine or, or, oriented a grammar of a language, and provide adequate description of any sentence of open texts. Actually current texts can not be handled by natural language processes.

And now, in contrast with this, corpus linguistics have a very different origin. And, after the early work in the, in the 60s progress became, was quite slow, because the computers of today was simply not up to it, neither their capacity nor the speed of processing, nor, their all the power of their operate systems. And it wasn't until 1980 that you become to get, a multi-million word corpora. And, as you can see from that on the whole, it was doubling every couple of years until about 1980s, 1996 or so since when I'm glad to say the development of large corpora has slowed down in west substantially and so we have only just over 500 million words now in the, in the Bank of English in London. Now this doesn't handful for the work of many linguists, many corpus linguists. Small corpora are very, useful for a very large number of purposes and particularly if you are willing to, do usual subjective judgment on the first output of the, of the corpus, there is a recently published book on the use of small corpora in the Language Teaching edited by Daisy which I recommend to you as containing quite a lot of very interesting materials.

but, if you have, and this is very begin, to look at the theory, if you have, as I have, a contextual theory of meaning, an idea that the meaning of the expression is bound up with the context in which it occurs, as goes back to the work of the British linguist J. R. Firth, of, more than 50 years ago. And then you find you are driven to look for larger collections of evidence in order to get the broader contexts and the wider combinations that you need in order to make adequate descriptions.

This is the context for all world wide web this is of course, an obvious contender for being the enormous contemporary of many languages. And I just want to say, at this point, that it is extremely useful, very reliable but we shouldn't mistake it for a corpus.

a corpus has got much more design and organization to it than the web. The web is a huge, repository for communications and has immense value but it, it does not have does not representative of any language in the way in which a corpus at least tries to be. so the basic

position of corpus linguistics is that, a, language user cannot retrieve all the meaningful structuration of a language that if it is if he or she actually use it if it is his or her language, or if it is a language with which he or she is familiar. So a corpus supplies evidence that it is not available by any other method. And, the from a the theoretical point of view and from the descriptive point of view, we have to eventually revise our pre-corpus theories and models in order to cope with the others. And I think that it is not the case and if the findings of the corpus linguistics actually according with the predictions made by the theories of the last century, then we would have no need for corpus linguistics as a sacred discipline. It would then be what anybody expects it is expected originally Nelson Francis and the others expected to be essentially a confirmation of what we've already known about the languages. And what is turned out to be it is not the enemy of the confirmation but a source of some very innovative and interesting and lightening inquiry.

Now if you look at it from the point of view of computer science, not from computational linguistics, but computer science, then, language text is very simple and straightforward phenomenon. It is almost exactly the kind of information that Allen Turing had in mind when he developed his original models from his computers that ever made except that it was a numerical models and not in terms of characters. But essentially all languages from the computational point of view are linear strings of characters which is very easy to handle in computational terms. And so, that's very basic and very simplistic, and they have to be combined in order to produce useful and meaningful units.

But although this is clearly based on a bottom-up model, that is to say a working-up from the data. It is very important to see that, the, when I put an NB sign in there linguists human beings that is to say, cannot actually abandon their linguistic intuition. We are not using a bottom-up model at all. We are using our intuitive understanding of languages in the presence of a large quantity of the relevant data. And that is essentially the opposition.

Now from the methodological point of view, taking this, from, starting from the computer science perspective, then, you only need to add very sparingly, some amounts of linguistic information. Because, so much of the processing, is independent of which particular language you are using. And so, you add this information, very rarely, very sparingly, small amounts of information can be very useful. For example, in a, arithmetic script, type of input, then the word space is clearly very important and rather different character from all the others. So, if you recognize that, and give it its value, in terms of its boundary qualities. And then, you have a very useful advantage to your research.

Now, from a theoretical point of view, from a theoretical point of view, I just want to, introduce some of the main lines that I expect to be a, the source of a lot of research and interest in the future. And the first one, I hope we'll be attending to it tomorrow at the workshop. this is particularly concerns spoken language. And I do urge a, very pleased there is and a renew interest in the study of spoken language in China. that, we start with spoken language because anything that you can demonstrate as being, a relevant descriptive category and in the spoken language, which also turn out to be useful in the written language, but not necessarily vice versa. And this is a problem in a large amount of our work present time, is that we have been imposing written models on the spoken language and then finding that it doesn't fit very well.

Well, it shouldn't be expected to because the written language is some of late sophistication language inputs from the outside. I say a late tradition I appreciate the language of the country, for which we espouse the earliest form of a written models but even then people have been speaking

for thousands and thousands of years before they have started the written language down.

Now, I would like to suggest that, there are essentially two characters of meaningful organization and this is where we should be looking for. two types of grouping of the, fundamental tokens like words or and, and whatever you, whatever basic token you use, the endocentric at the exocentric. The endocentric is where you see more than one token as grouping together forming around a single core to make a single unit of meaning and a lexical item, an idiom, a phrase, and something which, has got a single unit of meaning although it may well have several uh individual components and variable components. And the second one is the exocentric, and, which is clausal if you like. Well the essence of it is that you see there are two separate elements of which the obvious one is from traditional linguistics are subjects and predicates, but you perceive that there are two separates and units of meaning which are put together in this clausal or propositional way, in order to form a different kind of unit. And this two types of organization are, incompatible with each other. You have to decide one or the other and the overlap, so that, if you have some kind of an idiomatic phase, it could well be a single lexical item although it looks like something clausal. And, if you want to see as excessively independent choices. So this, this I think is taking things slightly more abstractly if you like than you normally have in grammar. But, I think this is the, the starting point of how you can, and, how you can organize patterns of corpus in terms of an organization which is reasonably similar to, familiar models.

Then we are following meaning all the time. And this is a very important point because, language is not a fully formal system, and cannot be because human beings, who are all, all, all very quirky and individualistic and they are interacting with each other in totally open and unpredictably ways. And so you can never pin the style in terms of total formalism. So what we are looking for is a semiformal way in which the meaning can be handled within the language. And here I would like to stress, there is already a well-known feature of language called paraphrase, which is where you can rephrase a meaning in very similar terms but slightly different and which allows one to relate different, relates meanings together without going outside the language.

And I think this is very important indeed because, linguistics has been subject to, imported models of semantics, ontological models, logical models, referential models and so on where meaning is supposed to lie outside of language. How do you understand the language or well you understand it with reference to the world. And I think this is an absurd position. you might be able to understand the world a little with reference to language, but not the other way around.

the world is, is, is, not subject to a even a semi-formal organization. logical relations are developed, of course, from language. There are systematically, there's, there are much more clear, clearly ordered and neat and precise, and natural language, and certainly tempting to try to describe a language with reference to them. But since they are derived from languages anyway, then again, that is going to lead you to either a vicious circle or a lot of other kind of logic load of certainty. And so paraphrase is the key, and I, how paraphrase has often featured very strongly in language teaching. And I never, I, I have to do it when I was learning English and also learning foreign languages. Always have me to paraphrase, and I used to wonder why, and why is, why are my teachers so keen that I should will to take this perfectly enough paragraph and rephrase it into another paragraph. I'm now beginning to understand that is a very important skill. now, this center of the, endocentric precedent is the lexical item.

And, it's a, this is a summary of what I take to be its essential structure. I've formed several

papers on this, and so I don't want to go over it in any great detail. But there is essentially a core, an invariable core, and the number it gathers because it is contextual, strength, it gathers other words and phrases around about it and sometimes so strongly that they become part of the item as we should see in a moment when I look it at an example. the, the, the two elements are the core that's a bit recognise in variable bit and the semantic prosody which is the, the, the pragmatic attitudinal meaning, which is the reason why you chose to express yourself in this particular way. And in this case, we have to re-examine the role of grammar, because grammars had an almost all than embracing, grip on the study of language for many years. And I think it has weakened itself as a result because it has been a, assigned roles which it does not actually have, so the roles for grammar that I think we should look for emerging from corpus work are threefold and I put in there.

First of all, the management of meaning. essentially the linear arrangement of language means that you can't say everything at once, so you have to put things down in a roll, and so in order to manage and to organize, or to say grammar has a crucial function.

Secondly, the assembling of constituents, that's, if you like, the sort of rules, the rules of agreement, the rules of concord, the rules of ordering and so on. these are, these are not meaningful, they are they are just simply, conventions. You do in this way and you don't do it in another way. You make a question in English by reordering some of the constituents, and that's just the way you're doing.

and the, and thirdly, the components of lexical items. and that's not a normal, a normally understood grammatical role. in this case, I will give you a little example of this, I can do it, yes, make it slightly bigger, yes. in the lexical item in English, which is to get somebody into trouble, or to get someone in trouble, either. I think probably the second one is a more American use first of all. I'm not sure.

You therefore have a choice of into or in as a preposition and it's not a grammatical choice, because a grammatical choice of a preposition is a choice of one preposition rather than the others giving you a different meaning. But here, there is no difference in meaning. the, the two are pretty well they might statistical alternatives. They might be variable alternatives, but you don't get different meanings. And so it's only a courtesy to call into and in prepositions here. They are playing just the same role as letters in the alphabet which serve to identify a particular lexical item to get someone to trouble. we go back, uh yes. A preposition is primarily a word that is in a mutual exclusive relationship with all the other prepositions, but since none of the other prepositions will fit in here, then it is not acting as preposition. So, this is a view of grammar which reduces its role in meaning creations, which I think it is quite correct. And, emphasizes that the main role in meaning creation is in the, lexical item, the development of lexical items, multiword lexical items, with their variations.

Now, in order to, just summarize the, importance of relevance of this foreign language teaching, because I am not, I am afraid, think very much about language teaching as well as you. these are, well, I would say, a new set of skills. I remember when I learned about language teaching, there were four language skills: reading, writing, listening, and speaking. I am sure those are still highly relevant, but I'd like to offer my own four, as being, operating in power as those in did you can cross and combine them.

But the first of them which is the ability to divide a text, spoken or written, into Chinese. This is something that we will do with in detail tomorrow. the ability to differentiate between

exocentric and endocentric structures. This is absolutely crucial. uh in any use of any language you have to be able, as you hear, say or you read the words come along, you have to know whether they are grouping themselves together to make a lexical item, or whether they are separating themselves into some sort of clausal pattern. thirdly, this again will emerge tomorrow in the workshop, the ability to recognize and to use language about language. Very important skills we have talked about the language, not just in a grammar class, particularly everyday conversation. You are constantly in fact inferring two aspects of your conversation and your writing. It is very important to, to have those skills. And fourthly, to have the skill of paraphrase and the ability to revise and rephrase, which I've going to is one of the most difficult things for, speakers of a language as a foreign or second language, particularly as a second language. Where you can express yourself very clearly and very fluently. But, it is very difficult that to make a few slight changes in that in order set to overcome communication hurdle or if one is making something misunderstood or something like that.

Now I want us to aware of these comments with a particular example and give you a few, pieces of information, about the word and I chose, a few difficult to explain why you choose the particular word for particular occasion. And, and I, I chose the word fortune for, for two main reasons, one is that I was, the we are very often accused in corpus linguistics of choosing words that have very negative meanings. the words that talk about desperate disasters, the awful thing happening and, and in details a paper published not so long ago which attempts, corpora is being extremely biased in this way, and so I talk about fortune has a nice, sound a nice, and it's also, a particular very small lexical item based on fortune called the fortune cookie, which is, associated in the United States with Chinese cuisine . I don't I don't get it in China much but it's, a little a little cookie inside there is a piece of paper which tells you fortune. It must originated in China somewhere. so we have very few examples of fortune cookie in the Bank of English but we got quite a lot of examples since you can see there also words "fortune", nearly 14, 000 occurrences of the word "fortune".

So uh I want you now to look at its collocational profile and I can do this I hope, yes. this might be rather difficult to, see, clearly but what I have done is, printed out the collocations of this word and in an order based upon the first column of figures which is the occurrence in within five words of the word "fortune" and so the indefinite article "a" occurs 6, 412 times in this, in the concordance of the word "fortune", the word "the" occurs quite often as well.

But if you notice the next figure which is a significance figure, which is a t-score, and then you see that whereas the indefinite article is gonna get a high t-score at nearly 15. That the word "the", the definite article, has got a minus t-score, a low t-score. so it's not, although it's all very common, it is nearly as important as the indefinite article, and so on. What I've done is picked out from this list, those that have the T-score of higher than 15, and the arbitrary figure, very high figure actually, t-scores of, more than 2 or 3 are usually important. But just to give a rather broad, general glimpse of this word, I have chosen the ones that occur, more than and the t-scores are more than 15 and obviously with a high frequency rating as well. So you see them the whole face, and I'll move that out from this column, and so those are the ones that we are going to look at. The one at the bottom, is "500". That's simply, there are some lists of companies called "The Fortune 500" index, and this comes up a great deal in the newspapers which I think is why it just comes just in the end of my list here.

So we go back to the looking at the, meaning of this collocational information. And first of

all, the indefinite article is very strong so we look at the combinations of "fortune" plus the indefinite article. We not talk here about positions and implications or just the two words. There the three collocates that come strongly are "made", "small" and "cost". I'm sure, you know immediately we are talking phrases like "made a fortune", "cost a fortune", and so on. And indeed that's quite correct. And the next most important common word is "his" and I will not go back to the profile for that. But, although these are very often found together, so that for "made" and "make", "You're making your fortune" and "He made his fortune". Soundly, it is the male species to make his fortune in the colony than the female as it reported in the financial crisis. And these occur for 85% of the incidences of the "fortune". And, with the cost small, his is not important at all. They hardly occur, but their enlarged the phrase and "made a fortune" and "made a small fortune", "cost a small fortune" and so on. They enhance the use of the indefinite article. So, what you get is, the lexical item around the indefinite article, and fortune, and which, is taken shape which we can conclude, is, this time, and "make his fortune" and much more restricted than "make a fortune".

Now, you also will notice it if you look at the "small fortune", it isn't fortune at all, and what is an ironic phrase for a lot of money, and more money than you would predict or expect in the second sense, but not a fortune in the sense of very large and some money. But if something you find is very expensive, then you say "cost a small fortune", as well a small fortune is not fortune at all. A "small fortune" is usually something rather expensive and either something with big a likely or more likely somebody has been over charge in you. And, so, and after this group of, collocates, of "a" and "his", the next one is "good", and fortune with good. And here you don't have an indefinite task at all, in fact ? fixing here. And the word half, and so you get a typically phrase like "have the good fortune", too, which has got an extra collocation that come in, like "the great good fortune", too. And other forms another lexical item around the word fortune, a different core in this case, not the indefinite article, but actually the forms are verbs "to have" and the words "to" and "the".

Then there are some minor items like "the squirrel of fortune", sometimes "fortune's squirrel" and that has got classic groups of the verbs that grow with you, and the squirrel fortune is just squirrel fortunes. And the core of the lexical item "fame and fortune", and that collocates with the set of verb which you can see that five "c" search ? and these are clearly associated in meaning, you won't find them together in a resource, in a conceptual resource, because they don't organize themselves so much conceptually, but they are clearly have semantically preference together of a certain way of approaching, the phrase "fame and fortune". Then, now thing is a very, very quick sketch, because I'm very nearly limited to my time, and a very quick sketch of the way in which the word "fortune" patterns in present day English.

Now I offer that now alongside a dictionary entry for it from Collins English dictionary. Now, I'm not, I think this is a not very useful, summary of the evidence that I put forward. But I want to say that I'm also the advise editor of this dictionary. So please call it corpus dictionary. So I have to take responsibility for it. but as you can see what it says is not untrue, but, no, it is not untrue. But it's I said of half truth. And, and it says a "small fortune" is a large sum of money, that's true. But it's not large in relation to fortunes.

And, and it says an amount of wealth of material prosperity, especially grave and qualified, a great amount. That's interesting because I'm qualified. And if you say "a fortune", that's a lot if you qualified in any way like a "small fortune" or a "considerable fortune" to some its length.

So the good point there is not, is not been properly made, and then is this power or force of a personalise regarded as being responsible for human affairs charges that, that doesn’t appear at all hardly. there is perhaps this "grand good fortune" is the nearest to that. I think that's really NO. 4 luck especially favorable. I think that in No. 4 this third word hardly appears at all. There is, an infrequent item with a core of fortune plus telling fortune tell us telling fortune, things like that. It's not very common use of the word and I've to say: It's not used, I think it goes for all of these things. This is, none of these is the meaning of the word "fortune". Each of these is a meaning of a phrase of which fortune in one of the elementary elements. But only one of those elementary elements and you can not say as it says in No.4, fortune equals luck. And, and it's only fortune if you include if the word good, for example. And that's, that's why, you see, it is noted there especially when favorable. All that means is it collocates with good, and you have to tell both good and fortune, in order to make its meaning.

So you see this is one, one example chosen nearly as random as I, as you can, and which, it's, it's supposed to illustrate the fact that this conventional dictionary entries its place is supposed to express the meaning of fortune. It's rarely it's advantage because there's no such meaning of fortune. fortune because only in a variety of combinations, each of which is a different lexical item and each of which has different meaning. So, although to such ? there is a bit of adjustment necessary in, in, inner what you start on the details of the adjustment then I think you will find that you are really beginning revolutionary. And you have to, you have to accept and appreciate that the results of this growing, impetus to corpus well. It's going to, it's going to overturn a lot of our present assumptions about languages and we must be ready for that and, we must be receptive to it and not, and not to be frightened by it perhaps. Thank you very much. I look forward to ?.

Yang Huizhong: There are some questions.

Sinclair: OK.

Yang Huizhong: John, professor, is always full of many genius ideas, I believe you have question to ask professor John Sinclair, he will be happy to answer, we have five minutes to ask questions.

Questioner: Is it, is it possible for us, for us to find common semantic elements in the word "fortune"? Is it possible for us to find, is it possible for us to find common semantics elements in the word "fortune" across these different phrases?

Sinclair: there is a general positive semantic prosody. That's to say one of the reasons I chose the word is in general uses of the word fortune are, are, give you an idea of rather good things. Whether, it whatever for now or it is a matter of a lot of money or, or, something good happening to you. So that is I think a common, semantic element, but it's not a semantic element in the normal classificatory semantics. It's a semantic element in the sort of pragmatic or attitudinal semantics. And that is what you do find quite commonly as a common element. But, whether, see if you regard, say making a large sum of money, has been somehow similar to being very lucky in other ways. Well, this is in danger, I think, of being handsome. This is in danger of, say well because these are both use of a word "fortune" for both of these, then they must have some associations that they got necessary out. So I think you'll find it in the pragmatic, the semantic prosodies, but I don't think you'll find it very often in the semantic preferences. It is not exactly accidental that the word "fortune" appears in all of these cases, but it's not a guarantee of a

uniform semantic, element.

Li Wenzhong: Professor John Sinclair, I am quite interested in the, your proposal for the language skills of a person. They are more like the language skills of a linguist. So my question is, do you think if we apply to the, your proposal of the language skills of a person to the EFL learners, so how we, how to make such language skills teachable to all the learners?

Sinclair: That is a very good point to raise because those do look like more analytic skills than performance skills. And it's a very important part of my belief that they are actually performance skills. I think paraphrase is one that is already amid to come out in the language teaching sites. I don't think I need to go into that in great till. But I think, for example, we take language about language, the ability to negotiate your own language in relation to what you are saying as you're saying it, is a primary discourse skill. You have to be able to be analyzing as well, but it is a primary skill of your operating just about every sentence of your speech. in the, we are negotiating that you are constantly referring to what you've just said or what somebody else has just said in all sorts of ways. And "that", as well. "That"! I've just said the word "that". That is language about language referring back to what I've previously said. Encapsulating it, and bringing it forward into the next theme. It's a crucial, essential language skill. The skill of chunking, we have to wait until tomorrow. And I hope that by tomorrow night you'll agree with me that it is a crucial and essential skill.

And, what's the fourth? Oh yes it was, the division, very important. The division between endocentric and exocentric. That I may be just, putting this in language that is unfamiliar to you. I do it quite deliberately, because I want this to be seen as a more abstract thing. But if you think it as endocentric means something like noun-phrases, and exocentric means something like clauses. Then I think you agree (laugh), you have be able to distinguishes between these. And in this I am supported from for me most unusual sector which is the most abstract or theoretical foreign linguistics, the work of McCarthy, on the complexity in language, which was published four years ago, and make it exactly simple that I am totally respective. And he said the unique feature of human languages is that they make noun-phrases and clauses. They don't need to, there are many other theoretical possibilities. But this is the feature of all human languages and it is a fundamental feature in ? as far as it is suggested. It is part of our mental wiry. Well I am not going to say that that's up to him not me. But what I do say is that it is a fundamental skill that we have, that we recognize all the time in the, as we hear words or as you read, we are constantly assigning these items into groupings. And those groupings are either grouping together and making a rich lexical item ?or separating them and seeing them as being essentially propositional unrelated to each other in the sense of an argument.

And, and so that I take to be fundamentals operational skill as well as of course as analytical skill. So thank you for that question because I think this point is very worth well making. And also you can of course take the four traditional skills, and you can take my four skills, and you have a very nice program of language teaching of 16 components because you can say what is you know what is how do you interpret this skill of, say paraphrase in terms of listening. How, how is that going to be brought up. You'll find in every case that they are very important to, teaching and learning capabilities in ? .Thank you.

Yang Huizhong: because of the time limit, I will allow the last question.

Questioner: If somebody say corpus linguistics is only a tool, or a research method but not a subject, discipline like functional grammar, phonology that kind of things, how would you

response?

Sinclair: Well I think, I think, I respond from, initially from my own experience, because I did twenty years ago, five years ago, I think, more or less than that, that is just a rich source of evidence, and indeed I in set up to write a dictionary which became the Cobuild dictionary, I set up to write that, and I, the whole operation was designed and was ?, thinking that we would be able to use, standard framework for as description of, of words lexicography, because I was working with a big dictionary house, Collins, and they already had lots of English dictionaries like the one I have, the other one, that's one of their dictionary. We thought all we were going to do is to take this evidence, sort it in to the existing frameworks that we have. And indeed the whole project was seriously in danger of being stopped, because after three years I have to confess that the evidence I had would not fit in with the lexicographical patterns that I was supposed to make it fit into, and the project has to be extended and legal battles, and also problems involved in it, so that for me it is a matter of experience, and there is absolutely a way in which you can expect corpus simply to provide evidence for theories and descriptions which have not taken corpus into account. And the reason for that is, as I said, before, that we do not, as human beings, as users of language, as students of language, as highly skilled, research linguists we have not the ability to, to recall and retrieve the evidence that we can find in the corpus. You can find in a corpus evidence of far more words than you cloud have read and in any case are we tend to abilities are nothing like those of the computer, and they are nothing like systematic, the depression about computer, is, I think, that it does not have intuitions, and so it is an ideal complement to the human being who does have intuition, the two together get a long fire, each on their own makes the myth. That would be my answer.

Note: The talk above is the transcript based on the video recording of the talk given by Prof. John Sinclair at the *International Conference on Corpus Linguistics*, Shanghai Jiao Tong University, Shanghai, China, October 25-27, 2003. The transcription was collaboratively done by 庄亮亮, 李昭锦, 苏子珺, 薛冰, 姚瑶, 李丽珠, 毕慧, 刘洁琳, 袁飞, 余香红, 龙江, 邹积铭, 颜雪飞, 陈茜, 孔蕾, 冯佳, 毕争, and 荀晓鸣, who are MA and PhD students at the National Research Centre for Foreign Language Education, Beijing Foreign Studies University. 龙江 helped collating the individual bits of work by all the other transcribers. A special thank goes to them all. The transcript in its current shape is an unedited one, which means it is a verbatim transcription of the original recording. So disfluencies are largely kept.

A Cognitive Model of Corpus-based Analysis of Chinese Learners' Errors of English

Gui Shichun

National Centre for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies

Abstract: The paper tries to introduce a cognitive approach to the corpus-based analysis of Chinese learners' errors of English based on the findings of CLEC (Chinese Learner English Corpus). The model is based on the competition model of MacWhinney and the cognitive approach to language learning by Skehan. A confirmatory factor analysis was conducted to test the hypothesis that the errors can be grouped under three different levels: lexical perceptual errors (like *spelling*, *number*), identifiable at single-word level; lexico-grammatical errors (like *substitution*), identifiable at into-word level; and syntactical errors (like *sentence fragment* and *construction deficiency*), identifiable at the sentential level. Correspondence analysis was run to show how error types and learner types are related to one another. The paper draws attention to the importance of language transfer in the writing of Chinese learners. The learners have two language systems (one more complete, one rather incomplete) at their disposal; which system to use depends very much on the writing task and the learner's certainty of fulfilling the task. As mature learners, they tend to fall back on the L1 linguistic system when they are required to express complex thinking. Errors are task-dependent, and they may not be an indication of their language proficiency.

Background

The present paper tries to adopt a cognitive approach to the corpus-based error analysis of Chinese learners' English by making use of the resources provided by CLEC (Chinese Learner English Corpus, CLEC). The corpus consists of one million words of written compositions by 5 types of learners: senior middle-school, tertiary college English (band 4), tertiary college English (band 6), tertiary majors in English (1st and 2nd years), tertiary majors in English (3rd and 4th years). The corpus is annotated with grammatical tags (automatically) and error tags (manually). It is available for public use at http://www.clal.org.cn/baseinfo/achievement/Achievement1.htm

The entire corpus is tagged according to an error-tagging scheme which divides errors into 11 classes, and 61 categories. For the sake of investigation, the categorization of errors is neither too exhaustive nor too simple, so that the markers can manipulate and the researchers can create a subcategory for his/her research purposes.

To set up our cognitive framework of error analysis we make use of only those errors whose frequencies are well above 1% of the total. There are altogether 21 error types.

Table 1: Table of Errors (above 1% of the total)

Type		st3	st3	st4	st5	st6	Total
fm1	Spelling	2424	3349	2556	2175	2063	12567
fm2	Word formation	439	522	531	270	405	2167
fm3	Capitalization	1853	851	491	826	213	4234
vp1	Transitivity	326	379	603	123	246	1677
vp3	Verb Agreement	470	610	950	325	401	2756
vp6	Tense	1465	414	377	452	263	2971
vp9	Modal/Auxiliary	140	319	338	51	105	953
np3	Noun agreement	254	288	302	251	230	1325
np6	Number	470	761	582	427	435	2675
np7	Articles	300	125	108	209	73	815
pr1	Reference	103	275	248	107	24	757
wd2	Parts of Speech	410	1081	935	270	306	3002
wd3	Substitution	1385	1901	2196	902	472	6856

wd4	Absence	737	965	537	480	532	3251
wd5	Redundancy	517	713	627	316	217	2390
wd7	Ambiguity	329	501	316	272	268	1686
cc3	v/n collocation	213	598	505	90	160	1566
sn1	Run-on sentence	527	694	699	141	54	2115
sn2	Sent. Fragment	535	459	367	158	98	1617
sn8	St. Deficiency	1392	527	1045	587	318	3869
sn9	Punctuation	1083	668	408	773	395	3327
总计		17760	18838	16869	10585	8343	62576

% of the total = 86.4

Basic considerations of the cognitive framework

In setting up our cognitive model we have taken the following points into consideration:

1. Errors are traditionally distinguished from mistakes. Corder (1967) associates errors with failures in competence, and mistakes with failures in performance, making use of Chomsky's distinction. But more and more cognitive psychologists have found it hard to separate one from the other: Aitchison (1998), who is heavily committed to Chomsky, "finds it quite odd that anybody is able to concentrate on one rather than the other of these factors, since they seem to her to go together rather closely." Corpus linguistics by nature deals with frequencies of data, which demonstrate language performance; language competence is only the derivation from our observation of language performance. Unless we have evidence to show that a learner keeps on making the same mistake, we have no way of determining whether it is a competence error or a performance error. As Johnson (1988) points out, if learners say or write a form that is wrong, it could be either of two reasons: either they lack the requisite knowledge (this is a case of ignorance) or they deploy knowledge they do have, but it happens to be wrong knowledge. So in our study, we use 'error' as a cover term for all ways of being wrong as an FL learner. Errors are results of "uncertainty" in language performance and "uncertainty" is a kind of probabilistic behaviour that is a gradient continuum. There are various kinds of uncertainty that can be traced back to cognition:

   False analogy: books, news > knowledges, informations
   Incomplete application of rules: development > advantagement
   Redundancy: "这是一间三层高的建筑">it was a three-story-tall building
   Overgeneralization: entered the classroom>returned the classroom

2. In terms of "emergentism", verbal behaviour (errors as well as linguistic structures) can be considered as an emergence process. The emergentist approach to language acquisition views language as a structure arising from interacting constraints, much as the shape of the coastline arises from pressures exerted by ocean currents, underlying geology, weather patterns, and human construction. According to this view of language learning and processing, the behaviors that we tend to characterize in terms of rules and symbols are in fact emergent patterns that arise from the interactions of other less complex or more stable underlying systems. In the Chinese context, English learning takes place in a non-English community, with Chinese being a language whose sounds have no connection with the written forms (characters), so errors display the interactions of social, cognitive, and even physical factors. According to the competition model (MacWhinney, 2000), language processing (including the occurrence of errors) involves the competition of cues. The learning of the system of form-function mappings is driven by cue reliability. There are four dimensions that contribute to cue strength:

   a) Task Frequency. The most basic determinant of cue strength is the raw frequency of the basic task. The task of determining the agent of the verb occurs with virtually every transitive verb in English. Similar to English, Chinese is also a language that makes more use of word order. Chinese learners of English have no difficulties in identifying the agent of the verb, but are ill at ease with determining the transitivity of the verb:

*I like listening all kinds of music.
*I have lived this village for many years.
*But no one came this island again.

b) Availability. Within a given task, cues will vary in their relative availability. We are more interested in knowing whether a cue has a contrasting effect (known as "contrast availability") than in just knowing whether it is present (known as "simple availability") or not. In the sentence "The cat chases the dog," both "cat" and "dog" are singular. The fact that the verb is marked for a singular subject tells us nothing about the status of the subject. The agreement cue is available, but not contrastively. Whereas the agreement cue in "The cat chases the dogs" throws light on the status of subject. Contrast availability is more significant to German learners of English that to Chinese learners, because German speakers rely on overt morphological forms as markers of syntactic relations. Whereas both Chinese and English lay more weight on word order. Chinese learners can identify the subject by default, but they have difficulties in observing verb agreement.

c) Simple reliability. The most important and basic cue validity dimension is the dimension of reliability. A cue is reliable if it leads to the right functional choice whenever it is present. It is equally reliable if it leads to the wrong functional choice in the case of second language learners, by way of making good prediction.

d) Conflict reliability. In addition to simple reliability, cues can also be characterized in terms of their conflict reliability *vis a vis* some other particular cue. In second language acquisition, different cues may lead to different kinds of error depending on their strength. When the learner is uncertain about using some expression, the cue with the strongest weight will come into effect.

In most situations, Chinese learners of English are adult learners who have already set up a well-organized neurolinguistic system, while they are in the process of setting up an L2 system. According to MacWhinney (2000), the learning of L2 is initially highly parasitic on the structures of the L1 in both lexicon and phonology. From our observations, Chinese learners have two language systems (one more complete, one rather incomplete) at their disposal; which system to use depends very much on the writing task and the learner's certainty of fulfilling the task. When the adult learner is faced with a task of expressing his or her complex conceptual representations in the target language, they will try to use their newly learnt language system. If their knowledge of the system is incomplete, or the task is too demanding, they are bound to fall back on their own L1 system. This task-dependent character becomes even more obvious in thematic writing in an English test, where it is more difficult to pursue the strategy of avoidance. So instead of saying "浪费了时间和精神", a learner would say "it is a waste of time and spirit.", and instead of saying "不想增加人", another learner would say "don't want to add people." This is what Schachter (1978) identifies as "resident errors."

3. The above examples above illustrate a process proposed by Skehan (1998) that L2 acquisition seems to follow the process of L1 acquisition: lexicalization→ syntaticalization →relexicalization. The only difference is that, once the critical period is past, as in the case of adult learners, there will be a greater predisposition towards the exemplar, memory-based system, and the internally-generated pressure for syntacticalization will not come into play. In other words, there is a danger that the L2 learner will not progress beyond the first of the three stages mentioned above. At the lexicalization level, language transfer seems to play a bigger role.

4. A cognitive approach to error analysis also suggests that errors can be divided into several levels.
   a) Lexical perceptual level. Errors at this are also known as "substance errors" (James, 1998), and they are related to perceptual representations, especially to memory, such as "memory failure" or "memory distortion". Typically these errors can be identified at single-word level, as *spelling* or *word-formation* errors, or by looking at the neighbors of a word as *absence of the article.*

   b) Lexico-grammatical level. Errors of these types result from misconception of target language

system. When looking at the errors of our learners, it is very difficult to isolate grammar and lexis into separate categories, because grammar does not exist on its own. James defines it as "text-level errors". Typically these errors can be identified at the inter-word level, by looking at the word and its neighbors.

c). Syntactic level. Errors can be identified at a broader context, at the sentential level. James chooses to call it "discourse-level errors", but we propose to reserve the word "discourse" for another upper level. L2 learners may often produce grammatical sentences, but their utterances still sound foreign. As what Pawley and Syder (1983) have pointed out, "native speakers just do not say things that way." They term "the capacity to sound idiomatic" as "native-like selection". In the case of error analysis, the question can be put in another manner, "If our learners correct all the errors at the three levels, will their writings be similar to those of the native speakers?" The answer is obviously no. They may still have problems at the discourse level. For the moment, we leave this level open to further investigation, and we retain the term "discourse-level" for later description of those errors.

5. To summarize what we have discussed, the following cognitive model of error analysis in connection with L2 acquisition is proposed:

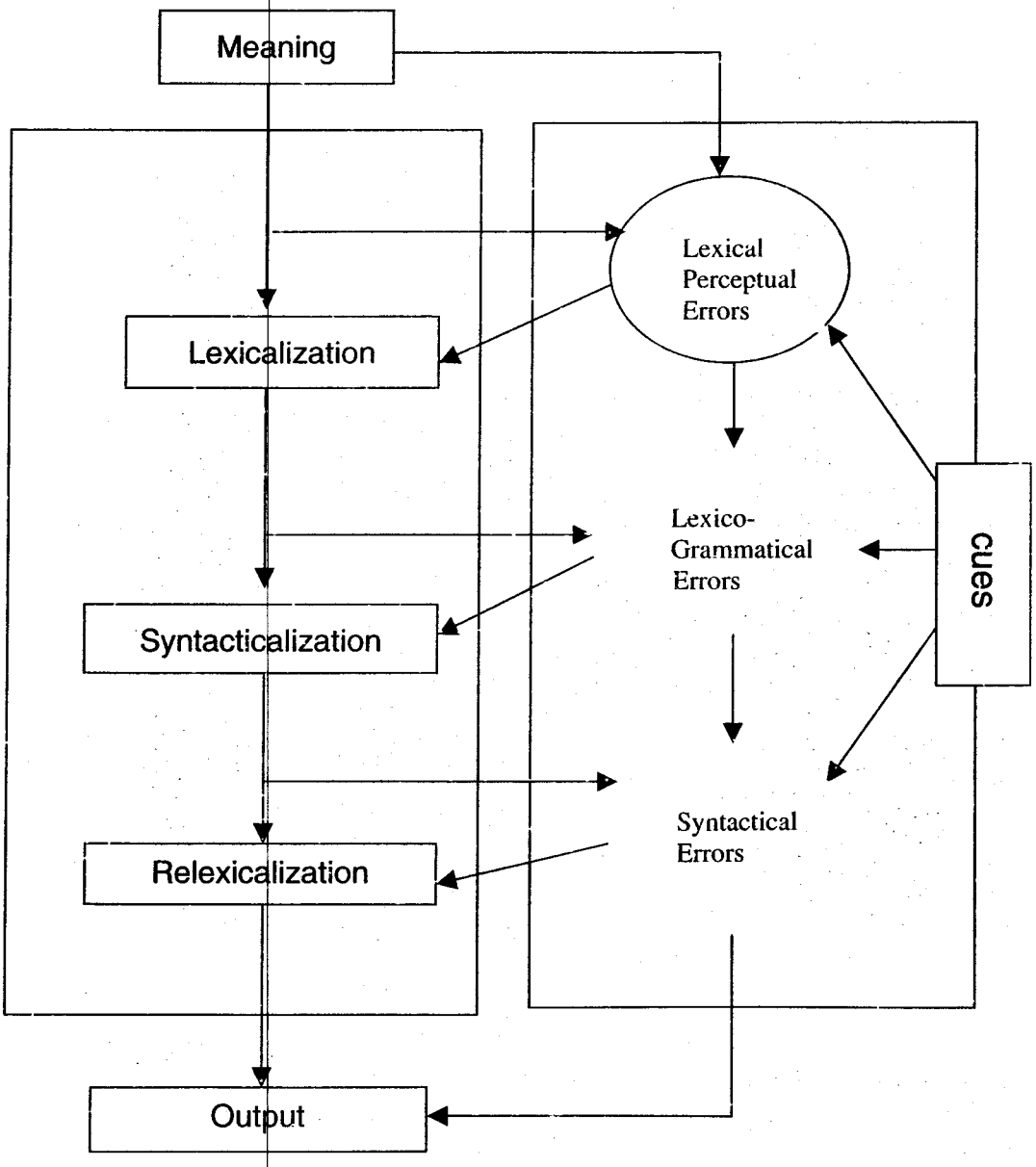


Figure One A Cognitive Models of Error Analysis and L2 Acquisition

Empirical Investigation of the Model

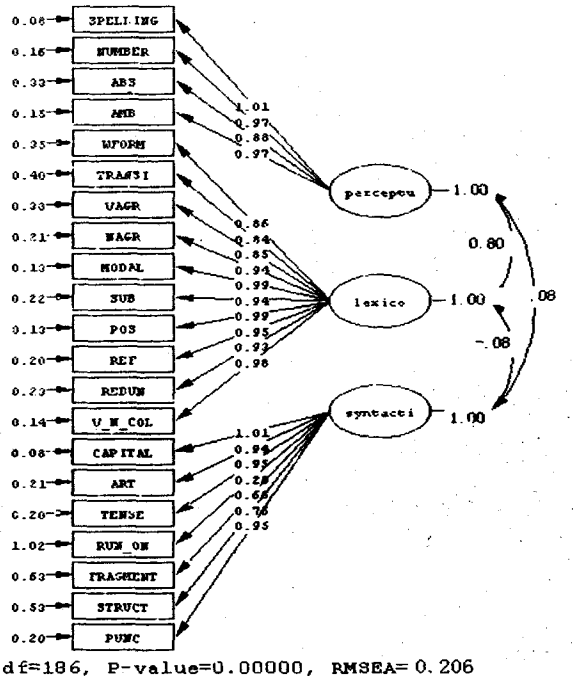
We shall now try to see whether the model fits in with the frequencies of errors collected from CLEC or not.

1. Confirmatory Factor Analysis

Confirmatory factor analysis of Table 2 was condueted by using Lisrel 8.50, which shows clearly that there are 3 factors, and they are grouped under 3 categories as what have been defined. Path analysis shows that all the parameters (values of λs)of the hypothetical paths are significant except run-on sentences. Since it is clearly a syntactic component, we shall put it in the third group (syntactical), and keep an eye on what will happen in our analysis.

Figure Two also shows that the lexical perceptual factor correlates with the factor of lexico-grammatical errors (.80), whereas the syntactical factor has no correlations with either of them. All these look reasonable except for the fact that the chi-square value is too large, showing that the model doesn't fit the data very well. Modification indices are offered in the output so that the model can be modified to fit the data better. Since we are only interested in looking at the groupings of errors, we shall not go any further.

Figure 2 Confirmatory factor analysis of 21 types of errors

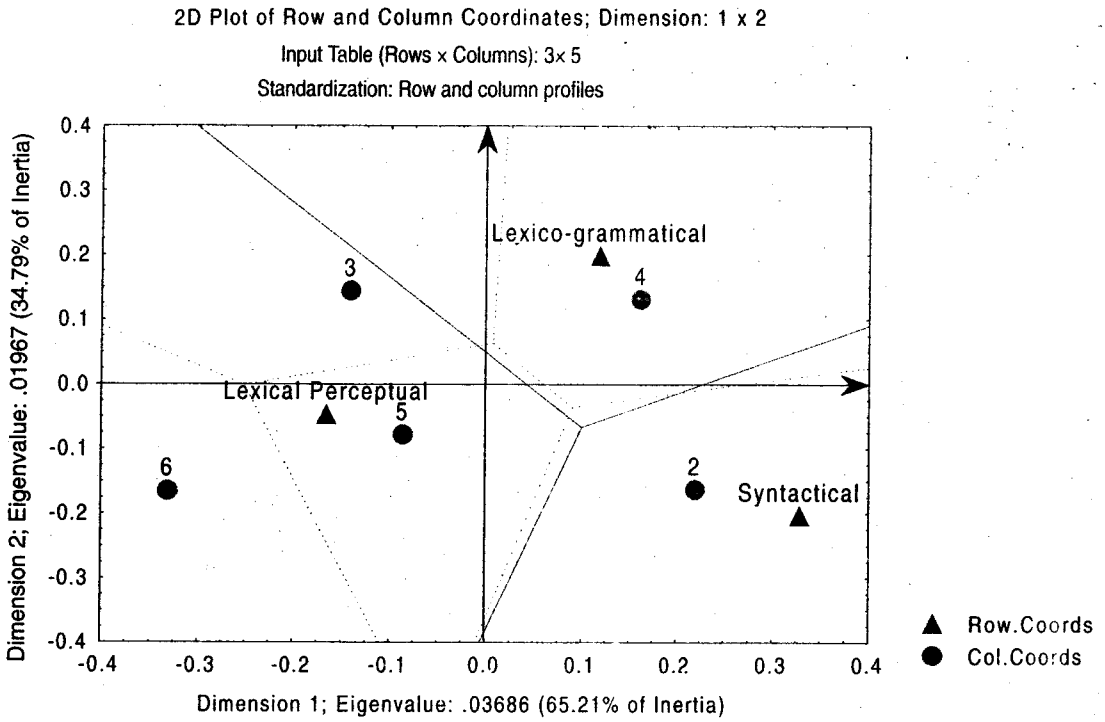


Chi-Square= 373.1 , df=186, P-value=0.00000, RMSEA= 0.206

We have also tried to put these errors into 3 groups and conduct correspondence analysis in order to find out how they are related to different types of learner. Correspondence analysis is a technique for describing contingency tables. (Lebart et al 1998) The description essentially takes the form of a graphic representation of associations among rows and among columns. Statistica 5.0 was used to make the graph, and Voronoi scatterplot can also be obtained to offer a better view of rows×columns relationship.

Table 3
Variables (columns): 5
Cases (rows): 3
Eigenvalues: .0369, .0197
X^2=1316.73, degrees of freedom=8, p=0.000

Figure Three Correspondence Analysis of Rows (errors) and Columns (learners)



We can see from the figure that there are 3 groups of errors, 5 groups of learners. St2 tends to commit more syntactical errors, whereas st4 tends to commit more lexico-grammatical errors. In terms of lexical perceptual errors, st3 and st5 can be grouped together. In terms of learners, St6 seems to be further away from the rest.

2. Analysis of Lexical Perceptual Errors

Spelling errors are the most typical lexical perceptual errors, because they can be readily identified at single-word level. They appear to be the most frequent errors of Chinese learners. *Number*, *absence* and *ambiguity* also belong to this category:

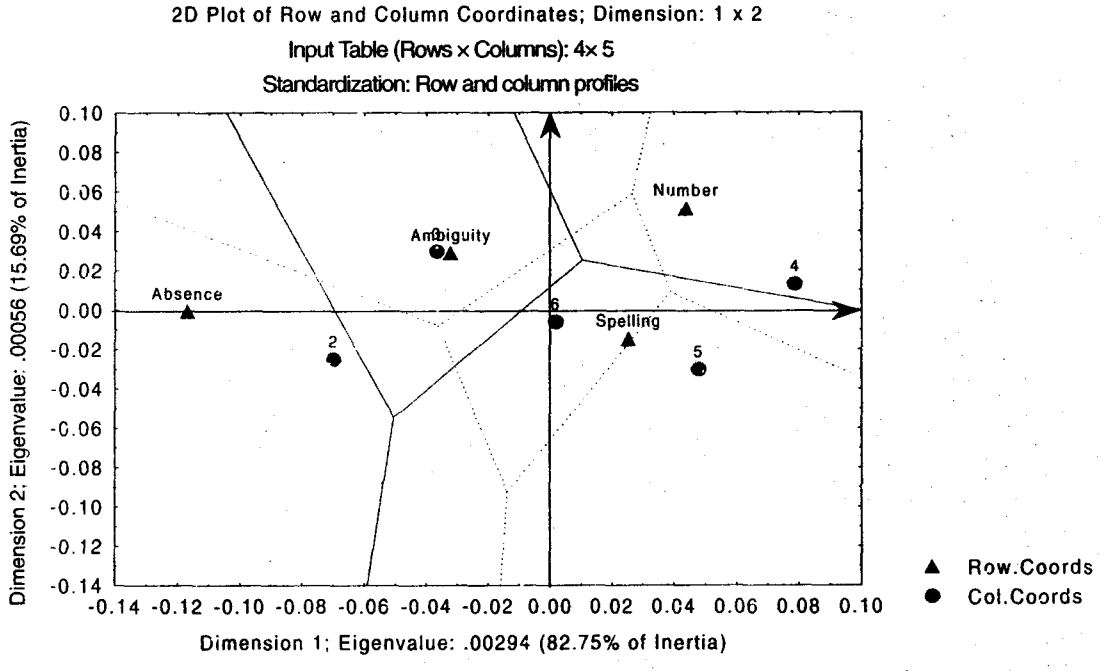
Table Four: Examples of Lexical Perceptual Errors

Error Type		Example
Spelling	**(Exchange of vowels)**	great>*graet*(10), cigarette>*cigeratte*(4), received>*recieved*(10)
	(Vowel)	benefit > *benifit* (32), soldiers > *soliders* (18), signature > *signiture* (10)
	(Addition or deletion of vowels)	mortality>*mortaility* (374), fresh>freash (9) ; beautiful> *beatiful* (20), create>*creat* (16)
	(Exchange of consonants or vowels)	etc>*ect* (32), first>*frist* (13), challenge>*chanllege* (9), environment>*enviorment* (14)
	(Consonant)	modern > *morden* (37), realized > *realised*(20), William > *Willian* (11),emerge>*energe* (10)
	(Addition or deletion of consonants)	develop>*develope* (24), college>*colledge* (24), can't>*cann't* (21); government>*goverment* (46), environment>*enviroment* (40) , studying>*studing* (31), knowledge>*knowledge* (30)
Number		information>*informations* / circumstances > *circumstance* / *several test*/conflicts and war/ *in* /\noisy city/ /\very good passage
Absence	**(determiner)**	the moon is /\brightest, my hope was/\same, I dressed myself in /\hurry
	(preposition)	*they went out /\the place, I sat back /\my chair, what is he*

		thinking 人?
Ambiguity	**(transfer)**	*They were neck to neck.* ("他们肩并肩"> They stood shoulder to shoulder);*China is one of the oldest of fours in the world*; ("中国是世界上四个最古老的国家">China is one of the four older countries in the world) ; *on the 12th of the third lunar calendar.* ("阴历三月12日">on the 12th of March in the lunar calendar)

Correspondence analysis shows that *spelling* errors are closer to the centre; whereas st2, st3, and st4 tend to be closer to *absence*, *ambiguity* and *number* respectively.

Figure Four Correspondence Analysis of Lexical Perceptual Errors



3. Analysis of Lexico-grammatical Errors

Lexis forms the core of lexico-grammatical level. I exical items have both lexical meanings and grammatical meanings, so they can only be identified by relating them to the context (at least the inter-word level). This level is the most predominant of the three levels. They occur somewhere between the process of lexicalization and the process of syntacticalization as shown in Figure One. There are three different sub-levels:

a) The first sub-level is more or less related to lexical perceptual level, for example, errors of *verb* or *noun forms; homeworks, ourself;* and the misuse of *parts of speech* and *affixation (harmness, bestly; unequality, unsimilar),* which can readily be recognized at single-word level. Learners at this stage also coin non-existent words, which typically reflect that they have already obtained some grammatical knowledge of L2 system. That is why we believe that word formation should be included at this level.

Table Five: Examples of the First Sub-level of Lexico-grammatical Errors (single-word)

Error Type	**Example**
Word Formation (irregular verb)	*rised, hitted*
(-s)	*factorys, sangs*
(coinage)	*admirment , darkmen , belowed , bestest , attacktion*

b) The second sub-level is most typical of lexico-grammatical level. Errors of this type have to be recognized at the inter-word level.

Table Six: The Second Sub-level of Lexico-grammatical Errors (inter-word)

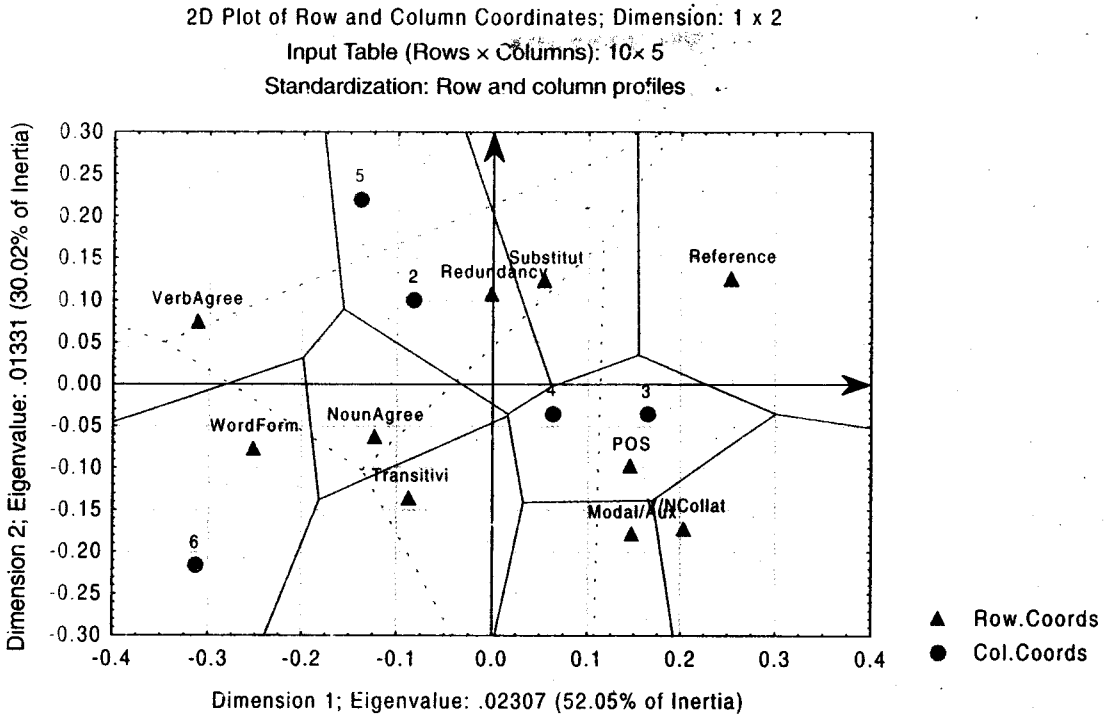
Error Type	Example
POS (adj)	*It is not difficulty that we can find.../ a great deal of raining water.*
(verb)	*the people who product fake commodities/ we should rich our knowledge*
(noun)	*There is some different in this sport meets.*
(adverb)	*to write a word beautifully and rapid./ painlessly death*
Redundancy (determiner)	*At 7:00 am I have a breakfast./ devote themselves into some certain jobs of the society.*
(noun)	*baby infant mortality*
(verb)	*The week is will pass soon.*
(adj)	*in their own given fieids*
(preposition)	*So he lives in a happy life.*
Substitution(transfer)	"如果您碰到难题">*If you match difficult problem*; "接触各种人群"> *touch all kinds of people*; "你必须有足够的条件"> *you must have enough conditions*
(word form)	*Weather effects (=affects) us in one way or another./ Your company is booking (=looking) for a secretary.)*
(Hyponymy)	*People take (=pay) more attention to it./ We must make (=take) measures to deal with this.)*
(phrase)	*Make these new words put in our hearts./The whole office was in a noise.*

c) The third sub-level of lexico-grammatical errors occurs in still broader context.

Table Seven: The Third Sub-Level of Lexico-grammatical Errors(broader context)

Error Type	Example
Transitivity	*Guangzhou was a good place to live. / But no one came the island*
(transitive or intransitive)	*again./ they cause to water pollution, / let us consider of it further.*
(complement)	*I felt it very cold. / I began felt very tired.*
Verb Agreement	*Cigarette smoking do a lot of harm./ I has graduated, so I hope I*
(subject)	*can have this chance./ Many young ladies or boys likes to stay in the western coffee shop. /Their outrage still receive severe punishment today.*
(people, every,etc)	*People argues that euthanasia or mercy killing is humane. / Every player were very hard.*
Modal/auxiliary (infinitive)	*I can became a useful woman./ It will brings luck./ I will to do something for people.*
(perfective)	*I have never see it before.*
(transfer of 能够、可以)	*Many good teachers can respect their students./ I must can do it well./*
	The student may consider that the teacher is too hard to get on with.
(to be)	*Juvenile delinquency is increasingly become a focus of social concern.*
Noun Agreement (noun)	*It has two door./ In a words, practice makes perfect.*
(other NP)	*We have to wash the clothes with our own hand./ You can find out the meaning of the new words.*
N/v Collocation (transitive)	*People like to eat Chinese tea./ They must listen to the lesson more carefully.*
(voice)	*When your friend or relative entered his job, your work was arranged at the same time.)*
Reference	*My aunt came to my home with his son./ As a student who majors*
(anaphora)	*in English, we can't just focus on the language itself.*
(it)	*If we do not use fresh water, we must shut it up/ I will remember to learn from our world from time to time and really put it into practice.)*

Figure Five Correspondence Analysis of Lexico-grammatical Errors



The figure looks a bit complex, because we have ten error variables. We can see that the five groups of learners are distinguished from one another, with st3 closer to st4, and st2 closer to st5. Learners of higher proficiency (st5 and st6) are further away from the errors. St2 tends to commit more *redundancy, substitution* errors; st3 tends to commit more *POS, modal, V/N collocate*, and *reference* errors; st4 is situated near the centre, it is related to *redundancy, substitution, reference, noun agreement and transitivity* errors, almost at equal distances.

4. Analysis of Syntactical errors

Errors at this level occurred at sentential level, either within a sentence or beyond the sentence boundary.

To begin with, we've put *capitalization* and *punctuation* under this category. Some of these errors are actually quite simple, like *I like Miss wu best, /...came to My uncle's house* and they *should belong to the lexical perceptual level.* Even with the sentence *During the Spring Festival. I had a happy day,* it involves just changing of the full stop into a comma. But these two errors are related to each other, errors can sometimes be corrected by either way. For example, *She talked well. first she let us listened...*We may capitalize the word "first", or change the full stop into a comma, and insert "and." In any case, it is necessary to look at the context of a sentence in full beyond we can correct it.

Table Eight: Examples of Syntactical Errors

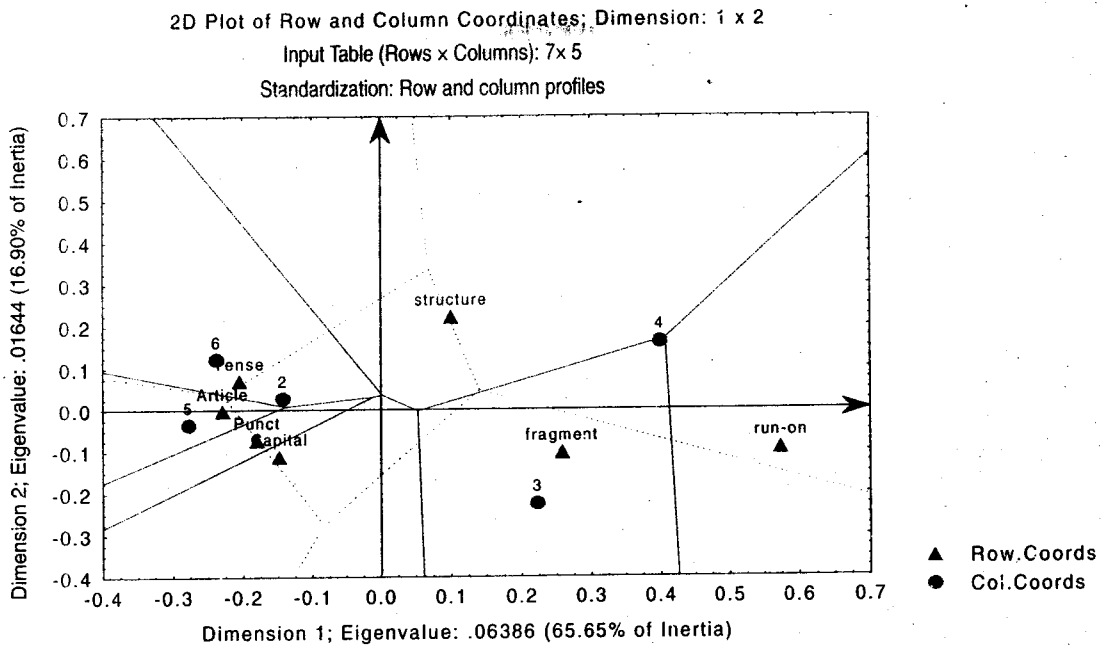
Error Type	Example
Capitalization	... he learned English and Russian and Wrote the Civil War in France./ ...the price is $2.50 for it. "well, can you make it a little cheaper?"
Punctuation	When playing football or basketball. You might be using 400 calories an hour./ If we know nothing about it. How we will survive? /
Article	The question is I like English, and I want to enter the institute of foreign languages./ Till now, no one can find out the solution to a riddle of this hole.

Error Type	Example
Run-on sentence (subordinate)	If I am not famous, it doesn't matter, I don't mind this./ Since the first person appeared on earth, there have been conflicts and wars because of strife for living, while people are dreaming of living in a world full of happiness and peacefulness, however, now the history of human being has come to the 20th century, the dream has not been realized, the shadow of war still permeates in the world./
(coordinate)	They carry the lantern and run in the street, they sing songs, and shout happily, they make the festival more lovely./The coverage of forest in the world is reducing rapidly, a lot of earth are being washed away, many acres of original fertile soil are becoming deserted, acid rain, air pollution appear in many places.
Fragment (subordinate clause)	As they do more exercises and often think deeply. / But have something one day./ Relaxing with friends you feel more interested./
(because)	Because we look upon them as a different kind of people who are acceptable in no place except prisons.
Structural Deficiency (re-phrase)	During I spent my holidays in Beijing about ten years ago,.../ Can you exist in society without money? Be sure not to be
(transfer)	世界上的战争使很多受苦的人无家可归，没有东西吃，没有衣服穿，没有水喝，甚至很多病>fighting to all over the world causes a lot of suffering-people are homeless, no food to eat, unwearing and no water to drink, even full of illness
Tense (adverbial)	Last week I spend the whole Mid-Autumn night with my roommate./ Before the computers are invented, people use abacus to count./ In the future I do my best to get to know this society.
(clause)	you had hoped to work for a job, so that you can earn a little money to buy something you want. /When I studied in a high school, there are national physical competitions.
(beyond the sentence boundary)	When I arrived there, I was very pleasantly surprised. There are many trees and flowers in it./ This winter holiday I go back town home with my presents. I had so much lucky money from the relatives. I'll take the money back./

Earlier we have mentioned that according to confirmatory factor analysis, it is better to place run-on sentences under the category of lexico-grammatical errors. Now we can see for ourselves that as syntactical errors they are quite special to Chinese learners. Some of them can be corrected by just changing the punctuation or capitalization, and in the spoken form, they seem to be acceptable. In modern Chinese writing, the conjunction can be omitted, and parallel constructions are often acceptable. So this is also a matter of language transfer.

Correspondence analysis reveals the following:

Figure Six Correspondence Analysis of Syntactical Errors



St2, st5 and st6 are close to one another, and they tend to make more *tense, article, punctuation,* and *capitalization errors*; whereas st3 and st4 are two separate groups and are of equal distance to *fragment* and *structure*. St4 also tends to write more *run-on sentences.*

Some Concluding Remarks

1. On the whole, identifying errors at 3 levels seems to be working well in our cognitive model. Our categorization of errors is based on the error-tagging scheme we have laid out, and so far we've not covered errors at the discourse level, because,

   a) It is difficult to set down the standards for "native-like selection" as defined by Pawley and Syder (1983);
   b) It is even more difficult for Chinese markers of errors to observe the standards.

2. The grouping of errors is not as clear-cut as what we've thought. Very often the same type of error can be put into different categories or the same type of errors can occur at 3 different levels depending on the situations. We can only say this is done according to the main tendency.

3. At every level, language transfer seems to play an important role. This is because the adult learners have set up their L1 (more complete) linguistic system and are in the process of setting up another linguistic system (rather incomplete). As mature learners, when they want to express their complex thinking, they often fall back on using the linguistic system that is more familiar to them.

4. Occurrences of errors depend very much on the writing task and the learner's certainty of fulfilling the task. They may not be an indication of the language proficiency of the learners. CET learners tend to commit more lexico-grammatical errors because their data were collected mainly from CET compositions.

5. Correspondence analysis offers a good visual graphing of the relation of error types against learner groups. However, we should be careful with the interpretation of the graph, because the mapping is done according to the frequencies of errors, and they, in turn, depend on the writing assignment, as pointed out in 4. All our interpretations are tentative and subject to change, when more and more data are collected.

References

[1] Aitchison, J. 1998. The Articulate Mammal: An Introduction to Psycholinguistics. The 4th Edition. London: Routledge. 184

[2] Corder, S. 1967. The significance of learners' errors, International Review of Applied Linguistics Vol.5 No.4:161-70. Reprinted in S.P. Corder(1981) Error Analysis and Interlanguage. Oxford: OUP

[3] James, C. 1998. Errors in Language Learning and Use: Exploring Error Analysis. London: Addison Wesley Longman. 129-172

[4] Johnson, K. 1988. 'Mistake correction', *English Language Teaching Journal*. Vol.42 No:2 89-97.

[5] Lebart, et al. 1998. Exploring Textual Data. Dordrecht: Kluwer Academic Publisher. 45-78.

[6] MacWhinney, B. 2000. The competition model: the input, the context, and the brain. In P. Robinson. (Ed.) Cognition and Second Language Instruction. Cambridge: Cambridge University Press. 69-90

[7] Pawley, A. & F. Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. Richards & R. Schmidt (Eds.) Language and Communication. London: Longman. 191-226.

[8] Schachter, J. 1978. Interrelationships between total production and error production in the syntax of adult learners. *Papers in ESL*. NAFSA

[9] Skehan, P. 1998. A Cognitive Approach to Language Learning. 89-92. Oxford: OUP

[10] 桂诗春、杨惠中. 2003. 《中国学习者英语语料库》.上海：上海外语教育出版社.

A Corpus-Based Study of Reporting Verbs in Fictions: A Translational Perspective

LIU Zequan HONG Huaqing

National University of Singapore

Abstract: This study sets out to compare verbs which are used in both original and translated Chinese and English fictions to "report" the characters' utterances, with an aim to investigate major linguistic differences between the two languages in their use of reporting verbs. The data investigated consist of two parallel corpora: (1) the Chinese *Hong Lou Meng* and its two English translations; and (2) the English *Tess of the D'urbervilles* and its Chinese translation. Concordance analysis of the data shows that since Chinese resorts to three or four implicit verbs in reporting, it relies on more than a dozen pre-modifiers to explicate the content, mode or nature of the reported utterances. In contrast, English dose not load attributive verbs with adverbs of manner in reporting. Instead, it tends to employ more than a dozen explanatory reporting verbs.

Key Words: reporting verbs, corpus, translation, fiction

1. Introduction

While reporting verbs refer to verbs that are used to tell the reader what someone has said, written or done, distinction must be made between those employed in academic writings and those used in everyday interactions and literary works. The former concerns verbs which academic writers use to introduce, especially for the purpose of literature review, information they have read from other authors/researchers. The latter, by contrast, covers verbs which people use to convey the speech or opinions other people have expressed in verbal communications. Fuzzy as it is, this distinction is necessary in that most verbs in the former category entail evaluation in reporting: "the writer is...under a conventional obligation to justify mentioning the author in the present context" (Thompson and Ye, 1991: 367). That is, the writer must show her authorial stance toward the validity of the reported information or opinion. This justifies why some verbs, e.g., *advocate*, *assert*, *claim*, *generalise*, *dispute*, to name only a few, seldom or even never occur in ordinary communication, while other verbs such as *ask*, *exclaim*, *order*, *protest*, *say*, *tell*, etc., rarely, if ever, appear in scientific reports or academic writings.

Following Swales's (1981) pioneering study of the introduction sections of academic papers for EAP (English for Academic Purpose) teaching of academic reading and writing, especially to non-native-speaker (NNS) students, a great deal of attention has been paid to reporting verbs that are used in academic writings. And this has understandably resulted in a number of research findings with significant pedagogic insights and implications, for instance, Swale's (1987, 1990) metaphor of "creating a research space" and identification of tense variation in reporting previous literature, Tarone *et al*'s (1981) analysis of verb forms used in journal papers, and Thompson and Ye's (1991) classification of verbs in terms of their denotation and evaluative potential in academic papers. These studies and their respective findings have not only contributed enormously to the teaching of the organization and presentation of introduction sections in academic genre to NNS students, especially in the sphere of course material development. Importantly, they have also helped to raise the NNS students' awareness to various reporting verbs in academic contexts, thus enhancing their ability in coming to grips with the use of such verbs in their reading and writing of research papers.

However, to the best knowledge of researchers, studies of the use of reporting verbs in conversational settings and literary works have scarcely been seen in either EAP or translation studies. This might be attributed to the fact that such verbs are so frequently employed in our oral communication and fiction texts and that they are easily overlooked. This is incompatible with the attention that has so far been given to reporting verbs in academic situations, and thus not conducive to the balanced acquisition and mastery of reporting verbs in a comprehensive manner. In order to examine how Chinese and English reporting verbs

as used in literary works are similar to or different from one another with reference to the content, mode, and nature of the reported utterances, this study sets out to make a corpus-based contrast analysis of the reporting verbs used in one Chinese and one English classic novel as source text (ST) and their respective translation(s) into the other as target text (TT).

The study can be justified from two perspectives. First, reporting verbs provide a culture-bound area in both English and Chinese texts. According to Ardekani (2002: 125), "[i]t is probably here that we might join the chorus of Sapir-Worf and Lotman that 'No language can exist unless it is steeped in the context of culture and no culture can exist which does not have at its centre the structure of natural language'." Second, from the perspective of translation studies, the study will, through a corpus-based approach, allow us first to record the strategies of translation which are repeatedly opted for by the respective translators to deal with the ST verbs in question, and then to make intersubjectively testable generalisations of translator behaviour, i.e., norms. It is hoped that the study will yield results that will contribute not only to the teaching of translation, but also to NNS language learning as well.

2. The Data

The data of the study consists of two parallel corpora: one is the Chinese classic *Hong Lou Meng* and its two English translations, and the other the British novel *Tess of the D'urbervilles* and its Chinese version. These two corpora are chosen out of the following considerations. First, the two novels are each a representative of the literary achievements of their respective languages. Second, they both contain sufficient instances of narrations which entail the use of a variety of reporting verbs. In addition, for the sake of the application of corpus-based analysis, the availability of electronic versions of both source language (SL) and target language (TL) texts is also concerned.

With respect to Hong Lou Meng or Dreams of the Red Chamber, the 850,000-word Chinese novel is a combination of eighty chapters of what is believed to be the original manuscript of Cao Xueqin (1724-1764) entitled *Shi Tou Ji* (*The Story of the Stone*) and forty chapters of continuation believed to be written by Gao E when it was published in 1791-1792. The selected two English translations are, on the one hand, David Hawkes and John Minford's version (830,000 words), and on the other, Yang Hsienyi and his British wife Gladys Yang's version (620,000 words) respectively. As far as Thomas Hardy's (1840-1928) *Tess of the D'urbervilles* is concerned, it contains 59 chapters with 151,000 words in the original, and has 300,000 words in its Chinese translation by Wang Zhongxiang and Nie Zhenzhao. In so doing, we believe that the two English versions of *Hong Lou Meng* will provide us with, apart from the inter-lingual results concerning the use of reporting verbs in translation of literary narration, some proved evidences of inter-translator agreement in the translational behavior in terms of the choice of such verbs.

3. Data Analysis

Given the short of automatic tools for such a study, a combination of adapted automatic computer processing and manual analysis serves as the method of investigation in this study. With respect to the sampling procedure, the WordSmith Tools (version 3.0) (Scott 1999) provide researchers with a query program to facilitate such data collection. Along with the EditPlus Text Editor (version 2.11) (2003), WordSmith is used to make KWIC (Keywords in Context) concordances of English reporting verbs in the texts under process. Apart from that, a specially-designed Chinese concordance program is used to process those reporting verbs in corresponding Chinese texts. For distributive analysis, descriptive statistics is used to capture the nature of distribution of reporting verbs in the sampled Chinese and English fiction texts concerned, and thus cross-tabulation results are provided with interpretation and discussion.

To facilitate our analysis, reporting verbs in the two parallel corpora will be classified, following Ardekani (2002), into four groups with some degree of fuzziness. Given the implicit nature of the Chinese verbs used, this classification applies more readily to the English verbs. Specifically, the groups are: (1) those concerned with the content of the reported speech, e.g., *argue*, *caution*, *continue*, *persuade*, *say*, *threaten*, *warn*, etc., (2) those related to the mode of utterance, such as, *cry*, *exclaim*, (*force a*) *smile*, *interpose*, *laugh*, *mumble*, *shout*, *whisper*, etc., (3) those pertaining to the nature of the utterance, e.g., *answer*, *ask*, *inquire*, *instruct*, *order*, *reply*, *scold*, *tell*, etc., and (4) implicit reporting verbs like: *nodded* ("*Yes*"), *frowned* ("*No*"), *seconded* ("*this*"), etc.

However, due to the great amount of data to be processed and the lack of automatic algorithms, the reporting verbs that are used in the SL corpora could not be matched at this stage on a one-to-one- or corresponding-basis with their respective equivalents in the TL corpora. To save the trouble of time-consuming manual checking, yet a global concordance processing is done to query the frequency of different kinds of reporting verbs used in the ST and TT respectively. Admittedly, this general, automatic processing may fail to identify a translator's individual preferences in transferring a SL reporting verb into the TL. It also may fall short of providing contextual evidence concerning both linguistic and pragmatic factors to be considered in Chinese-English (C-E) and English-Chinese (E-C) translation practice. However, this inadequacy necessitating future studies, this research is nevertheless justified, not only in terms of the insight that reporting verbs that are used in literary works be incorporated into NNS teaching materials. More significantly, the results of even such a coarse-grained comparison as these will also throw light on emerging corpus-based translation studies that are aimed at making intersubjectively testable generalizations of translation behaviour (Baker, 1993).

4. Results and Discussions

4.1 *Hong Lou Meng* and its English translations

Automatic processing of the *Hong Lou Meng* ST corpus shows that the Chinese novel untiringly exploits the monosyllabic verb道 (literally meaning "*tell*") as a blanket term for reporting. Compared with its total occurrences of 6431 times, the verbs说 ("*say*"), 说道 ("*speak*") and问 ("*ask*") with 1671, 927 and 358 instances respectively, can just be seen as the supplementary uses of道. However, a closer scrutiny of the context where道appears reveals that more than a dozen adverbials of manner precede the over-used reporting verb. Concordance analysis of the ST corpus lists the following more frequent modifiers of道:

Table 1: Occurrences of some modifiers of the verb道in *Hong Lou Meng*

Modifier	Literal Meaning	Occurrences	Modifier	Literal Meaning	Occurrences
笑	*laugh*	2213	问	*ask*	**288**
回	*return*	124	便	*thus*	**105**
冷笑	*Forced smile*	97	忙	*hurriedly*	**79**
叹	*sigh*	77	哭	*cry*	**49**
劝	*persuade*	47	骂	*curse*	**47**
啐	***spit***	**34**	**答**	***reply***	**29**

Further analysis of the way whereby说, 道and问are used tells us that some of the pre-modifiers listed above are also frequently resorted to in describing the manners by which the contents of the three verbs are reported. For instance, 说is found to be following便 ("*thus*") 158 times, 回 ("*return*") 89 times, 又 ("*again*") 81 times, and因 ("*therefore*") 57 times. By contrast, 问is used after因for 72 times, after便for 64 times, after忙 ("*hurriedly*") for 49 times, and after又for 43 times.

Turning to the two English versions of the Chinese novel, a different story is found with reference to the way the characters' utterances are reported. Basically, both the manner of "speaking" as reported by the ST with the help of pre-modifiers, and the concentration of reporting verbs on a limited number of lexis, are no where to be found. And this is regardless of a lack of ST-to-TT one-to-one correspondence analysis. As far as the former ST feature is concerned, not only much fewer adverbials of manner are used in both translations. Surprisingly, it is found that even these limited number of adverbs are used in a rather fixed manner, i.e., merely in association with a couple of reporting verbs. In Hawkes and Minford's version, more than a dozen adverbs ended with the suffix "*-ly*" are found used only together with the verb "smile"; and these include: *anxiously, coldly, courteously, deprecatingly, enigmatically, gratefully, ironically, modestly, mysteriously, nervously, patiently, sarcastically, unconcernedly*. In the Yangs' versions, two entirely different sets of words are used to describe how people in the novel show their opinions by a physical rather than verbal action. They may "*nod*" their *agreement, appreciation, approval,* and *consent*. Alternatively, they may "*nod*" *appreciatively, approvingly, repeatedly, thoughtfully,* or *in silence* or *sympathy*.

As far as the second ST feature is concerned, i.e., the repeated use of few reporting verbs with implicit nature, they are replaced with about twenty verbs that report in explicit terms. This seems to confirm Ardekani's (2002: 125) claim that "English and some other European languages specify the content of the report, the mode of utterance, the nature of the report, etc." However, this specification is done not by means of adverbials of manner, but rather by means of what Ardekani (ibid) calls "explanatory reporting verbs" per se. And this phenomenon goes hand in hand with Strunk and White's (1972: 68) advice, "Do not explain too much ... Be spearing, for instance, in the use of adverbs after "he said", "she replied", and the like: (he said consolingly; she replied grumblingly)."

While this is true of both English versions, their choices of "explanatory" reporting verbs do not actually converge. For purpose of comparison, these verbs are tabulated below in accordance with Ardekani's (2002) classification.

Table 2: Occurrences of reporting verbs in the two English versions of *Hong Lou Meng*

Verbs	Occurrences		Verbs	Occurrences	
	Hawkes	*Yangs*		*Hawkes*	*Yangs*
A. argue	8	7	**C.** answer	51	**241**
caution	1	8	ask	1015	**1384**
continue	293	97	inquire	50	**37**
say	5815	2131	reply	411	**285**
threaten	9	24	scold	9	**105**
warn	15	67	shout	84	**101**
B. cry	139	392	tell	886	**1104**
exclaim	137	257	**D.** frown	18	**9**
interpose	5	41	nod	310	**316**
laugh	385	264			
mumble	7	4			
smile	259	230			
whisper	**52**	**75**			

The table shows that, despite the differences in the total amount they actually employ, both Hawkes (and Minford) and the Yangs coincide with each other in the most frequent reporting verbs they choose in their translations. That is, they all use *say*, *ask* and *tell* as their most favorite choices to relate the nature of utterances reported. Therefore, we see the three verbs register the most instances of use in their respective works: 5815, 1015 and 886 in Hawkes as compared to 2131, 1384 and 1104 in the Yangs. This finding seems to suggest that both translators are bound, to certain degrees, by the ST's use of reporting verbs in the process of translation. This tendency can also be accounted for by some other frequent verbs they both choose, i.e., *laugh/smile*, *reply*, *nod*, etc., to explicate the mode of utterances reported.

4.2 *Tess of the D'urbervilles* and its Chinese translation

Coming to the English original of *Tess of the D'urbervilles*, concordance processing tells us that *say*, *ask* and *tell* also come on top of the ST author's list of reporting verbs in relating what his characters are conceived to have said. A comparison of Table 3 below with the most frequently used verbs in the two English versions of *Hong Lou Meng* as listed in Table 2 above seems to further confirm both Ardekani's claim and Strunk and White's advice. This can be seen not only from the three most frequent "implicit" verbs both the translators and the English writer choose, but also from the most frequent "explanatory" reporting verbs they use, such as *cry*, *reply* and *continue*. However, the ST writer of *Tess* rarely uses some of the "explanatory" reporting verbs which the *Hong Lou Meng* translators resort to. For instance, the verbs *laugh*, *smile* and *nod* which seem indispensable to the translators' explication of the speakers' verbal manner, do not seem popular with the English writer at all. The translators' reliance on these verbs might be explained on account of their bondage by the ST's over-use of such manners of speaking as specified by 笑/冷笑("*with a (cold) smile/laugh*"), 便 ("*thus*" or "*in accordance*"), and回 ("*in reply*"), etc. The English writer of *Tess*, however, should be at liberty in his choice of explanatory verbs.

Table 3: Occurrences of reporting verbs in the English version of *Tess*

Verbs	Occurrences	Verbs	Occurrences
A. argue	2	**C.** answer	**30**
caution		ask	**115**
continue	46	inquire	**13**
say	595	reply	**35**
threaten		scold	
warn		shout	**1**
B. cry	44	tell	**65**
exclaim	17	**D.** frown	**1**
interpose		nod	**2**
laugh	5		
mumble			
smile	2		
whisper	**24**		

With regard to the Chinese translation of *Tess*, it is found that the monosyllabic verb说constitutes the translators' unvaried option in "reporting". Altogether, it amounts to 1692 occurrences. Faced with this result, we tend to liken the verb with道used in the Chinese *Hong Lou Meng* since both rank as the single most frequent choice in the respective novels. While it is tempting to make this association, it would be wrong to assume that说is also heavily weighted with adverbs describing the various manners of speaking just as道is in the Chinese *Hong Lou Meng*. As shown in Table 4 below, explanatory adverbials are only used sparingly to the attributive verb in the translated *Tess*. At this we cannot help wondering: why is it that the Chinese literary language of *Tess* which is written two centuries later than that of *Hong Lou Meng* is less weighed with adverbs? Is it because of the fact that the language used is strictly based on that of its SL original?

Table 4: Occurrences of modifiers of the verb说in the Chinese version of *Tess*

Verb	Literal Meaning	Occurrences	Verb	Literal Meaning	Occurrences
回答	*reply*	51	嘟哝着	*mumble*	**12**
继续	*continue*	21	解释	*explain*	**8**
大声	*loudly*	18	温和地	*gently*	**5**
小声	*softly*	**6**	恳求	*plead*	**5**

5. Conclusion

Intra-corpus contrast of reporting verbs used in the Chinese novel *Hong Lou Meng* and their translations in two English versions found that, while Chinese literary narration is liable to weight reporting verbs with adverbs of manner, its English translation tends to replace the ST reporting verbs with explanatory reporting verbs. This accounts for the occurrences of more than a dozen reporting verbs in places where only less than half a dozen Chinese verbs are used for reporting in the ST. This means that, whereas various adverbs function to explicate the content, mode and nature of utterances in Chinese, explanatory verbs are employed to perform the function in its English translation. For instance, the Chinese original repeatedly loads reporting verbs with adverbs of 笑and回, its English versions rely on explanatory reporting verbs *laugh, smile* and *reply* to reciprocate the mode of utterance. While inter-corpus comparison of the reporting verbs used in the English versions of *Hong Lou Meng* and those used in the English *Tess of the D'urbervilles* again proved this finding, it is therefore advisable to take this linguistic difference into consideration in both the teaching and practice of C-E and E-C translation.

On the other hand, while intra-corpus analysis of the two English versions of *Hong Lou Meng* revealed the normative translational behaviour of different translators with reference to the use of reporting verbs, it also

showed some translatorial differences existing between the two versions. For instance, while Hawkes makes more use of *say* and *continue* than the Yangs, the latter make more use of *cry*, *answer*, *exclaim*, and *warn* than the former. This finding is significant in that, although the translators are all bound by the same SL text, they exercise their own freedom in their choices of reporting verbs.

Besides, it is also interesting to find that, although the Chinese translation of *Tess of the D'urbervilles* is written two hundred years later than the Chinese *Hong Lou Meng*, it makes use of both fewer reporting verbs and adverbs of manner than the latter. Since this study was conducted in a rush of time, it did not attempt to make a one-to-one comparison between a ST and its TT with reference to the translation of individual reporting verbs that are used in the ST, thus failing to provide contextual evidence as to either the correspondent reporting verbs that are used in the TT in replacement of their ST counterparts, or the actual translational behaviour of the relevant translators in the process of translating. It is thus hoped that future studies be done in these respects, so that more insights can be made in the emerging and promising interdiscipline of corpus-based translation studies.

References

Ardekani, M.A.M. 2002. The translation of reporting verbs in English and Persian. *Babel* 48 (2): 125-134.

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* (pp. 233-250). Philadelphia and Amsterdam: John Benjamins.

EditPlus Text Editor (version 2.11). 2003. Available at http://www.editplus.com.

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

Strunk Jr, W. and E.B. White. 1972. The Elements of Style. New York: Macmillan Publishing Co.

Swales, J. 1981. Aspects of Article Introductions. Research Report No.1. Birmingham: University of Aston.

Swales, J. 1987. Utilizing the literatures in teaching the research paper. TESOL Quarterly 21:41-68.

Swales, J. 1990. Genre Analysis. Cambridge: Cambridge University Press.

Tarone, E., S. Dwyer, S. Gillette, and V. Icke. 1981. On the use of the passive in two astrophysics journal papers. ESP Journal 1:123-40. Reprinted in Swales (ed.) 1985: Episodes in ESP (pp. 191-205). Oxford: Pergamon.

Thompson, G. and Ye Y. 1991. Evaluation in the reporting verbs used in academic papers. Applied Linguistics 12 (4): 365-382.

Collocational Characteristics in the Written English of Chinese University Students

Yuanwen Lu
National University of Singapore

Abstract: This paper carries out a corpus-based study of the use of "Verb + Noun" collocations in the written English of Chinese university students in order to reveal its collocational characteristics. It is based on two POS-tagged corpora: LOCNESS (the writing of British and American university students) and MLC (the writing of Mainland China university students). The "Verb + Noun" collocations in this study are divided into three sub-categories: "verb + noun", "verb + determiner + noun" and "verb + modifier + noun". Analysis of the two corpora shows that there is a big difference in the frequency of occurrence of "Verb + Noun" collocations between LOCNESS and MLC. The pattern used most frequently by Chinese students – "verb + modifier + noun" – is the least used in LOCNESS. Conversely, the most frequent pattern for British and American university students – "verb + determiner + noun" – is the least used in MLC. In terms of the total frequency of occurrence of the three patterns, Chinese university students employ such collocations in their English writing more than their Western peers do. The study also shows that Chinese students use the collocations with little variation and it is argued that they tend to repeatedly use the collocations with which they are most familiar. This familiarity may result either from the influence of the Chinese language or from the emphasis on these expressions in English instruction. If they attempt other varieties of collocation, the result tends to sound unidiomatic to native speakers. The implications of the present study for English language teaching and learning are discussed at the end of this paper.

Key words: collocation, written English of Chinese university students, POS-tagged corpus, "Verb + Noun" collocation.

Introduction

It is generally believed that the main task involved in learning English in People's Republic of China (henceforth referred to as China in this paper) is for students to master the grammatical rules and memorize the meaning of words. However, the reality is that their English proficiency does not improve significantly even after several years' study (at least six years in secondary school and two years in college). Apart from the very complexity of the grammatical rules, another hurdle is for students to produce not only grammatically correct sentences, but also idiomatic English in their writing. Their difficulty especially lies in the integration of grammar into lexis so that they can express themselves naturally and idiomatically in English, a naturalness which is typically represented by collocation.

Many linguists have noticed a typical error in foreign language learners' production, as pointed out by Allerton (1984):

> 'So often the patient language-learner is told by the native speaker that a particular sentence is perfectly good English...but that native speakers would never use it.' (Allerton, 1984: 39)

This kind of error is characterized by the grammatical but unidiomatic sentences produced by language learners, which can be best explained from the perspective of collocation. With the advent of learner corpora, many corpus-based studies on learner English have been carried out, some of which focus on collocation and error analysis. However, relatively few studies have been carried out on POS (part-of-speech) tagged corpora[1] (Aarts and Granger, 1998; Meunier, 1998). The advantage of exploring a POS tagged corpus lies in the grammatical information it provides. Syntactic patterns can be automatically extracted from a POS tagged corpus. It is therefore advisable to investigate collocation in combination with

[1] In a POS tagged corpus, each word is tagged with its part of speech. For example, the noun "book" is tagged as "book_NN1", which means it is a singular common noun. The "NN" stands for noun, the number 1 for singular form.

syntactic patterns, which is believed to yield a more complete picture of the use of lexical items in learner English.

This paper carries out a POS-tagged corpus-based study of the use of "Verb + Noun" collocations in the written English of Chinese university students in order to reveal its collocational characteristics. The "Verb + Noun" collocations in this study are divided into three sub-categories: "verb + noun", "verb + determiner + noun" and "verb + modifier + noun". It aims to answer the following questions: Is there any difference in the quantitative use of "Verb + Noun" collocations between the written English of British and American university students and that of Chinese university students in terms of the frequency of occurrence and the type/token ratio? What "Verb + Noun" collocations are most frequently used by Chinese students? What are the characteristics of these collocations?

I will begin with the definition of collocation in the context of the present study and the method used for the analysis, followed by the statistical results, then move on to both quantitative and qualitative analyses and discussions. Conclusion and pedagogical implications will be presented in the final section.

Defining Collocation

Collocation has been used and interpreted in various ways. For the purpose of this study, it refers to the word-combinations where one verb recurrently co-occurs with one or more nouns as the only choice or one of the few choices. These word-combinations are conventional and this commutability includes the following two types:

a) *Freedom of one component and some substitution on the other component, e.g.* "collect information", "collect stamps", "do business", "run business", etc.;

b) *Some substitution of both components, e.g.* "draw attention", "draw a conclusion", "pay attention", "reach a conclusion", etc.

Method

The present study is based on two tagged corpora: LOCNESS and MLC. The former is a corpus of British and American university students' writing, the native speaker component of the International Corpus of Learner English (ICLE[1]). It consists of essays written by both British and American university students for their assignments and examinations. MLC is a corpus of non-English major Mainland Chinese university students' writing, one of the sub-corpora of Chinese Learner English Corpus (CLEC). It consists of selected compositions written by Chinese university students for their College English Test Band-4 (CET-4) and Band-6 (CET-6). For the sake of convenience, the terms "English writing of Chinese university students", "written English of Chinese university students" and "Chinese learner English" will be used interchangeably in this paper to refer to the same corpus. "Chinese university students" and "Chinese learners" will also be used interchangeably to refer to the students whose essays comprise MLC.

CLAWS4, a POS tagger[2], is used in the present study to tag the two corpora, LOCNESS and MLC. The tagging accuracy rate is 99.3% for LOCNESS, and 98.1% for MLC. Since the wrongly tagged words have also been manually corrected in the course of identifying collocations, it is therefore reasonable to assume that the errors in tagging are not significant enough to affect the statistical findings for the study.

After the POS tagging of the two corpora, the software WordSmith Tools[3] is used to automatically extract the three patterns of "Verb + Noun". The tokens are the total number of collocations identified manually from the word-combinations for each collocational pattern in both LOCNESS and MLC. The types are the total number of the collocation types comprising each collocational pattern. All the variations within a collocation (such as number, tense, and determiner before nouns) belong to the same type. In order to

[1] More information on ICLE is available at: http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html
In the present study, only essays of British and American university students in LOCNESS are analysed for the sake of comparability with the English writing of Chinese university students.

[2] A POS (part-of-speech) tagger is a piece of software used to attach each word in a text with a POS mark. More information on CLAWS4 is available at: http://www.comp.lancs.ac.uk/computing/research/ucrel/claws

[3] More information on WordSmith is available at: http://www.lexically.net/wordsmith

differentiate the type/token ration in the present study from the general type/token ratio, the ratio for each collocational pattern in the present study will be tagged as "TTRC", i.e. collocational type/token ratio. Due to the different sizes of LOCNESS and MLC, the total number of collocations identified for each pattern is normalised to 10,000 words. For the same reason, the logarithmic type/token ratio is preferred to the type/token ratio, since this log value "will remain constant" for samples of different sizes (Granger and Rayson, 1998: 121; De Cock, *et al.*, 1998: 72). Therefore, the logarithmic TTRC will be used in the present study.

Statistical Results

The normalised numbers of "Verb + Noun" collocations are presented in Table 1.

Table 1: Number of "Verb + Noun" Collocations per 10,000 Words

Lexical Pattern	LOCNESS	MLC
verb + noun	20	33
verb + det + noun	22	32
verb + mod + noun	19	39
Total	61	104

("det" stands for determiner; "mod" stands for modifier)

The log TTRCs of "Verb + Noun" collocations are presented in Table 2.

Table 2: Log TTRC of "Verb + Noun" Collocations

Lexical Pattern	LOCNESS (%)	MLC (%)
verb + noun	87.9	74.5
verb + det + noun	91.4	79.6
verb + mod + noun	92.7	76.1

Quantitative Analysis and Discussion

Taken together, Tables 1 and 2 show that Chinese university students use "Verb + Noun" collocations more than their Western peers do (61 in LOCNESS *vis-à-vis* 104 in MLC), but their usage lacks variety. A careful examination of Table 1 reveals an interesting phenomenon. The most frequently used pattern by Chinese students is "verb + modifier + noun", but this is the least used pattern in LOCNESS (39 in MLC *vis-à-vis* 19 in LOCNESS per 10,000 words). The most frequently used pattern by British and American university students – "verb + determiner + noun" – is the least used in MLC (22 in LOCNESS *vis-à-vis* 32 in MLC per 10,000 words). It would be premature to draw any conclusions about the use of determiners and modifiers only from the above figures, but it is worth investigating in future study, for example, whether this is due to the omission of determiners or the overuse of a certain type of modifiers in Chinese learner English.

Although native speakers have a much wider range of "Verb + Noun" choices, this does not necessarily mean that they actually use many difficult verbs and nouns in their writing. In fact, most of the verbs and nouns native speakers use are listed in the Chinese College English Syllabus-Vocabulary (2000, henceforth CES-V). For example, native speakers use "pose/raise/ask question(s)" in their writing; however, there is only one type appearing in MLC, that is, "ask...question(s)". MLC has no instances of "pose...question(s)" and "raise...question(s)", even though these are also simple and idiomatic collocations.

The reason for the smaller number of "Verb + Noun" collocations in LOCNESS (61 in LOCNESS *vis-à-vis* 104 in MLC) is that as alternatives to such collocations, native speakers also employ other types of expressions such as prepositional phrases and part-of-speech transformation to refer to the same meanings. For example, "lose/losing/lost...job(s)" occurs 12 times in MLC. "Fire" used as a verb with the similar meaning appears 3 times in MLC. While in LOCNESS, apart from "lose" and "fire", it is found that "out of a job/jobs" (4) and "out of work" (2) are also used to express similar meanings. Another example is the word "question". It can be used as either a noun or a verb. In LOCNESS, there are two occurrences of "question" used as a verb, but there is no such use in MLC. Consider the following concordance from LOCNESS:

I erm looking European citizen will also *beg in to* **question** whether sovereignty is a valuable com

2 ndividual ones as we currently have. We *have to* **question** whether independant action, such as B

The lack of much variation in MLC results from the frequent use of some "Verb + Noun" collocations. Chinese students have a collocational preference to use the words which may have a Chinese translation. For instance, "lose...job" can be translated from Chinese "*shi qu gong zuo*" (失去工作) or "*shi ye*" (失业); "ask...question" into "*wen wen ti*" (问问题) or "*ti wen*" (提问), etc. These are the words with which they become most actively familiar, simply because Chinese university students have formed the prototypes of these expressions in their mind since the beginning of their English study.

In the quantitative analysis of "Verb + Noun" collocations, it is also found that the noun "attention" appears 188 times in MLC. The most frequent verbs it collocates with are "pay" (137) and "paid" (18); others are "take" (4), "focus" (3), "put" (3), "draw" (3), "attract" (3), and "receive" (2). No matter what other verbs "attention" collocates with, it is evident that "pay/paid" (155 occurrences) takes on an overwhelmingly higher proportion – 82.45% – than others do in MLC. This again illustrates the point that repeated use of some collocations can lead to both a large number of "Verb + Noun" collocations and a lower log TTRC of "Verb + Noun" collocations in Chinese students' writing. On the other hand, the frequent use of "pay attention to" may result from its emphasis in English instruction. This can be seen from the fact that "focus attention on" and "concentrate attention on" can be literally translated from Chinese "*ji zhong zhu yi li*" (集中注意力), but the former is not used much, and the latter does not occur in the written English of Chinese university students.

Qualitative Analysis and Discussion

In addition to the less diversity of "Verb + Noun" collocations in MLC, Chinese university students tend to coin some "collocations" in their writing such as *"*pay attentions to*", *"*provide medical conditions*", *"*join social action*", etc. Let us take "pay attentions to" as an example. "Pay attentions to" occurs 4 times in MLC, apparently with the same intended meaning as that of "pay attention to". Consider the following concordance taken from MLC:

1. times the GNP in 1960, the govement pay a lot attentions to the health of the people.
2. re…With a word, the health has been payed more attentions than time to live with their children
3. f the economy, more and more people pay their attentions to health. Of course, other factors
4. war. Second, the developing countries pay great attentions to the people health. If one people

It is the word "attentions" whose lexical behaviour is worth examining. According to the Collins Cobuild English Dictionary (1995, henceforth CCED), the plural noun "attentions" refers to 'someone's efforts to help you. Or the interest they show in you, ... especially if you dislike or disapprove of them.' It can be seen that, in most cases, the semantic prosody of "attentions" is negative. The instances in the CCED support this point:

> The only way to escape the unwanted attentions of the local men was not to go out...
> The meeting was held away from the attentions of the media...
> Some men are flattered by the attentions of a young woman.
>
> - [CCED, 1995: 96]

These instances show that the most frequent verbs with which "attentions" collocates are "escape", "flatter", etc. Obviously, Chinese university students do not realize the difference between "attention" and "attentions". Furthermore, the expression of *"*pay attentions to*" also implies that the grammatical change in a collocation has not drawn enough attention from Chinese learners and they tend to treat the components in a collocation as separate units, not as a whole. Nation (2001: 329-331) points out that "grammatical fossilization" is one of the scales indicating what is involved in learning collocations. This scale ranges from "no grammatical variation" to "changes in part of speech", with "inflectional change" as a mid-point (ibid: 333).

Another example is the noun "society/societies", which is one of the most frequently appearing nouns in MLC. This could be attributed to the topics of the compositions in MLC, since most of them are common social phenomena. What deserves special attention here are the verbs with which "society/society" collocates. The most frequent verbs "society/society" collocates with are "know", "learn", "serve", etc. However, the

expressions of *"*know the society*", *"*learn the society*", *"*touch the society*" and *"*understand the society*" sound quite odd to native speakers, especially *"*touch the society/societies*", with the intended meaning of "start to work and live in the adult world". These expressions show that the Chinese language is a strong possibility for these expressions, which may be literally translated from Chinese.

Conclusion and Pedagogical Implications

The results of the above systematic analysis of "Verb + Noun" collocations reveal that one of the major problems in the written English of Chinese university students is that they use collocations with considerably less variety. If they attempt other varieties, the result tends to sound unidiomatic to native speakers, as shown in both quantitative and qualitative analyses in the present study. In order to tackle these problems, some suggestions on English language teaching and learning are made below.

(1) *Focus on unacceptable collocations to raise awareness of English collocations and explicitly provide students with the typical English ones.* For example, *"*touch the society*" is one of the most common mis-collocations found in MLC. It is more helpful to explain why it is unacceptable than just to let them know it is wrong. Teachers can tell them that, for instance, when the verb "touch" is used with the meaning of "*jie chu*" (接触) in Chinese, it normally refers to "feel somebody or something physically" and is followed by concrete objects such as "skin", "face", "painting", "her", etc. It is interesting to note that the meaning of "touch" in the expression "touch her" depends on its subject. Considering the following sentences: "His story touched her". "Touch" in this sentence means "affect her feelings". It is therefore suggested that typical uses should be highlighted and taught explicitly in English instruction to raise learners' awareness of English collocations (Lu, 2002). However, students should be encouraged to use them with more variety in their writing, rather than focus on a particular one and cause its overuse.

(2) *Make as many collocations of a word as possible.* In order to have more collocations at hand, students should be encouraged to make as many collocations of a word as possible. This can be done by comparing concordances of language production between native speakers and learners. This activity helps not only to spot students' problems in using certain words and expressions (for example, repeated use of particular collocations with little variation), but also to overcome them. For instance, apart from the most common collocation "lose a job", other alternative collocations such as "out of a job" and "out of work" are most likely to be found in a concordance produced by native speakers, which will certainly expand students' collocational knowledge.

In addition, one of the advantages of exploring the collocational field is to help learners distinguish the 'supposed difference of certain words such as synonyms, etc.' (Hill, 2000: 61). They should be encouraged to discover collocational similarities and differences between L1 and L2 in these collocations. For example, whether or not students really know the difference between "question" and "problem" can be seen from the collocations they make using the two words. It would be reasonable to believe that they have known the difference if they can make the following collocations:

question: "raise a question", "pose a question", "ask a question", "answer a question", "reply to a question", etc.

problem: "solve a problem", "deal with a problem", "tackle a problem", "cause a problem", "create a problem", "pose a problem", etc.

(3) *Teach collocation with reference to the Chinese language.* This suggestion may sound unreasonable to some EFL practitioners. In fact, as the above analysis shows, the Chinese language plays an important role in the English writing of Chinese university students. Although Chinese students write essays in English, what they are talking about is their everyday life and events in China. It is therefore advisable for teachers to let students know the typical English expressions for the frequent Chinese ones. Meanwhile, it is also important for Chinese learners to realize the difference between L1 collocations and L2 collocations. The Chinese collocations do not necessarily have their corresponding ones in English. In this way, errors in the English production of Chinese students would be greatly reduced. For example, "接触社会" (*jie chu she hui*) is one of the most frequent collocations in the English writing of Chinese university students. However, there is no its corresponding English collocation. Chinese learners have to paraphrase its meaning in their writing.

Clearly, there are many suggestions on how to tackle the problems in the English writing of Chinese university students and how to teach collocations in the setting of China. The above suggestions are highly recommended to EFL teachers in China.

Acknowledgements

This paper is largely based on my PhD thesis; I am deeply indebted to my supervisors Dr Vincent Ooi and Dr Rosemary Khoo for their invaluable comments and suggestions. I am also grateful to Professor Yang Huizhong at Shanghai Jiao Tong University for permission to use MLC, a sub-corpus of the Chinese learner English corpus (CLEC).

References

Aarts, J. and Granger, S. 1998. Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In Granger, S. (ed.), *Learner English on Computer*: 132-141. London: Longman.

Allerton, D. J. 1984. Three (or Four) Levels of Word Cooccurrence Restriction. *Lingua*, 63: 17-40.

Benson, M., Benson, E., and Ilson, R. (revised edition). 1997. *The BBI Dictionary of English Word Combinations*. Amsterdam: John Benjamins.

College English Syllabus – Vocabulary. 2000. Beijing: Higher Education Press.

Collins Cobuild English Dictionary. 1995. London: Collins.

De Cock, S., Granger, S., Leech, G. and McEnery, T. 1998. An automated approach to the phrasicon of EFL learners. In Granger, S. (ed.), *Learner English on Computer*: 67-79. London: Longman.

de Haan, P. 1997. An experiment in English learner data analysis. In Aarts, J., de Monnink, I. and Wekker, H. (eds.), *Studies in English Language and Teaching: in honour of Flor Aarts*: 215-229. Amsterdam: Rodopi.

Garside, R. and Smith, N. 1997. A Hybrid Grammatical Tagger: CLAWS4. In Garside, R., Leech, G., and McEnery, T. (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*: 102-121. Harlow: Addison Wesley Longman.

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

Granger, S. 1998b. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In Cowie, A. P. (ed.), *Phraseology: theory, analysis, and applications*: 145-160. Oxford: Oxford University Press

Granger, S. and Rayson, P. 1998. Automatic profiling of learner texts. In Granger, S (ed.), *Learner English on Computer*: 119-131. London: Longman.

Hill, J. 2000. Revising priorities: from grammatical failure to collocational success. In Lewis, Michael. (ed.). *Teaching Collocation*: 10-27. Hove: Language Teaching Publications.

Howarth, P. 1998. The Phraseology of Learners' Academic Writing. In Cowie, A. P. (ed.), *Phraseology: theory, analysis, and applications*: 161-186. Oxford: Oxford University Press.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Lewis, M. (ed.). 2000. *Teaching Collocation: Further Developments in the lexical approach*. Hove: Language Teaching Publications.

Lu, Y. 2002. Linguistic Characteristics in Chinese Learner English. In Tan, M. (ed.) *Corpus Studies and Language Education*. Bangkok: IELE Press.

Meunier, F. 1998. Computer tools for the analysis of learner corpora. In Granger, S. (ed.), *Learner English on Computer*: 19-37. London: Longman.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Sinclair, J. 1996. The Search for Units of Meaning. *Textus*, IX: 75-106.

A Corpus-Based Analysis of Connectors in Non-English Major Graduate Students' Writing

Pan Fan Feng Yuejin

Huazhong University of Science and Technology

Abstract: Until now, most corpus-based empirical work on learner language has mainly concentrated on high school students and undergraduates. Advanced learners like non-English major graduate students haven't received enough attention from researchers. To fully understand the development of interlanguage of learners, it is necessary to study advanced learners. To bridge this gap, we have built a 120,000 non-English major graduate students' writing corpus (Master Writing Corpus) and investigated the connectors in the corpus both quantitatively and qualitativly. A contrastive approach is used on the basis of the comparison between a native speaker corpus and the learner corpus. The findings are interesting and surprising. First, advanced Chinese learners tend to underuse all connectors in their English writings. Second, while they share similar preferences for connectors in comparison with native speakers, they tend to overuse and underuse certain connectors. Third, the phenomena of overuse and underuse may also be found at the semantic level. Advanced Chinese learners and native speakers differ greatly in using the same connectors for serving different logical-semantic functions. The implications to English teaching in China and suggestions for future research are also discussed.

Key words: corpus-based; connectors; overuse; underuse; frequency

1. Introduction

In the West, there has been a sharp rise in the number of learner corpora and studies based on learner corpora since the 1990's. Corpus-based investigation into learner language has also attracted wide attention. The description of learner language has always been of primary concern to second language acquisition (SLA) researchers. Learner language provides the researcher with insights into the process of acquisition. If we have a better understanding of the second language (L2) acquisition process, then we can apply the findings to a variety of practical aspects of language teaching: syllabus design, material design, task design, testing, and so on.

Previous work on leaner has mainly concentrated on the adult beginners learning English as a foreign language. These studies are usually limited in scope, as they focus on a very limited number of learners, usually high school students and undergraduates. There have been comparatively few studies involving advanced adult learners. Until now, scanty attention has been paid to non-English major graduate students in corpus-based learner study home and abroad. In fact, non-English major graduate students in China are required to pass CET 6 (Coliege English Test Band 6, an officially held English Proficiency Test), which shows that they represent the highest English level of Chinese non-English major students to a certain degree. In this sense, if the development of learners' interlanguage is seen as a chain (from beginner level to advanced level), then graduate students will be the end of the chain. Investigation into their interlanguage helps to understand the developmental process of learner interlanguage. To bridge this gap, we built a small non-English major graduate students' writing corpus and made an attempt to find out the linguistic features in advanced Chinese English learners' writing.

In this study we will be concerned with the use of connectors by advanced Chinese EFL learners. Our experience tells us that even advanced Chinese learners tend to misuse as well as underuse connectors.

2. Research aim and method

We have gained an intuitive impression from years of teaching experience that the use of connectors is problematic for Chinese language learners. It is a common phenomenon that English writings of students

usually look disconnected and lack of coherence. As we know, effective communication (including both reading and writing) requires coherence and clarity. One way of achieving this is to signal logical or semantic relations between units of discourse by means of connectors such as *but* (to indicate a contrast), *because* (reason), *therefore* (result), *in addition* (exemplification), etc (Altenberg & Tapper, 1998). Connectors can be said to function as cohesive "signposts" in discourse, helping the listener or reader to relate successive units to each other and thus making sense of the text. A number of studies have shown that the use of connectors is problematic for language learners, in particular foreign language learners (e.g. Granger, 1994). Therefore, it may be proposed that it's just because of the underuse and misuse of connectors that reduces the comprehensibility of learners English writings.

Considering that using inadequate connectors and being unable to use connectors properly are two typical errors in Chinese students' writings, the paper intends to investigate the use of connectors in this learner's writing corpus by adopting both quantitative and qualitative approach. That means we are not only concerned with the quantitative aspects of connector usage but also its qualitative aspects--how connectors are actually used by the learners in comparison with native writers. In this sense, the approach adopted here is also contrastive: the analysis is based on a comparison of a corpus of non-native learner and a corpus of native speaker. We will try to answer the following questions:

1) Do advanced Chinese learners use connectors to the same extent as native English speakers?

2) Do they use them to express the same semantic relations as native speakers?

To achieve this goal, the research is designed as follows. First, twenty connectors are investigated in terms of relative frequency to answer the first question. Second, *And*, as a special form of *and*, is chosen as an example to answer the second question. The distribution of semantic relations indicated by *And* in two corpora is studied. Overuse and underuse are two key words in our research. They are expected to occur at these two levels of investigation.

The learner corpus used in the study is called Master Writing Corpus (120,000 words), developed at Huazhong University of Science and Technology (HUST). The Master Writing Corpus consists of 831 English essays written by non-English major graduate students in HUST. All the essays are argumentative in character. That means, besides presenting facts, the essays also have the aim to explain, analyze and interpret these facts and, usually, to argue for a certain standpoint. That will involve a great need of using connectors in their writings. The writers are all university graduate students and most of them are over 22. They can be regarded as "advanced learners" in the sense that they have finished at least 10 years of English study and they are studying for their Master degree.

3. Comparison in terms of frequency

3.1 Overall relative frequency

Let us first look at the overall frequency of connectors in two corpora (Table 1). Since the two corpora differ in size the relative frequency is given here (e.g.1:41 means that there will be an *and* every 41 words). The second and the third column offer the relative frequency of connectors used by Chinese learners and native speakers respectively. The fourth column offers the frequency ratio between Chinese learners and native speakers.

Table 1: Overall quantitative comparison between Chinese learners and Native speakers

	Chinese learner	Native speaker	Ratio
and	1:592	1:41	1:14.4
because	1:8,390	1:992	1:8.46
but	1:2,683	1:292	1:9.2
for example	1:37,995	1:2,559	1:14.85
for instance	1:379,951	1:10,763	1:35.30
furthermore	1:120,893	1:25,623	1:4.72
however	1:17,049	1:1,456	1:11.71
in addition	1:60,446	1:7,514	1:8.04
in fact	1:46,660	1:6,359	1:7.34

in other words	1:265,966	1:26,004	1:10.23
indeed	1:98,505	1:5,426	1:18.15
nevertheless	1:531,932	1:17,927	1:29.67
not only... but also	1:28,294	1:5,895	1:4.80
on the other hand	1:32,044	1:12,244	1:2.62
so	1:981	1:511	1:1.92
then	1:17,270	1:922	1:18.73
therefore	1:50,182	1:4,886	1:10.27
though	1:32,434	1:2,279	1:14.23
thus	1:69,991	1:3,260	1:21.47
while	1:27,996	1:1,659	1:16.88

It can be seen that the Chinese learners use all connectors much less frequently in their essays than the native speakers (e.g. 1:592 vs.1: 41). That is, in the sense of overall relative frequency, Chinese learners tend to underuse all connectors in comparison with native speakers and no overuse of connectors is found. The frequency ratio between Chinese learners and native speakers ranges from 1.92 to 35.30. Two extremes are *so* and *for instance*. A statistical interpretation is that Chinese learners and native speakers are similar in using *so*, but differ a lot in using *for instance*. Native speakers use *for instance* as many as 35 times that of Chinese learners. For a better understanding of the phenomenon as underuse of connectors by Chinese learners, the connectors are classified into different groups according to their ratio of frequency (Table 2).

Table 2: List of underused connectors by Chinese learners and ratio

Ratio	Connectors	Sum
1-10 times	because, but, furthermore, in addition, in fact, not only...but also, on the other hand, so	8
10- 20 times	and, for example, however, in other words, indeed, then, therefore, though, while	9
20-30 times	nevertheless, thus	2
30-40 times	for instance	1

Table 2 provides a clearer picture of the frequency differences in using connectors by native speakers and Chinese learners (statistically speaking). This supports our impression that advanced Chinese learners of English tend to underuse connectors in the English.

Two factors may explain the significant differences between Chinese learners and native speakers in terms of relative frequency. First is the negative transfer from the learners' mother language to the target language. In Chinese, cohesion is emphasized and achieved mainly by the internal structures of sentences via rhetoric devices such as parallelism and contrast. Invisible logic is hidden in lines and understood in its context. Nevertheless, coherence and clarity are emphasized in English and achieved by lexical devices like connectors. The logic between lines is clear and easy to follow. Second factor is related to the cultural differences between two types of speakers. Chinese people tend to talk in an indirect and vague way. Listeners or readers are supposed to extract enough information from the context and understand by themselves. So Chinese is a high-context language in which clarity and logic are not necessarily required. On the contrary, English is a low context language. Readers or listeners are supposed to know little background information about the topic. Therefore, massive information is usually offered in a clear and logical way to them. That's why various types of connectors are widely used in English. For these two reasons, Chinese learners use significantly fewer connectors than native speakers in English.

3.2 Top twenty connectors

The investigation into the overall relative frequency of connectors by two types of writers focus only on answering the question: do advanced Chinese learner use connectors as frequently as native speakers? Naturally, another question arises: do they have same preferences for connectors as native speakers in writings? In other words, do they prefer some connectors to others like native speakers do? To make this clear, the rank order of twenty connectors studied in Master writing Corpus and the native speaker corpus is arranged in Table 3 according to the frequency list in table 1.

Table 3: Top twenty connectors in two corpora

	Chinese learner	Native speaker
1	and	and
2	so	but
3	but	so
4	because	then
5	however	because
6	then	however
7	while	while
8	not only, but also	though
9	on the other hand	for example
10	though	thus
11	for example	therefore
12	in fact	indeed
13	therefore	not only, but also
14	in addition	in fact
15	thus	in addition
16	indeed	for instance
17	furthermore	on the other hand
18	in other words	nevertheless
19	for instance	furthermore
20	nevertheless	in other words

As we can see from Table 3, the rank order of most connectors listed in two columns doesn't show great differences as expected. In the top ten connectors, eight connectors are identical except *not only... but also* and *on the other hand*. This means that both the Chinese learners and the native speakers rely heavily on roughly the same connectors. The strong reliance on *and, but* and so as connectors by both native speakers and Chinese learners is especially striking: *and* ranks first in both corpora. Yet, despite this general tendency to share a common set of preferred connectors, there are some notable differences between the corpora. If we concentrate on the items that differ in rank order, we find that two connectors *not only...but also, on the other hand* are significantly overused by the Chinese learners, while *thus, indeed* are significantly underused. *Furthermore, in other words* are slightly overused while *for instance* is slightly underused. Looking back to our teaching practice, we will find that *not only...but also, on the other hand, furthermore, in other words* are usually emphasized and encouraged by most teachers of English in China while they prefer to replace *thus* with *so, indeed* with *in fact, for instance* with *for example*. We may conclude that learners' preferences for choosing connectors are greatly influenced by English teaching practices.

The finding that advanced Chinese learners and native speakers don't differ greatly in the preferences for connectors seems surprising because it is inconsistent with Granger's findings (1994). In her study, great differences were found between learners and native speakers in terms of their preferences for connectors. The top ten list of native speakers is quite different from that of learners. The phenomena as overuse and underusee of certain connectors seem striking in her study.

Two reasons may account for the inconsistency in our study and her study. First, learners are from different backgrounds. The influence from different mother languages may lead to the differences in learner language (that may also confirm the influence from mother languages on learners). The second reason is related to the differences in the English levels of learners. As they are graduate students, the advanced Chinese learners are required to read authentic English materials more extensively. More chances to be exposed to an English environment may have put them more frequently under the influence of real English. Thus, they are more likely to follow a similar preference order for connectors as native speakers when writing in English. But limited by their English proficiency level, they are still not familiar with using appropriate connectors to achieve clarity and coherence.

4. Comparison in terms of semantic relations

4.1 Classifications of semantic relations of *And*

The above analysis has provided a general description of the overuse and underuse of connectors by Chinese learners. Then comes the second question: to what extent Chinese learners use connectors to mark

the same semantic relations as native learners? In other words, do the phenomena as overuse and underuse also exist in the semantic functions expressed by the connectors? To answer the question, we now turn our attention to the use of individual connectors in the material, then "*And*"(at the sentence beginning) is chosen as one example. The classifications of semantic relations of *And* are given here.

a. **Additive relation**: giving further illustrations on the basis of last sentence.
b. **Resultive relation**: introducing cause-effect relationship or antecedent-consequence relationship between sentences.
c. **Adversative relation**: introducing a different topic or a different point. It equals to *however* or *but* here.
d. **Contrastive relation**: introducing a comparison or a contrast.
e. **Progressive relation**: indicate the progressive move towards the topic or the action after another action.
f. **Introducing relation**: introducing another topic, usually followed by a question (informal).
g. **Explanatory relation**: provide explanations or reasons for the facts in last sentence.
h. **Listing relation**: listing two or more facts or examples to illustrate the same topic.
i. **Summative relation**: summarize the above illustrations or make comments or draw a conclusion.

4.2 Comparison of the distribution by semantic relations

There are 3004 *And* in the native speaker corpus and 51 And in Master Writing corpus. Sampling method is used here before analyzing the logical-semantic relations of *And* in the native speaker corpus. One concordancing line is extracted every ten lines. 300 *And* samples are extracted and then classified into groups according to their semantic relations. 51 *And* in Master Wring Corpus are all analyzed. The results are as follows. (Table 4) .

Table 4: Distribution of semantic types of connectors

	Native speaker Frequency	Percentage	Chinese learner Frequency	Percentage
additive	21	7%	1	2.0%
resultive	27	9%	10	19.6%
adversative	33	11%	6	11.8%
contrastive	18	6%	1	2.0%
progressive	27	9%	12	23.5%
introducing	60	20%	5	9.8%
explanatory	15	5%	3	5.9%
listing	51	17%	8	15.7%
summative	48	16%	5	9.7%
non-logical	0	0	0	0
sum	300	100%	51	100%

As shown in Table 4, the distribution of the different semantic categories in two corpora is quite different from each other. In the native speaker corpus, introducing and listing relations are most common, contrastive and explanatory relations are rare. But in Master Writing Corpus, resultive and progessive relations are most common, additive and contrastive relations are rare.

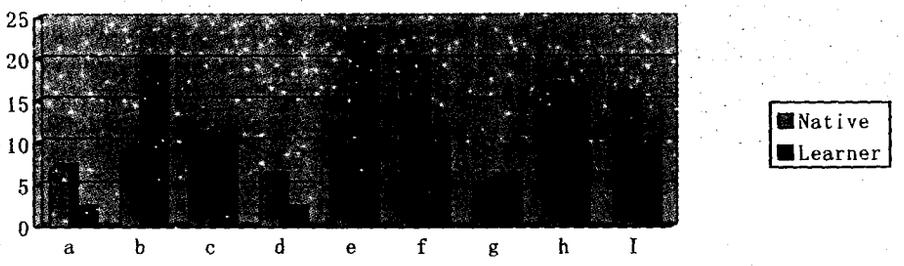


Diagram 1: Distribution of semantic types of connectors
(Notes: a-additive; b-resultive; c-adversative; d-contrastive; e-progressive; f-introducing; g- explanatory; h-listing; i-summative)

The differences between the corpora are statistically significant. The Chinese learners tend to overuse *And* to

express resultive (9% vs. 19.6%) and progressive relations (9% vs. 23.5%) than the native speakers but underuse *And* to express introducing relation (20% vs.9.8%). The native speakers strong reliance on *And* as introducing connector and Chinese learners strong reliance on *And* as progressive connector are especially striking.

Table 5: Semantic relations underrused and overused by Chinese learners

	Semantic relations
overused	resultive, progressive
underused	additive, contrastive, introducing, summative

Another finding is that all *And* indicate a sort of logical relation in Chinese learners writings (Table 4). At first sight, this finding seem to contradict the finding made by Chen (2001) in his study on *And* usage of high school students. He found that 27.5% *And* used by Chinese learners doses not serve any logical functions in the sentences. If two findings are put together, the inconsistency diminishes. On one hand, 27.5% *And* used by Chinese beginners is non-logical; on the other hand, 100% *And* used by advanced Chinese learners indicates logical–semantic relations. This fits the fact that the learner's capacity to use connectors increases with his language competence. It also denies the possibility that their mother tongue have influence on their use of non-logical *And* and confirms the necessity of emphasizing text-based teaching in high school. Text-based teaching may help the students improve their consciousness of "text" and clarify the logical relations in their writings.

5. Conclusion

The main conclusions of this study are as follows. First, advanced Chinese learners tend to underuse all connectors in their English writings, which confirms our own impression of Chinese students' essay writing. Second, while they share similar preferences for connectors in comparison with native speakers, advanced Chinese learners tend to overuse and underuse some connectors. Third, the phenomena of overuse and underuse also exist at the semantic level. Advanced Chinese learners and native speakers differ in using the same connector for serving different logical-semantic functions.

Though these conclusions are very tentative, they also have some implications to English teaching in China. Underuse and overuse of connectors and semantic relations may partly account for why Chinese learners cannot write idiomatic English writings. It seems that it's necessary for English teachers to lead them to use connectors according to the "real picture" of connectors (how the connectors are actually used by native speakers), especially those connectors and semantic relations that are not familiar to them. Learners can be frequently given specially designed exercises emphasizing and clarifying the roles of various connectors and their logic-semantic functions. As much exposure as possible to native language should be emphasized in teaching practices as well, which will help learners to approach and get familiar with the real language.

The findings also shed some light on the development of interlanguage of Chinese language learners. The differences between Chinese beginners and advanced Chinese learners will tell the beginning and ending of the same developmental process. They help eliminate the influence from the mother tongue and reveal what may be achieved by optimizing English teaching practice. In future research, two things can be done to push the present study further. One is to base the analysis on another comparable advanced native learner writing corpus instead of the "expert corpus" used in the present study. The other is to make a series of comparisons between Master Writing Corpus and other comparable writing corpora written by high school students and undergraduate students. More valuable and convincing findings will be got from the research practices.

References

Altenberg Bengt , Tapper Marie. 1998. The use of adverbial connectors in advanced Swedish Learners' written English. In *Learner English on Computer.* Edited by Sylviane Granger. 81-92.

Chen Rongxin. 2001. The study of logical connectors And in corpus and implications to EFL teaching. *Fujian Foreign Languages.* Vol.69.Issue 3.39-43. (Translated from Chinese).

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

'Small-words' in EFL Learners' Spoken Corpora

He Anping

South China Normal University

Abstract: This study investigates a group of discourse markers named 'Small-words' in the corpus of EFL learners' spoken English. It is based on four sub-corpora of the *International Corpora of EFL Learners' Spoken English* (LINSEI), namely LINSEI-China, LINSEI-Japan, LINSEI-France and LINSEI-Italy. The research addresses two issues:

1. How EFL learners of the four countries use Small-words differently in comparison with native English speakers? This relates to an investigation of Small words in terms of its type range, frequency and discoursal / interactional functions.
2. How is the learners' use of small-word associated with their fluency in spoken English performance? This relates to a comparison between more fluent and less fluent Chinese EFL learners in terms of their degree of fluency (i.e., the timing index including speech rate, filled/unfilled pause and mean length rate) with their use of Small-words (i.e., type range and frequency).

As the data-driven analysis goes on, some other salient oral English features of the EFL learners, Chinese EFL learners in particular, are also found. The findings will highlight the functions of Small-words in developing speaking fluency and implication for EFL teaching and oral testing.

Key words: Spoken corpus, small words, fluency, disfluency

What is Small words?

This study investigates a group of discourse markers named 'Small-words' (SW) in EFL learners' spoken English interlanguage. SW are a group of 'words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself (Hasselgren,1998: 155). The most typical SW includes *well, sort of, you know, I mean, I think*, etc.(for more detail see Table 2) . A number of scholars abroad have studied SW in its components as well as functions, including:

> Dawley & Syder (1983) who regards such words as 'lexicalized sentence stems and other memorized strings that form the main building blocks of fluent connected speech.'
>
> Bygate (1987) who highlights the functions of these words as 'a stock of devices for facilitating speech routines for structuring speech and procedures for negotiating meaning.'
>
> Sinclair (1991) who proposes SW to be semi-preconstructed phrases that constitute a single choice.
>
> Nattinger & De Carrico (1992) who identifies such subgroup of lexical phrases as 'discourse devices' with further subcategory of fluency devices.
>
> Stenström (1994) who gives inventories of 'interactional signals (e.g., *well, I mean, you know*)' which play a crucial role in smooth interaction and 'discourse markers (e.g., *right, well, anyway*) which help the speaker organize the discourse.
>
> (cited from Hasselgren,1998)

The theoretical base for SW study can be Sperber and Wilson's (1995) relevance theory that focuses on how a listener interprets – through inference – what is being communicated. It assumes that human cognition tends to be geared to the maximization of relevance and every act of ostensive communication communicates a presumption of its own optimal relevance. A speaker who wants to achieve some particular effect should give whatever linguistic cues are needed to ensure that the interpretation consistent with the principle of relevance is the one she is intended to convey (Sperber & Wilson, 1995:249, 260). SW is among such linguistic cues and devices, for SW can help the hearers to wok out the communicative intention of the speaker and to make the right interpretation of speakers' utterances.

Database and research questions

The study is based on the Louvain International Database of Spoken English Interlanguage (LINSEI). LINSEI is a complementary project of ICLE (International Corpus of Learners' English (Written) headed by Prof. S. Granger at the Center of English Corpus Linguistics of Louvain University in Belgium (Granger, 2001). It started in 1995 and is now jointed by a number of other countries for different mother tongue backgrounds, including Japanese, Chinese, Swedish, Spanish, Italian, and Bulgarian, etc. The corpus includes data of informal interviews between a native English speaker and an EFL learner for about 15 minutes. 50 interviews were involved in each sub-corpus under the same topic within the same time length. It thus provides a fine database for comparison of EFL inter-language and native language, and also for identification of universal and L1-specific features of oral inter-language. The present corpus for the study is made up by four sub-corpora: LINSEI-Chinese, LINSEI-Japanese, LINSEI-French and LINSEI-Italian, about 100,000 words for each (see Table 1).

Table 1: Corpus used in the study

Corpora	Words	
CHIN	58,919*	Spoken English of Chinese Advanced EFL learners in 2001
JAP	36,999	Spoken English of Japanese Advanced EFL learners in 2001
FRAN	90,857	Spoken English of French Advanced EFL learners in 2001
ITA	58,656	Spoken English of Italian Advanced EFL learners in 2001
ICE-GB	246,166	Spoken English by British adults in 1990s
COLT	500,000	London teenagers' spoken English in the 1990s

(* the word count in this column only include the EFL learners' spoken words but excluding the interviewers' (i.e., the native English speakers) words)

The research addresses two questions:

1) How do EFL learners in the 4 sub-corpora use SW differently from native English speakers and from each other?
2) How is the learners' use of SW associated with their fluency in spoken English performance?

Methods and procedures

1. To get the general idea of EFL learners' use of SW, the first investigation is made, with the help of *Concord* tool of *Wordsmith*, on 19 typical SWs in the EFL learners' speaking parts in each sub-corpora, retrieving SW's type range, frequency and discoursal / interactional functions. The results are then compared with two native English spoken corpora: ICE-GB (spoken section) and COLT.

2. Since more than 60% of the 19 items are two-word phrases, the second investigation is made on the two-word list in the same corpora, using the *Wordlist tool* of *Wordsmith*. It is to observe the EFL learners' use of SW from another aspect and reveal some other features in their spoken English performance.

3. To explore the association between SW and oral English proficiency, the third investigation is made on two groups of Chinese EFL learners: the top ten best graded students and the bottom ten opposites. Comparison is made between their degree of fluency (demonstrated by the timing index including speech rate, filled/unfilled pause and mean length rate) and their use of SW (i.e., type range and frequency).

Results and discussion

With the help of Wordsmith, 19 types of most typical SW were identified and retrieved according to their discourse meaning and turn position in the EFL learners' speech (for detail see Aijmer, 1996: 200-233), thus it can be compared with the native English speakers' speech in terms of frequency and type range (see Table 2).

Table 2: The Frequency distribution of SW in each Corpus (/10,000)

Type of SW	COLT	ICE-GB	CHIN	JAP	FRAN	ITA
well	**23**	**68**	7	**19**	**114**	**40**

okay	15	15	**24**	**24**	6	**20**
like	23	6	3	14	**40**	8
right	27	33	3	2	4	0
oh	**75**	**59**	**26**	**20**	**28**	**11**
ah	14	06	6	**52**	4	5
just	55	**51**	**62**	11	**32**	**17**
all right	2	6	0	**0**	0	4
a bit	8	10	0	0	12	4
I think	11	34	**63**	**48**	**45**	**64**
I mean	14	**42**	18	3	17	6
I see	1	1	0	0	0	0
I know	9	0	0	0	1	2
you see	2	5	2	0	2	0
you know	36	**36**	**20**	5	22	9
not really	2	1	0	0	6	0
or something	5	6	1	3	10	0
sort/kind of	9	25	6	8	12	11
and things/everything /staff/that	2	6	0	0	2	2
Total of SW tokens	334	410	248	211	358	208
Total of SW types	19	18	13	12	17	14

Comments:

- SW in the 4 EFL learners' corpora are all less than that of the native English speakers, either in terms of total tokens or in terms of type range.
- The 2 oriental countries' corpora (CHIN & JAN) appear to use SW types less than that of the 2 European EFL learners' corpora, but no less than Italian corpus in terms of total tokens.
- SW in CHIN is similar to that of JAN, but it is less than FREN in general frequency, and less than the ITA in type range.

These preliminary findings lead us to further investigate the specific types of SW overused or underused by EFL learners and the results are in Table 3.

Table 3: The Top 5 SW in each Corpus

Corpus	1st	2nd	3rd	4th	5th	%
COLT	oh	just	you know	right	well	65
ICE-GB	well	oh	just	I mean	you know	65
CHIN	**I think**	just	oh	okay	you know	79
JAP	ah	**I think**	okay	oh	well	77
FRAN	well	**I think**	like	just	oh	72
ITA	**I think**	well	okay	just	oh	73

Comment:

- Native English speakers' most frequently used SWs are different in relation to age: The most frequent type for English Adults is *well* while for English teenagers, it is *oh,* indicating *well* could be a signal for more mature native English speakers.
- All EFL learners' preferred SW is 'I think', particular for those in CHIN and ITA.
- Chinese EFL's SW types are least, for the top 5 SW has taken up almost 80% of the range.

There rises the question: if EFL learners use SW less, how do they keep their speech flow continued and connected? One of the ways is to investigate the disfluency signals (such as repetition and filled pauses, see Lennon: 1990) in their speech. Table 4 demonstrates the sharp contrast of two-word SW occurring in the two-word lists between native English speakers' speech and the EFL learners'; it also reveals some salient disfluency signals among the EFL corpora.

Table 4: The Top Ten Two-word Phrase in Each Corpus

	COLT	ICE-GB	NSC*	CHIN	JAP	FRAN	ITA
1	***you know***	***I mean***	it was	**I think**	so I	and er	in the
2	I don't	***you know***	***you know***	in the	I think	it was	**I think**
3	**I know**	I m	***sort of***	The picture	and I	I I	the I
4	in the	***I think***	**I mean**	I I	I I	**I think**	I don't
5	do you	***sort of***	I was	want to	I was	I don't	the the
6	**I mean**	in the	**I think**	and the	went to	in the	and so
7	are you	I don't	in the	er the	I went	yes yes	don't know
8	I was	of the	and I	and then	when I	I was	it was
9	and then	going to	I don't	and I	is very	I dunno	of the
10	and I	and I	and then	the the	I like	er I	and I

Table 5: The Frequency of filled pauses in each corpus

Corpus	ICE-GB	CHIN	JAP	FRAN	ITA
Total words	246,166	58,919	36,999	90,857	58,656
Tokens of filled pause	9,822	4,434	2,687	7,828	5,617
/1,000	40	75	73	86	96

Comments

- 3 and 4 SWs are among the most frequently used two-word phrase in native English speakers' corpora, but there is only 1 in EFL learners', it is the same as 'I think' .
- There are more single word repetition (e.g., *I I the the*), hesitation markers (e.g., *er* , *and*), in the EFL corpora, CHIN in particular
- EFL corpora also have more filled pauses than native English speakers' corpus, indicating that the frequency of SW is negatively correlated to that of the disfluency signals such as repetition and filled pause.

This finding drives us to further investigate the correlation between the use of SW and the speaker's fluency, as is shown in Table 6 and 7.

Table 6: Fluency Index and SW within 3 Minutes Episodes of each Student in High/Low Proficiency Groups

Higher proficiency group							Lower proficiency group					
Sts'	Fluency index			SW			Sts'	Fluency index			SW	
No.	SR	UP	MLR	Token	Type		No	SR	UP	MLR	Token	Type
01	144	10	7.9	11	6		01	57	39	5.3	1	1
02	121	29	6.9	11	8		02	151	19	8.6	9	3
03	171	14	7.2	6	2		03	77	31	7.3	1	1
04	154	6	9.7	6	3		04	109	17	5.8	2	2
05	140	4	8.9	5	4		05	111	15	4.3	6	2
06	128	15	9.5	9	3		06	108	19	4.8	12	6
07	133	19	7.7	8	5		07	90	17	5.4	5	4
08	131	2	8.6	11	4		08	100	08	5.7	5	2
09	164	5	10.1	14	6		09	116	18	5.2	6	2
10	136	14	7.8	11	6		10	92	23	5.2	4	2
Ave.	142	3.9	8.4	9.2	4.7		Ave.	101	6.9	5.8	5.2	2.4

(SR=total words per minute. UP=total pause per minutes, MLR= average words with in undisrupted)

Table 7: Distribution of SW among High & Low proficiency groups

SW	3m'episod		Whole talk	
	High	Low	High	Low
Total words	4,266	3,033	14,775	10,980
well	8	0	34	1
okay	9	6	37	20
Like	6	0	13	1
right	0	0	1	0
oh	8	4	20	24
ah	2	3	4	3
Just	12	7	90	34
all right	1	0	1	0
I think	27	22	72	45
I mean	6	5	29	11
I see	1	0	3	0
you know	6	3	27	8
or something	1	0	4	0
sort/kind of	4	2	4	1
SW tokens	93	52	350	179
SW types	13	8	14	10

Comments:

- Students who have higher index of fluency (i.e., speak more words, speak faster, pause less and hesitate less) tend to use more tokens and types of SW, and vise versa.
- As to the type of SW, High group tend to use more *well, just, I mean, you know*; while Low group just use more *oh*. This indicates that the use of a certain type of SW is associated with the fluency of a speaker, it thus can be one of the descriptors in the evaluation of EFL learners' oral proficiency.

Implications

SW has long been neglected in EFL teaching. For example, some textbooks deliberately cut SWs in the dialogues so as to make the grammar patterns more prominent or clearly formed. Some EFL teachers even regards the SW occurring in learners' speech is a marker of disfluency, hence the EFL learners' unnatural speaking performances. This study demonstrates that SW has a highly frequent use in natural speaking and contributes to spoken performance both in connecting utterances and conveying speaker's intention. The findings in here again highlight the functions of SW in developing speaking fluency and has implication for EFL spoken training and oral testing. Further analysis is to be made on the actual use of SW in relation to different level of English proficiency so as to find out the sequencing and ordering of SW acquisition.

References

Aijmer, K. (1996) Conversational Routines in English—Convention and Creativity Singapore: Longman.

Granger, S. 2001 http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html

Hasselgren, A. (1998) *Smallwords and Valid Testing*. Unpublished PhD thesis in Bergen University, Norway.

Lennon, A. (1990) Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 3: 387-417.

Sperber, D. And D. Wilson. (1995) *Relevance*. Oxford: Blackwell.

Stenström, A. (1994) *Introduction to Spoken Interaction*. London: Longman.

A Preliminary Report on the COLSEC Project

Wei Naixing
Shanghai Jiaotong University, China

Introduction

For many years, SLA studies in China have exclusively focused on the learners' receptive skills, that is, reading skills and listening skills. Around the turn of the century, the EFL learners' English compositions have been used as basis, albeit sporadically, to investigate their acquisition of grammatical structures and lexical collocations. With the advent of the new millennium, many systematic studies were carried out, on the basis of the newly-constructed CLEC (Chinese Learners' English Corpus) to investigate the patterns of Chinese learners' written English production. But the learners' patterns of behavior in spoken language production have always been neglected. This paper attempts to redress the long-standing situation, by presenting a progress report on the study of the characteristics of Chinese learners' spoken English. The study is based on COLSEC, College Learners' Spoken English Corpus, which is a joint venture between several key universities of China and is funded by the Chinese National Social Science Research Foundation. At present, the corpus construction has reached its third year and over 300,000 words of spoken texts have been transcribed and annotated. The study to be reported is a component part of the project, with a view to describing and generalizing the characteristic features of the learners' spoken English and their tendencies of behavior in conversation. This will include the learners' tendencies in producing linguistic forms, particularly, the characteristic errors they tend to commit in pronunciation, the characteristic patterns of discourse organization, and the use of pragmatic strategies. For this purpose, a special corpus search software package CAST (Corpus Analysis and Statistic Tools) has been developed, by means of which linguistic forms, tagged pronunciation errors, conversational turns, and discourse signals have been retrieved and processed. Both quantitative and qualitative analyses have been conducted. As a result, some preliminary findings have emerged.

The study has found that Chinese learners tend to commit characteristic errors in pronouncing words of English and that they tend to over-use certain lexical sequences or chunks while under-using others. It has also found that turns in the learners' conversation can be categorized into several types and some larger discourse patterns recur. Another finding is that Chinese learners have strong inclinations to adopt certain means and techniques in managing conversation, but on the whole, their pragmatic strategies for conversational management are very much under-developed.

For reasons of space, what is printed below are just the two parts of this preliminary study.. The first part presents a short introduction to the COLSEC project and the basic statistics of the corpus at the present stage of development. The second part deals with major tendencies of pronunciation errors.

1. The COLSEC project and basic statistics of the corpus

The COLSEC, College Learners' Spoken English Corpus, project was funded by the Chinese National Social Science Research Foundation and launched in the year 2000. The purpose of the project is to construct a medium-sized corpus of the learners' spoken English, with approximately 700,000 words, which can serve as resources for investigating characteristics of Chinese learners' spoken English and, thus, providing insights and implications for English learning and teaching in the country. The project leader is Professor Yang Huizhong of Shanghai Jiaotong University, and many core corpus researchers have joined in the effort. COLSEC is also designed to be a sister corpus of CLEC (Gui Shichuen and Yang Huizhong, 2002). Completed in the year 2000, CLEC is a written corpus consisting of 1,000,000 words of the Chinese learners' English compositions and has proved to be a valuable resource for inter-language studies. Raw materials for COLSEC are episodes from the spoken test part of CET (College English Test), which is administered nation-wide twice a year. Each episode from the test consists of three sections, including an interview section, in which the examiner (a teacher) and a examinee (a student) perform question-answer tasks concerning the examinee's academic study, personal life and other familiar topics, a discussion section in which three examinees are having a discussion or debate over certain social issues of common interest, and , finally, a further discussion section in which the examiner and a examinee re-discuss , from a different angle, particular

questions which have just been discussed in the previous section. All the test episodes have been video-recorded. The episodes are selected according to the examinees' grades in the test, the topics of discussion and the geographical regions of examinees. A balanced approach has been taken in sampling the episodes. Then, the selected video-recorded test episodes are transcribed and annotated according to a set of general guidelines, procedural specifications and methodical requirements. All the important aspects of the English conversation, including conversation turns, international contours, various types of mispronunciation, errors of word stress and non-linguistic sounds are faithfully transcribed with related signs and symbols. At present half of the corpus has been completed, and the corpus has had a size of over 300.000 words. Table 1 below presents the overall statistics of the corpus at the present stage of development.

As can be seen from Table 1, the COLSEC under construction is providing basic data for studies of Chinese EFL learners' spoken language production. As the corpus continues to grow, more valuable information will be provided for the learners' overall productive language ability. Though it may be too early to draw any conclusions about the learners' productive language ability at the present stage, some preliminary studies can be carried out on the basis of the present corpus, to investigate the characteristics of the learners' spoken language. In the following sections of this paper, we will attempt to describe and generalize the general tendencies and major characteristic of the leaner spoken English.

Table 1: Statistics of CLSEC

Tokens	321,918	2-letter words	75,448
Types	7,155	3-letter words	69,363
Type/Token Ratio	2.22	4-letter words	59,514
Sd. Type/Token	28.89	5-letter words	30,573
Ave. Word Length	3.85	6-letter words	17,699
Sentences	22,741	7-letter words	16,735
Sent.length	14.15	8-letter words	9,550
Sd. Sent. Length	13.74	9-letter words	7,712
Turns	4980	10-letter words	4,739
Students' turns	3139	11-letter words	2,178
Teacher's turns	1841	12-letter words	992
Average turn length	64.64	13-letter words	591
1-letter words	26,690	14(+)-letter words	108

2. Characteristic Errors in Pronunciation

Pronunciation is one of the essential components of linguistic competence when we study the learners' spoken English. Other components of linguistic competence may include the use of syntactic structures, collocations, lexical chunks and idioms. For reasons of space, we will focus on the pronunciation in this section and leaves the use of lexical chunks to be discussed in the next section. All the other aspects will be specially addressed in separate papers. With all evidence extracted from the corpus and examined, the study now reveals that in pronouncing words of English, Chinese learners tend to commit four obvious characteristic types of errors. They are mispronunciation, sound addition, sound deletion, and stress shift. In mispronunciation, a certain phonological sound of a word is mispronounced as other incorrect sounds. In sound addition, a phonological sound is added to the pronunciation of a word. In sound deletion, a phonological sound is deleted from the normal pronunciation. In stress shift, the normal stress of a word is shifted to a preceding or succeeding syllabic sound. In the transcription specifications, four initial letters W, P, M and S are used to stand for the four types of error, respectively, all placed within a square bracket, and with detailed information attached, as in [We-ai], [Pd-er], [Mn] and [S2], the meanings of which are to be explained later in this section. Statistics show that mispronunciation has occurred 3043 times, sound addition 1080 times, sound deletion 554 times and stress shift 238 times, as is shown in Table2, in which [W*], [P*], [M*] and [S*] cover all the instances of pronunciation error of all sub-categories in each type.

Tables 2: Data for Pronunciation Errors

Errors	Occurrences
[W*]	3043
[P*]	1080
[M*]	554
[S*]	238

With a corresponding bar graph and a corresponding pie graph, it can be seen more clearly that mispronunciation has had the highest frequency of occurrence, accounting for 62% of all the pronunciation errors, thus, showing the strongest tendency in the students' pronunciation error. The next strongest tendency is sound addition, but its frequency of occurrence is reduced dramatically, only a little more than 1/3 that of mispronunciation, accounting for 22% of all the pronunciation errors. The rest two types of error, sound deletion and stress shift, taken together, account for 16% of all the errors, showing the two weak tendencies.

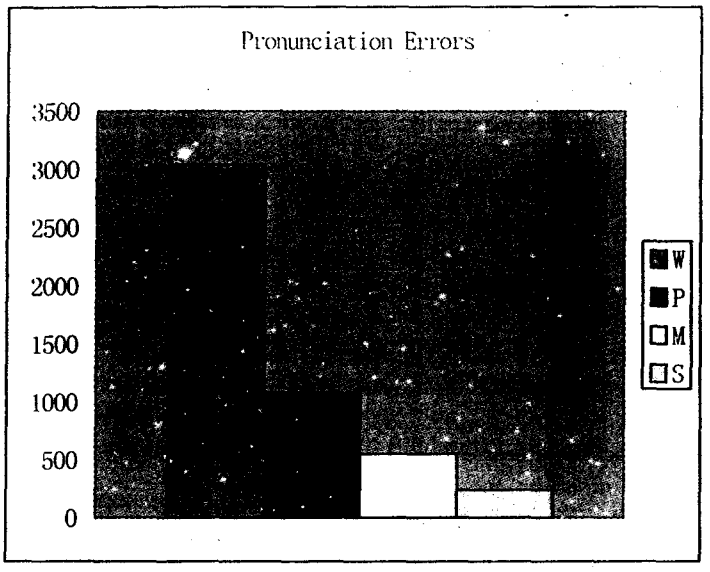


Figure 1: Frequency of occurrence of Pronunciation Errors

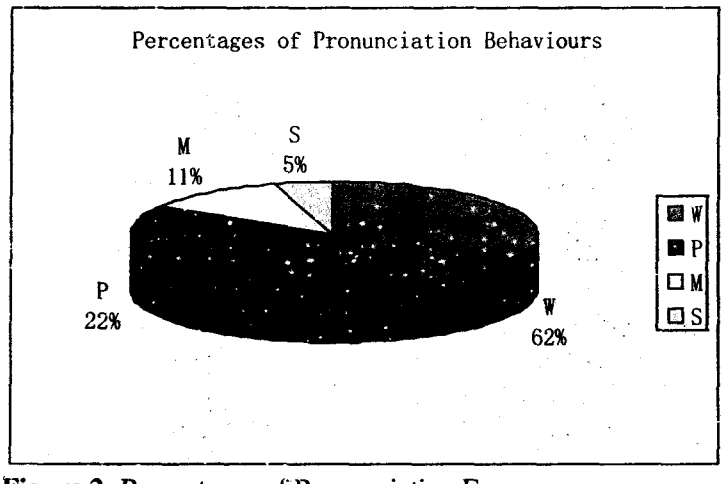


Figure 2: Percentages of Pronunciation Errors

2.1 Mispronunciations

2.1.1 Categories of mispronunciation and phonological tendencies

Mispronunciation has occurred 3043 times, which can be classified into over 70 sub-categories, such as [Wth-z], which means that the two letters "th", usually pronounced as the voiced dental fricative [ð] or the voiceless dental fricative [θ], has been mispronounced as the voiced alveolar fricative [z], as in the cases of *others, their, those, think, then, method, themselves* and *father*. More exemplar sub-categories are [Ws-s], which means that the letter "s" has been mispronounced as a voiceless alveolar fricative [s] when it should be pronounced as a voiced alveolar fricative [z], as in the cases of *Waters, rivers, matters, problems, numbers*, and [Wl-r], which means that the letter "l", usually pronounced as the voiced alveolar liquid [l], has been mispronounced as another voiced alveolar liquid [r], as in the cases of *place, absolutely, cloud, problem, lead, download, plane*. An important point to be noted in this respect is that not all the

sub-categories of mispronunciation are of equal weight in describing the learners' behavioral tendencies. Some have occurred far more frequently than others. If we make a cut-off line and disregard all those sub-categories of mispronunciation that have less than ten occurrences, we get 26 major sub-categories of mispronunciation, data of all of which are shown in Table 3 below.

Table 3: Data for Major Mispronunciations

N	Sub-categ.	Occur.	Accumulated.		Instances
1.	[Wth-z]	616	616	20.2%	others, then, themselves
2.	[Wth-s]	449	1065	35%	think, through, youth
3.	[Ws-s]	139	1204	39.6%	Waters, rivers, matters
4.	[Wv-w]	138	1342	44.1%	very, videos, traveling
5.	[Wl-r]	69	1411	46.4%	place, absolutely, cloud
6.	[Wr-l]	64	1475	48.5%	bright, several, pressure
7.	[Wth-d]	59	1534	50.4%	They, them, then
8.	[Ww-v]	58	1592	52.3%	way, well, world
9.	[We-i]	57	1649	54.2%	Better, rest, penny
10.	[Wu-a]	38	1687	55.4%	Industry, products, must
11.	[Wl-n]	37	1724	56.7%	slow, qualified, allowance
12.	[We-ei]	36	1760	57.8%	Special, protection, cigarette
13.	[Wi-i:]	27	1787	58.7%	Live, it, picture
14.	[Wa-ae]	24	1811	59.5	travel, demands, nation
15.	[Wa-e]	23	1834	60.3%	dangerous, salesman, chase
16.	[Wp-p]	20	1854	60.9%	Especially, respect, sports
17.	[Wa-ai]	20	1874	61.6%	Bad, habit, hackers
18.	[Wv-f]	19	1893	62.2%	value, divorce, service
19.	[We-e]	17	1910	62.8%	Business, media, marketing
20.	[Wa-ei]	14	1924	63.2%	Pirated, talent, establish
21.	[Wa-i]	14	1838	60.4%	Nature, place, phenomena
22.	[Wc-k]	14	1852	60.9%	discuss, discussion, magnificent
23.	[Wa-a:]	13	1965	64.6%	Many, finance, latest
24.	[We-ai]	12	1977	65.0%	Smell, internet, sceneries
25.	[Wt-d]	11	1988	65.3%	Matter, invite, city
26.	[Wf-v]	10	1998	65.7%	yourself ,of, fix

As has been shown in Table 3, the 26 sub-categories have had a total of 1998 occurrences, accounting for 66% of all the mispronunciation instances. And if we just focus on the sub-categories which have had 30 and 30 plus occurrences, then we get 12 such sub-categories, which have had a total of 1760 occurrences, accounting for almost 60% of all the pronunciation errors, from which we can generalize the major error tendencies in terms of phonological sound.

In all the pronunciation errors, those connected with the sound of the sequence "th" suggest the most noticeable tendency. The strongest tendency is to utter the voiced alveolar fricative sound [z] for its voiced dental fricative counterpart [ð]; next, the learners have a very strong inclination to utter the voiceless alveolar fricative sound [s] for its voiceless dental fricative counterpart [θ]; learners also tend to utter the voiced alveolar stop [d] for its voiced dental fricative counterpart [ð]. All the major pronunciation error tendencies can be generalized in terms of phonological sound as follows:

(1) For the sequence "th": a voiced alveolar fricative for a voiced dental fricative
(2) For the sequence "th": a voiceless alveolar fricative]for a voiceless dental fricative
(3)Fro the letter "s": a voiceless alveolar fricative for a voiced alveolar fricative
(4) For the letter "v": a voiced labiovelar glide for a voiced labiodental fricative
(5)For the letter "l": a voiced alveolar liquid for another voiced alveolar liquid
(6)For the letter "r": a voiced alveolar liquid for another voiced alveolar liquid
(7) For the sequence "th": a voiced alveolar stop for a voiced dental fricative
(8)For the letter "w": a voiced labiodental fricative for a voiced labiovelar glide
(9)For the letter "e": a high front vowel for a Mid front vowel
(10)For the letter "u": a low back vowel for a high back vowel
(11)For the letter "l": a voiced alveolar nasal for a voiced alveolar liquid
(12)For the letter "e": a mid front diphthong for a Mid front monophthong

2.1.2 Mispronunciation in types

So far, all the calculations of the categories of pronunciation errors are made in terms of tokens, that is, the absolute number of occurrences of errors. Each occurrence means a token. A word mispronounced 10 times will mean that it has10 tokens. If we want to know the number of different words which have been mispronounced, we have to use the term "type". A type means a mispronounced word and it may have one token. But, in most cases, more than one tokens are connected with a type. Let us take [Wth-z]. [Wth-z] has had 616 occurrences, which are connected with 40 different mispronounced words. So we say that [Wth-z] has had 616 tokens but just 40 types. Table 4 below shows such data.

Table 4: Tokens and Types of Major Sub-categories

Sub-category	Tokens	Types
[Wth-z]	616	40
[Wth-s]	449	42
[Ws-s]	140	57
[Wv-w]	138	53
[Wl-r]	69	39
[Wr-l]	64	30
[Wth-d]	60	9
[Ww-v]	58	27
[We-i]	57	13
[Wu-a]	38	6
[Wl-n]	37	23
[We-ei]	36	18

Data in Table 4 can be translated into a bar graph through which difference between the error types of each sub-category can be observed more clearly.

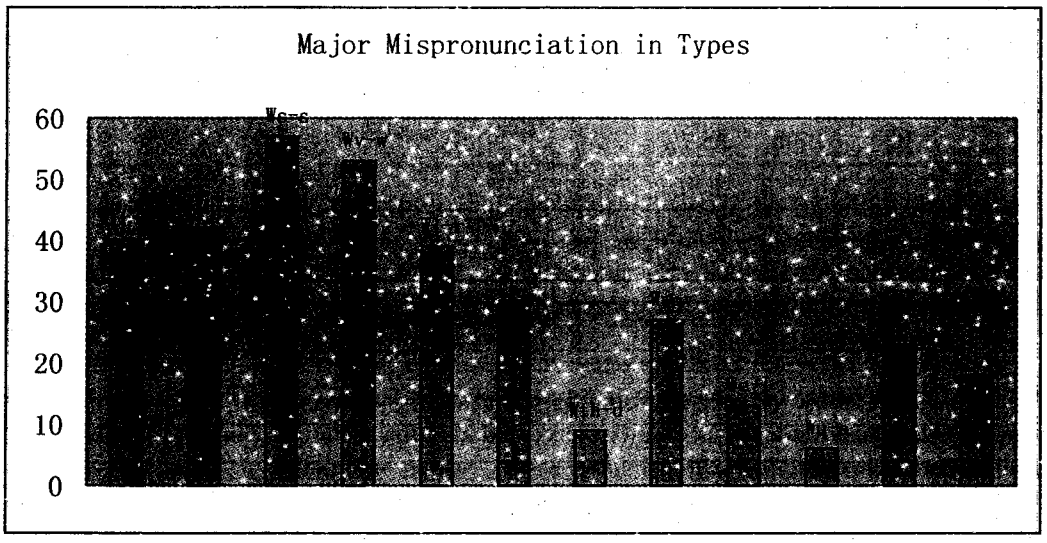


Figure 3: Major sub-categories of mispronunciation in types.

2.2 Sound Addition

Sound addition has occurred 1080 times in the corpus. Evidence reveals that there is a much stronger tendency for learners to add a phonological sound to the end of a consonant than to the end of a vowel, though both exist in the corpus. There are about ten sub-categories of sound addition, in which a vowel is added to the end of a consonant. Of all the occurrences, [Pd-er], in which a [er] is added to the end of the consonant[d], and [Pt-er], in which a [er] is added to the end of the consonant [t] have the largest shares. [Pd-er] has occurred 227 times, as the following concordances show:

1 there is there shows a map [Pp-er] and [Pd-er] er there is a plane [Wl-r] flyin
2 have ??? as well as movie, you may find [Pd-er] beauty in both of them [Wth-z].
3 interesting cultures so that we should [Pd-er] cherish every chance to go to go

4 ended to mm study hard er er for a good [Pd-er] foundation for my future career.
5 very necessary [S2] for some of us, but [Pd-er] I think if ¡er¡maybe maybe I w
6 if I have a chance, I will be very glad [Pd-er] to be a volunteer to do anything
7 at i study is very is very hard. I need [Pd-er] to relax.yes, I th
8 Er¡smoking, I know smoking is very bad [Pd -er] for¡er¡is¡do harm to our healt
9 have all fruit replace all of your food [Pd-er], I think it¡¯ s bad to your heal
10 the in the night. We should go to bed [Pd-er]¡er¡a little earlier, because we

[Pt-er] has occurred 130 times, as is shown in the concordances below:

1 o people have to go a long way to get [Pt-er] water, en, to get water to use.
2 ronment pollution, and the [Wth-z] next [Pt-er] important factor is that er every
3 ve with her and [Pd-er] she er¡brought [Pt -er] me brought me up. And [Pd-er] mm
4 is Lin Yan. An¡and [Pd -e] it¡¯s great [Pt-er] pleasure to meet you here. And er
5 is not impossible at at the moment. But [Pt-er] hm in the long [Wl-n] run it will
6 formed [Pd-e], and TV can can be a most [Pt-er] powerful way to give us a lot of
7 advantage of using cars, erm if we want [Pt-er] to erm to erm if we want to do e
8 st festival in China, and er usually at [Pt-er] that [Wth-zh] time, every family
9 our life conditions have changed a lot [Pt-er]. And we I think I know I remembe
10 s not very development [Mt] so we must [Pt-er] er introduce new technology and

Tendencies in sound addition need to be further examined and described.

2.3 Sound Deletion

Sound deletion has occurred 554 times, in which a vowel or a consonant has been deleted from the normal pronunciation. For example, in **[Mo]**, the phonological sound of the letter "o" is deleted in such words as "violence" and "economy", in **[Mt]**, the phonological sound of *t* is deleted from the normal pronunciation of such words as ***about, heart, benefit*** and ***must*** . Of all the sub-categories of sound deletion, [Mt] has occurred 62 times, and [Mn] 25 times. The following are concordances for [Mn].

1 ou put together. And you All the human [Mn] can enjoy the information and makes
2 [Pd-er] his and and her her performance [Mn] in the hour is excellent. Mm¡-
3 calls also calls ??. So¡in the second [Mn] picture, we shoul d, I know, we shou
4 can use we can use buses, cars, planes [Mn], air planes [Mn], trains, etc. But
5 so er it is very difficult to eliminate [Mn] the the fake products er and many c
6 f graduate course is not that essential [Mn]. not essential?
7 pass the examination. Er In my opinion [Mn], it is not wise to choose that kind
8 l er burden our horizons [Wo-ae] [Wr-m] [Mn], because we can learn new culture,
9 uh¡mm I can lea I can lea:n [Wear -ii] [Mn] little from them, and [Pd-e] uh pro
10 it, mm playing chess is er a hard mind [Mn] labor which can develop our many ab

Tendencies in sound deletion need to be further examined and described.

2.4 Stress Shift

Stress shift has occurred 238 times. There are four sub-categories of stress shift, [S1], [S2], [S3] and [S4], referring to cases where the stresses of words have mistakenly shifted to the first, the second, the third and the fourth syllable respectively. For example, in the learners' speech, the stress of such words as *although, compute, contribute, increased, coordinate, electronic, encourage, horizon, finance, financial, interrupt, kilometer, identify, improve, provide, urbanization* and *indifferent* has fallen on the first syllable, and, thus, they are all labeled as [S1]. Examples of [S2] include: *atmosphere, industry, consequence, economics, contact, necessary, relative, meanwhile, competition, colleague, algebra* and *academic*. In [S3], we may find *economy, intelligence, corporate, anatomy, psychology, Singapore, sophomore, supermarket, atmosphere, internet, literature, cigarette, aerobics, agriculture* and *advantage*. [S4] has the fewest instances: *communicate, curiosity, university*. Of the 238 occurrences of stress shift, 59 have fallen into the category of [S1], 126 into the category of [S2], which is the largest share. And of the 126 occurrences of [S2], the word "industry" has appeared 33 times. There are 44 occurrences for [S3], in which the word "economy" has appeared 10 times. [S4] has 9 occurrences, taking the smallest share. Figure 5 below shows

the general picture.

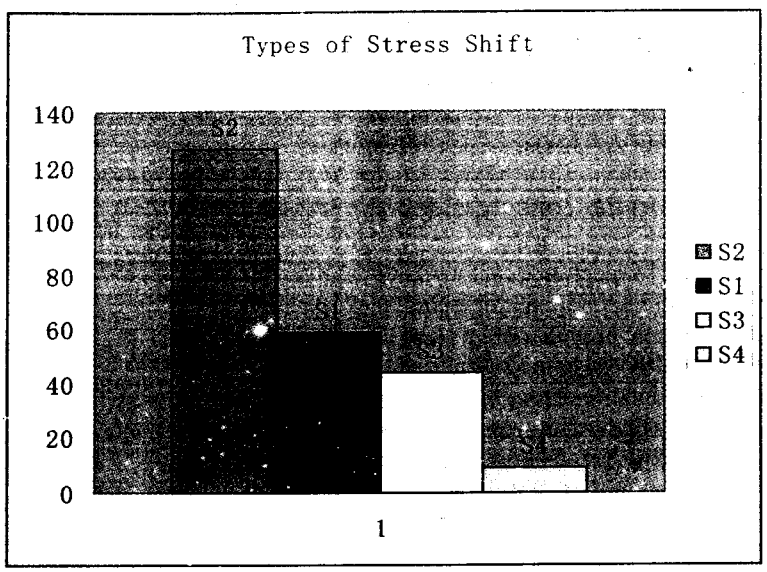


Figure 4 Occurrences of categories of stress shift.

As the data shows, the strongest tendency for the learners is to shift the normal word stress to the second syllable. And this usually happens to a multi-syllable word where the normal stress should be on the third syllable from the end (the first syllable in the case of a 3-syllable word). Also stress shift may happen to words like "contact" and "increase", which can be used as both a verb and a noun, but their stresses differ with syntactic category. In such cases, learners may confuse a noun stress with a verb stress, or vice versa. The instances of stress shift cover a variety of complexities, and data needs to be further explored so as to arrive at the more specific tendencies.

Conclusions

The COLSEC corpus has provided, and will continue to provide, valuable data for a systematic study of Chinese learners' spoken English. This preliminary study has examined the corpus evidence available so far and made a full-scale inquiry into the related issues. All the data and discussion have shown that there are some generalizable characteristics in the learners' spoken English. They tend to commit characteristic errors in pronouncing words of English, including mispronunciation, sound addition, sound deletion and stress shift. The single most frequent error is mispronunciation, in which there are very strong tendencies for them to utter a closely related phonological sound for a certain correct sound. The learners have also shown characteristic patterns of behaviour in the use of lexical chunks, discourse patterns and conversation management strategies, details of which will be reported in separate papers in due course.

References

杨惠中，2002，《语言库语言学导论》，上海外语教育出版社
桂诗春，杨惠中，2003，《中国学习者英语语料库》，上海外语教育出版社
濮建中，2000，"中国学生英语动词语法和词汇型式使用特点初探"，《现代外语》，第一期，24-44
李文中，2003，"基于英语学习者语料库的主题词研究"《现代外语》，第三期，283—293
娄宝翠，2001，"中国学生英语写作中的造词现象"，《外语教学与研究》，第一期，63-68
卫乃兴，2002，"语义韵研究的一般方法"，《外语教学与研究》，第四期，300-307

A Corpus-Based Analysis of the Use of Frequency Adverbs by Chinese University English Majors

Wen Qiufang, Ting Yenren
Nanjing University

1. Introduction

Spoken and written language contains large quantities of frequency adverbs such as "often," "sometimes" and "never." This paper reports a study on how Chinese college English majors use the 20 most frequently used frequency adverbs, or, the top twenty frequency adverbs (TTFAs; see Leech, *et al.* 2001), in their written and spoken English and how they use these TTFAs differently from the way native speakers use them as recorded in the British National Corpus (BNC). These TTFAs can be classified into three levels in terms of their occurrence frequencies: 15 of them belong to the 1000-word level, three to the 2000-word level and two to the academic word list (See Table 1).

TABLE 1: Twenty Top Frequency Adverbs (TTFAs)[1]

Level of vocabulary	TTFA	Number
1000-word level	never, always, often, ever, sometimes usually, once, generally, hardly, no longer, increasingly, twice, in general, occasionally, mostly	15
2000-word level	frequently, rarely, regularly	3
Academic word list	normally, constantly	2

The previous corpus-based studies found that advanced EFL learners show a tendency to overuse high-frequency words and tend to employ a spoken type of discourse in their English writing (Cobb, 2002; Ringbom, 1998; 文秋芳，丁言仁、王文宇，2003). However, these studies did not investigate the use of frequency adverbs, nor did they make a comparison between speech and writing by advanced EFL learners with regard to the use of frequency adverbs. Therefore, this study was designed in an attempt to investigate whether Chinese English learners demonstrate a similar or different pattern in using the TTFAs as they use other high-frequency words and whether they use TTFAs similarly or differently in speaking as they do in writing. Specifically, this study addressed the following questions:

1) Do Chinese English majors overuse or underuse these TTFAs?
2) Do they overuse or underuse the TTFAs differently between speech and writing?
3) Do they differ more from native speakers in writing or in speaking with regard to the use of the TTFAs?
4) Do they demonstrate a similar pattern of writing-speaking difference as the native speakers in the use of the TTFAs? If not, what are the differences?

2. Data and methodology

The learner data used in this study included a written English corpus, 481,635 words of essay writing by English majors from the Chinese Learner English Corpus (CLEC) developed by Gui and Young (2003), and a spoken English corpus, 473,408 words of speech transcription from the Spoken English Corpus of Chinese Learners (SECCL), which Nanjing University has been trying to build on the basis of the performance of sophomore English majors in the national Spoken English Test for English Majors (Band 4) from 1999 to 2002. With these two corpora combined, the general learner corpus, or, the Chinese English Major Corpus (CEMC), contained 955,043 words. Hereafter, the spoken learner corpus is referred to as CEMCS, the written corpus as CEMCW.

[1] It is unclear why Leech *et al.* (2001) include the phrases "no longer" and "in general" in the TTFAs. For the convenience of discussion, these two phrases are referred to as "words" without a distinction between a word and a phrase.

The control corpus in this study was the British National Corpus (BNC), which contains 90 million words of written English and 10 million words of spoken English, all by native speakers. The spoken and written parts of BNC are referred to as BNCS and BNCW, respectively. Table 2 provides a summary of the corpora used in this study.

TABLE 2: Learner and Native-Speaker Corpora

Type of corpus		Size of the corpus	Total
The learner corpus: Chinese English major Corpus (CEMC)	Spoken (CEMCS)	473,408 words	955,043 words
	Written (CEMCW)	481,635 words	
The native-speaker corpus: British National Corpus (BNC)	Spoken (BNCS)	10 million words	100 million words
	Written (BNCW)	90 million words	

Data analysis included the following steps:

1) Use WordSmith tools to find the occurrence frequencies of each TTFA in the learner corpus and then in the learners' spoken and written English corpora, respectively.
2) Compare the general learner English corpus CEMC with the BNC to see overall differences between the two so as to measure how learners overuse or underuse the TTFAs.
3) Compare CEMCS, the learners' spoken corpus, with BNCS, the spoken English portion of BNC.
4) Compare CEMCW, the learners' written corpus, with BNCW, the written English portion of BNC.
5) Measure the extent to which the differences in the TTFA use between the CEMCS and the CEMCW (the learners' spoken and written corpora) deviate from those between the BNCS and the BNCW (the spoken and written portions of BNC).
6) Compare the TTFAs in the CEMC that were register neutral, written-register sensitive and spoken-register sensitive, respectively, with those in the BNC so as to measure the extent to which these sets of TTFAs in CEMC differ from those in the BNC.

3. Results

3.1 TTFA use in the learner corpus

The first research question of this study concerns the overuse or underuse of the TTFAs by Chinese English majors. Table 3 shows that in the CEMC, there are 215 occurrences of the TTFAs per million tokens while in the BNC, there are only 151 occurrences per million tokens; the CEMC has 64 occurrences more than the BNC. However, a close examination of the frequencies of each of the TTFAs shows that the picture is not so simple. Among the TTFAs, Chinese students overuse

always	once	sometimes	usually	often
never	hardly	no longer		

but underuse

normally	increasingly	ever	twice	frequently
rarely	occasionally	generally	regularly	in general

Therefore, both the overuse and underuse of TTFAs can be found in the performance of Chinese English majors.

As shown in Table 3, although learners underuse more TTFAs (10) than they overuse them (8), the difference in average normalized frequency between the CEMC and BNC is far greater in the overused TTFAs (214 occurrences per million words) than in the underused TTFAs (43 occurrences per million words). In other words, the overusing tendency is stronger than the underusing tendency, and such overuse concentrates on only 8 TTFAs.

TABLE 3: An Overall Comparison Between CEMC and BNC

	TTFA	CEMC Normalized frequency	BNC Normalized frequency	Difference	Tendencies
1	always	1206	462	744	**8 overused TTFAs** Average normalized frequency difference: **214**
2	once	470	183	287	
3	sometimes	433	205	228	
4	usually	387	191	196	
5	often	539	376	163	
6	never	604	559	45	
7	hardly	113	88	25	
8	no longer	110	88	22	
9	mostly	39	39	0	Identical or similar
10	constantly	24	31	-7	
11	in general	25	42	-17	**10 underused TTFAs** Average normalized frequency difference: **43**
12	regularly	13	39	-26	
13	generally	83	116	-33	
14	occasionally	7	40	-33	
15	rarely	7	42	-35	
16	frequently	18	58	-40	
17	twice	12	63	-51	
18	ever	207	259	-52	
19	increasingly	4	66	-62	
20	normally	8	83	-80	
Average normalized frequency		**215**	**151**	**64**	

3.2 TTFA use in learners' spoken corpus

Table 4 shows how Chinese English majors overuse and underuse TTFAs in speech. In the learners' spoken corpus CEMCS, there are 231 occurrences of the TTFAS per million words while in the native speakers' spoken corpus BNCS, there are 132 per million words; the CEMCS has 99 occurrences more than the BNCS. Similar to the overall comparison between CEMC and BNC, a close examination of the frequencies of each of the TTFAs also shows the complexities of the picture. Chinese English majors overuse

always	**once**	often	**sometimes**	usually
hardly				

but underuse

normally	never	ever	twice	generally
in general	occasionally	no longer	constantly	increasingly

This finding indicates that in speech, they have a tendency to overuse certain TTFAs but underuse others. Moreover, the data demonstrate a pattern similar to that found in the overall comparison between CEMC and BNC: although learners underuse more TTFAs (10) than they overuse them (6), the difference in average normalized frequency between the CEMCS and BNCS is far greater in the overused TTFAs (407 occurrences per million words) than in the underused TTFAs (48 occurrences).

TABLE 4: A Comparison Between the CEMCS and BNCS

	TTFA	CEMCS Normalized frequency	BNCS normalized frequency	CEMS-BNCS Difference	Tendencies
1	always	1493	597	896	**Overused 6 TTFAs** Average normalized frequency difference: **407**
2	once	699	114	585	
3	often	640	175	465	
4	sometimes	454	199	255	
5	usually	300	144	156	
6	hardly	133	46	87	

7	frequently	11	11	0	Similar or identical
8	regularly	13	16	-3	
9	rarely	4	9	-5	
10	mostly	19	27	-8	
11	increasingly	0	14	-14	**Overused 10 TTFAs** Average normalized frequency difference: **48**
12	constantly	27	13	-14	
13	no longer	6	29	-23	
14	occasionally	0	23	-23	
15	in general	0	24	-24	
16	generally	25	59	-34	
17	twice	6	62	-56	
18	ever	186	275	-89	
19	never	606	700	-94	
20	normally	4	112	-108	
Average normalized frequency		**231**	**132**	**99**	

3.3 TTFA use in learners' writing

Table 5 shows how Chinese English majors overuse and underuse TTFAs in writing. In the learners' written corpus CEMCW, there are 45 more occurrences of TTFAs per million words in the CEMW than there are in the native speakers' written corpus BNCW. This suggests that there be a slight tendency to overuse the TTFAs by Chinese English majors in writing. However, close examination of the normalized frequency difference in each TTFA in the CEMW and BNCW shows that both overuse and underuse of TTFAs can be found in writings by Chinese English majors. Furthermore, although the number of overused TTFAs (8) was slightly fewer than that of underused TTFAs (10), the overusing tendency is stronger than the underusing tendency since the difference in averaged normalized frequency in the overused TTFAs between the CEMCW and BNCW (122 occurrences per million words) was much greater than that in the underused TTFAs (50 occurrences).

TABLE 5: A Comparison Between the CEMW and BNCW

	TTFA	CEMW normalized frequency	BNCW normalized frequency	CEMW-BNCW difference	Tendencies
1	always	914	446	468	**Overused 8 TTFAs** Average normalized frequency difference: **122**
2	sometimes	411	206	205	
3	no longer	216	95	121	
4	never	602	542	60	
5	once	237	191	46	
6	often	436	399	37	
7	generally	141	122	19	
8	mostly	60	41	19	
9	in general	50	44	6	Similar or identical
10	hardly	93	93	0	
11	constantly	21	33	-12	**Underused 10 TTFAs** Average normalized frequency difference: **50**
12	occasionally	15	42	-27	
13	ever	228	257	-29	
14	regularly	12	41	-29	
15	rarely	10	46	-36	
16	frequently	25	63	-38	
17	twice	19	63	-44	
18	increasingly	8	73	-65	
19	normally	12	79	-67	
20	usually	475	197	-150	
	Average	**199**	**154**	**45**	

4. Comparison of the learners' speech with their writing in TTFA use

Table 6 summarizes how Chinese English majors differ from native speakers in speaking and writing with regard to the use of the TTFAs. In general, they overuse 6 TTFAs in speech (CEMCS) and 8 TTFAs in writing (CEMCW). In terms of the difference in average normalized frequency, however, the overusing tendency is much stronger in speech than in writing; the CEMCS shows 407 more occurrences of the TTFAs per million words than the BNCS while the CEMCW only show 122 more occurrences than the BNCW. As for the underusing tendency, Chinese English majors underuse 10 TTFAs both in speech and in writing. In addition, for the underused TTFAs, the difference in average normalized frequency between the CEMCS and BNCS is 48 per million words, and that between the CEMCW and BNCW is 50 per million words; the two are similar. There are 4 TTFAs with identical or similar frequencies in the CEMCS to the native speaker's norm while there are only 2 in the CEMCW. In general, Chinese English majors differ from native speakers more in speaking than they do in writing with regard to the use of TTFAs, primarily because in speech, they drastically overuse a few TTFAs.

It is also worth noting that some overused TTFAs are identical in speech and writing (i.e., "always," "once," "often," and "sometimes"). So are some underused ones (i.e., "normally," "ever," "twice," "occasionally," "constantly," and "increasingly"). Other TTFAs are overused or underused differently in speech and writing.

TABLE 6: A Comparison of Two Sets of Frequency Differences

	Overuse		Underuse		Identical or similar	
	Number of TTFAs	Average normalized frequency difference	Number of TTFAs	Average normalized frequency difference	Number of TTFAs	Average normalized frequency difference
CEMCS-BNCS	6 (always, once, often, sometimes, usually, hardly)	407	10 (normally, never, ever, twice, generally, in general, occasionally, no longer, constantly, increasingly)	-48	4 (frequently regularly rarely mostly	-4
CEMCW-BNCW	8 (always, sometimes, no longer, never, once, often, generally, mostly)	122	10 (usually, normally, increasingly, twice, frequently, rarely, regulariy, ever, occasionally, constantly)	-50	2 (in general hardly)	3

5. Speaking-writing differences in TTFA use in the CEMC and BNC

This section reports findings that address the last research question of this study: whether or not Chinese English majors demonstrate a similar pattern in speaking-writing difference as native speakers with regard to the use of TTFAs.

Table 7 shows that the speaking-writing difference in the native speaker corpus BNC is negative (-22); that is to say, native speakers use fewer TTFAs in speech than they do in writing. In contrast, the speaking-writing difference in the learner corpus CEMC is positive (32); Chinese English majors use more TTFAs in speech than they do in writing. Therefore, the two speaking-writing difference patterns are skewed in opposite directions.

A close examination of the TTFAs shows that native speakers are more likely to use

never always normally ever

in speech than in writing, but more likely to use

often once no longer generally increasingly usually
frequently hardly rarely regularly constantly in general
occasionally mostly

in writing than in speech. They are equally likely to use

twice sometimes

in speech and in writing. In other words, in the native speaker corpus, 4 TTFAs are spoken-register sensitive; 14, written-register sensitive and 2, register neutral.

TABLE 7: A Comparison of Register Differences

	Word	Difference (BNCS-BNCW)		Word	Difference (CEMCS-CEMCW)
1	never	158	1	always	579
2	always	151	2	once	462
3	normally	33	3	often	204
4	ever	18	4	sometimes	43
5	twice	**-1**	5	hardly	40
6	sometimes	**-7**	6	constantly	**6**
7	mostly	-14	7	never	**4**
8	occasionally	-19	8	regularly	**1**
9	In general	-20	9	rarely	-6
10	constantly	-20	10	normally	**-8**
11	regularly	-25	11	increasingly	**-8**
12	rarely	-37	12	twice	-13
13	hardly	-47	13	frequently	-14
14	frequently	-52	14	occasionally	-15
15	usually	-53	15	mostly	-41
16	increasingly	-59	16	ever	-42
17	generally	-63	17	in general	-50
18	no longer	-66	18	usually	-75
19	once	-77	19	generally	-116
20	often	-224	20	no longer	-210
	Average	**-22**		**Average**	**32**

However, the learner corpus demonstrates important differences in the use of TTFAs in terms of register. As shown in Table 7, Chinese English majors are more likely to use

always once often sometimes hardly

in speech than in writing, but more likely to use

twice frequently occasionally mostly ever in general
usually generally no longer

in writing than in speech. They are equally likely to use

constantly never regularly rarely normally increasingly

in speech and in writing. In other words, in the learner corpus, 5 TTFAs are spoken-register sensitive; 9,

written-register sensitive; and 6, register neutral.

Table 8 compares the three sets of TTFAs (spoken-register sensitive, written-register sensitive and register neutral) in the learner corpus CEMC with their counterpart in the native speaker corpus BNC. As the table shows, of the 5 TTFAs that are spoken-register sensitive in the CEMC, only one ("always") is also spoken-register sensitive in the BNC. Of the 9 TTFAS that are written-register sensitive in the CEMC, by contrast, most are also written-register sensitive in the BNC, with the exception of "ever" and "twice." The set of register-neutral TTFAs demonstrate a similar pattern as that of the spoken-register sensitive TTFAs. Of the 6 register-neutral TTFAs, none occur in the same category in the BNC. This finding suggests that Chinese English majors still do not have a clear register awareness in their choice of TTFAs.

TABLE 8: A Comparison of TTFAs in Registers Between the BNC and CEMC

	Spoken-register sensitive	Written-register sensitive		Register neutral
BNC	never always normally ever **(4 TTFAs)**	often, once, no longer, generally, increasingly, usually, frequently	hardly, rarely, regularly, constantly, in general, occasionally, mostly **(14 TTFAs)**	Twice Sometimes **(2 TTFAs)**
CEMC	always once often sometimes hardly **(5 TTFAs)**	no longer generally usually in general **ever**	mostly occasionally frequently **twice** **(9 TTFAs)**	constantly never regularly rarely increasingly normally **(6 TTFAs)**

4. Discussions and conclusions

4.1 Summary of the findings

The major findings from this study can be summarized as follows:

1) Chinese university English majors tend to overuse and underuse certain TTFAs in their speech and writing. The overusing tendency is stronger than the underusing tendency in both speaking and writing.
2) The overusing tendency is more marked in their speech than in their writing while the underusing tendencies in speech and writing are similar to each other in terms of their frequencies. Some of the overused or underused TTFAs in speech are the same as those in writing but others are different.
3) Chinese English majors demonstrate a pattern of speaking-writing difference that is opposite to that shown in the native speakers' corpus: they tend to use more TTFAs in their speech than in their writing. This shows that they use TTFAs without awareness of their register differences.

4.2 Possible reasons for the overuse and underuse of TTFAs

From the corpora alone, it is difficult to pin down the reasons for the differences between Chinese English majors and native speakers in the use of the TTFAs. However, several factors might be underlying what was observed in corpus analysis:

1) University English majors in China have a much smaller vocabulary than native speakers and, therefore, tend to overuse the words they know. They would use TTFAs where native speakers might use less frequent but more precise time adverbials
2) When the fluency is limited, Chinese English majors have a greater need than native speakers for certain "time buyers" such as "always," "often" "sometimes," "usually" and "once." For that matter, TTFAs might be used where they do not have to be used.
3) Chinese English majors tend to overuse "always," "often" "sometimes," "usually" and "once,"

TTFAs to which equivalents are readily available in their mother tongue; they tend to underuse "normally," "ever," "increasingly" and other TTFAs whose Chinese equivalents are used at low frequencies in their daily speech in the first place.

4.3 Suggestions

It is obvious from the discussion that future research may be directed at studying the linguistic, textual contexts under which Chinese English learners overuse or underuse certain TTFAs as compared with their native speaker counterpart. Such investigation will shed light on the study of the reasons for the TTFA overuse and underuse by Chinese English learners.

References

Altenberg, B & Granger, S. 2002. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22: 173-194.

Altenberg, B & Tapper, M. 1998. The use of adverbial connectors in advanced Swedish learners' written English [A]. In S. Granger (ed.). Learner English on computer [C]. London & New York: Longman.

Cobb, T. 2002. Analyze late interlanguage with learner corpora: Quebec replications of three European studies [J]. *Canadian Modern Language Review* (In press).

Granger, S (ed.) 1998. Learner English on computer. London & New York: Longman.

Leech, G., Rayson, P. & Wilson, A. 2001. Word frequencies in written and spoken English. London: Pearson Education Limited.

Nation, I.S.P. 1990. Teaching and learning vocabulary [M]. Massachusetts: Newbury House.

Nation, I.S. P. 2001. Learning vocabulary in another language. Cambridge: Cambridge University Press.

Ringbom, H. 1998. Vocabulary frequencies in advanced learner English [A]. In S. Granger (ed.). Learner English on computer. London & New York: Longman.

WordSmith Tools, 1996. Oxford: Oxford University Press.

文秋芳、丁言仁、王文宇，2003，中国大学生书面语中的口语化倾向，《外语教学与研究》第4期。

桂诗春、杨惠中，2003，《中国英语学习者语料库》，上海外语教育出版社。

The Effects of the Command of Formulaic Sequences on Oral English Performance

Ting Yenren, Wen Qiufang
Nanjing University

Abstract: This paper reports a study that used oral English transcriptions contained in a learner corpus that is being built. It analyzed the formulaic sequences employed by 70 second-year English majors at the retelling task of the 2002 National Spoken English Test for English Majors (Band 4) as well as the linguistic accuracy of their performance at this task. Results show that 1) there is a wide discrepancy between students in the use of formulaic sequences in their oral English; and 2) the learners' command of formulaic sequences, rather than their command of grammar, has a direct effect on their oral English scores. Therefore, learning formulaic language should be an important part of learning language.

Keywords: formulaic sequence, formulaicity, oral English, accuracy, fluency, idiomaticity

Introduction

Corpus linguistics shows that in everyday language use, people do not normally construct sentences out of discrete words and grammar rules; rather, they extensively employ strings of collocated words that they have learned, retrieved and used as single units (see Wray 2002). Research into the use of these formulaic sequences is increasingly drawing the attention of applied linguists and language teaching specialists. Some of these sequences are not grammatically analyzable, but even for those that are analyzable, language users usually have neither the need nor time to make such analysis. According to Sinclair (1991), language users follow the idiom principle and the open choice principle, the former referring to stringing formulae in constructing language, the latter to choosing words and applying rules. He claims that people primarily apply the idiom principle and that even if they switch to the open-choice principle, they quickly switch back again.

Corpus linguistics is changing the view of language. From Saussure to Chomsky, linguists often view language as only systematic, rule-governed behavior and set out to study this system of *langue* or competence, a system that is responsible for the generation and comprehension of novel sentences. The notion of formulaic language leads to a dual nature view of language: language both has rule-based analyticity and memory-based formulaicity (Skehan 1998); it is both closed and open-ended.

The dual nature view of language in turn leads people to see the importance of the knowledge of formulaic sequences in second language acquisition. Widdowson (1989) views the learning of formulaic chunks as more important than that of grammar rules. He argues that a great deal of knowledge of language seems to consist of "formulaic chunks, lexical units completely or partially assembled in readiness for use" (1989: 135). By contrast, rules, he argues, are only "variably applied" because their role is only to adapt formulaic chunks to syntactic constraints and contextual requirements; they are "not generative but regulative and subservient" (1989: 135). Other researchers (e.g., Cowie & Howarth 1996; Howarth 1998; Lewis 1993; Nattinger & DeCarrico 1992; Pawley & Syder 1983; Weinert 1995; Wray 2000) also rightly point out the importance of learning formulaic language and point out the weaknesses in the knowledge of formulaic language on the part of L2 learners. Nevertheless, not much research has investigated how L2 learners learn and use formulaic language and how the different learning strategies they employ affect their use of formulaic language. Within China proper, there has been little research on how Chinese L2 learners learn and use formulaic language.

This paper reports a study that attempted to investigate the effects of Chinese English majors' knowledge of formulaic sequences and knowledge of grammar on their oral English performance. It was hypothesized that students with a better command of formulaic English would have higher scores in the oral English test, and those with a better command of grammar would also have higher scores in the test.

Methodology

The data for this study were collected from a portion of the Spoken English Corpus of Chinese Learners (SECCL), which Nanjing University has been building by transcribing the audio recordings of the performance of sophomore English majors in the National Spoken English Test for English Majors (Band 4) from 1999 to 2002.

The band-4 oral English test consists of three items: (1) Retelling a 300-word story after listening to it twice; (2) Giving a three-minute talk on a given topic (after a three-minute preparation); and (3) Holding a four-minute conversation with a partner on a given issue (after a three-minute preparation). Each test-taker's performance is recorded on tape, and the cassettes from different universities are randomly divided into packages of 35 each. Two experienced teachers grade each tape separately according to its content, organization, accuracy, fluency, pronunciation and intonation. Then, another two teachers rank the tapes in a package, listened to tapes the two previous teachers have ranked differently, and decide on the final grades: 4 (Excellent), 3 (Good), 2 (Pass), or 1 (Fail). In the 2002 test, these four grades, respectively, made up 21%, 61%, 14% and 3% of the test-takers.

This study investigated the use of grammar and formulaic sequences in the retelling task of the 2002 test by 70 English majors (two packages) as recorded in the transcription. The original story the students were asked to retell (see Attachment) contained 354 words. An analysis by the software Range (Version 1.12) showed that 83.6% of these words belong to the most frequently used 1,000 words and 13.3% belong to the second most frequently used 1,000 words. The story was about an unpleasant experience of a visitor after he checked into a hotel under renovation. Data analysis included the following steps:

1) Select from the original script sequences and phrases that were related to the content and should have been noticed by the students. Altogether, 79 sequences were selected, ranging from very short ones such as "might be" and "at all" to rather long ones such as "overlooks a beautiful bay" and "heard someone hammering loudly." The selection was indeed rather arbitrary and simplified since, in the first place, the boundaries between formulae and non-formulae remained "open to debate" (Hunston 2002: 147); formulae may overlap and one formula may contain another.
2) Delete from the transcriptions markers for pause, hesitation and false start and correct the spelling that was intended to imitate learner pronunciation. For instance, some students mistakenly pronounced "win" or "ring" when they meant to say "wing." The spelling was put back to "wing" so as to correctly measure their command of the sequence "to build a new wing."
3) Identify, in the transcription for each student, the phrases that had occurred in exact wording in the original script. This was a narrow way of identifying formulaic sequences. With this method, if a student used "he stays," which occurred in the original, it was counted as one correctly used formulaic sequence, but if she used "he lives" or "he stayed," it was not counted. The number of sequences a student used was tallied.
4) Identify, in the same material, the phrases that could be seen as paraphrases of their counterpart in the original script, e.g., "louder and louder" and "louder than before" for "louder than ever." Similarly, "he stayed" was also treated as a corrected paraphrase of "he stays" when the story was talking about the habitual behavior of the protagonist. This was a broad way of identifying formulaic sequences.
5) Divide the transcription for each student into T-units. A T-unit would be counted as wrong if it was not in keeping with the content of the original story or if it contained errors in tense, number, word form, word order or fragmentation. It would be counted as wrong, for instance, if a student used "noisy" for "noise" or "dust" for "dusty," but it would be counted as correct if she used the past tense when the story was introducing the background. The rate of T-unit correctness was calculated for each student.
6) Apply the statistic procedures to studying how the oral English scores as a dependent variable varied with the quantity of the formulaic sequences and the T-unit correctness rate as independent variables.

Findings and Results

1) The quantities of the formulaic sequences used in the oral English test

TABLE 1: Formulaic Sequences Found in the Oral English Transcriptions (N = 70)

Formula Identification	Min	Max	Mean	Std Dev
Narrow way	0.00	25.00	7.61	4.76
Broad way	0.00	31.00	8.90	5.60

Table One shows that there was a wide discrepancy between the 70 students in the use of formulaic sequences in their retelling performance. The mean was 7.61 by the narrow way of formula identification and 8.90 by the broad way, which resulted in a slightly larger number of formulaic sequences for each student. This was a small number in comparison to the 79 sequences that had been selected from the original script. On the one hand, this indicates that although the students had listened to the story on tape twice before they started to retell it, most of them did not try to memorize the original text but used their own language. On the other hand, it also demonstrates the problems with the students' spoken English, which was not accurate or idiomatic, with many errors and unidiomatic usages.

The discrepancy in the use of formulaic sequences indicates the discrepancy between the students in the learning strategies they use when trying to improve their listening and reading comprehension. Many students, while listening and reading, only go after meaning comprehension but pay no attention to the way language is used. A few students, however, can go beyond comprehension and attend to the idiomatic sequences used in the texts. Such practice enabled these students to notice and memorize more formulaic sequences than others and used them in their speech.

2) The accuracy rate as measured by the T-unit correctness rate

Table 2: T-Unit Correctness Rates of the Oral English Transcriptions

No.	Min	Max	Mean	Std Dev
70	0.00	0.81	0.3144	0.16150

Table Two sums up the T-unit correctness rates for the 70 students. It shows that the average accuracy rate was only a little bit above 30%, while the standard deviation exceeded 15%, indicating that many students could not correctly use their knowledge of grammar or could not pay attention to grammar when trying to convey ideas. This, again, reveals the weaknesses in English teaching in China's universities.

The same problem was found in the examination of the transcriptions. The students could only correctly used some simple sentences such as "he began to cough," "he felt very uncomfortable" and "he said that he didn't mind." Once the sentence was long, errors would occur. Many students used the Chinese conjunctive patterns like "although...but..." and "because...so..." Besides, many used "it's" for "it was" and "there's" for "there was," not knowing that "it's" and "there's" could only be used for simple present tense.

3) The effects on oral English of the knowledge of formulaic language and grammar

With the oral English test scores as the dependent variable and the number of formulaic sequences and the T-unit correctness rate as the independent variables, the linear regression (stepwise) procedures show that the T-unit correctness rate did not reach the level of statistical significance and therefore could not enter the regression model. In other words, the knowledge of grammar had no effect on the quality of the students' oral English.

The model kept the other variable: the number of formulaic sequences. For the narrow way of formula identification, the standardized coefficient (Beta) was: .676. The F value of the model (ANOVA) was: 57.284; the significance level was: .000.

For the broad way of formula identification, the standardized coefficient (Beta) was: .689. The F value of the model (ANOVA) was: 61.424; the significance level was: .000.

In either case, the null hypothesis could be rejected, and the model was valid. In other words, the number of formulaic sequences was significantly related to the oral English test score. It accounted for over 45% of the variance (R^2) in the oral English test score. This was a significant finding given the fact that this study only investigated the use of formulaic sequences in one of the three tasks in the National Spoken English

Test for English Majors (Band 4) whereas the score of this test was based on the student's performance on all these three tasks. To a great extent, the knowledge of spoken English is the knowledge of formulaic sequences; grammar is of secondary importance. The finding, therefore, supports Widdowson's view of learning lexical chunks as being more important than learning grammar (1989). It also supports Sinclair's view of the precedence of the idiom principle over the open choice principle (1991).

4) The variation between formulaic sequences used in the retelling task

Of the 79 formulaic sequences selected from the original script, each was by average used by 7.203 out of 70 students, and the standard deviation was 7.648. This indicates that there was a wide discrepancy between these sequences in the number of times each was used. For 19 sequences, more than 12 students used them in the task. Table Three shows that 11 of these 19 sequences, a majority, were adjectival, nominal, prepositional and verbal phrases that were closely related to the content of the story and held salient positions in the story. Five of them were time adverbials in the beginning of sentences, also a salient position.

TABLE 3: The Formulaic Sequences in the Script Used by More than 12 Students (No. = 19)

Type	Sequence	No. of students who used it
Closely related to the content of the story	cheap, clean and comfortable	24
	the same room	31
	a little noisy	24
	to build a new wing	18
	a little dusty	18
	borrowed a book	25
	from the hotel library	21
	very uncomfortable	14
	began to cough	27
	the whole building	16
	complain to the manager	13
Sentence-initial time adverbials	Whenever Mr. Smith	17
	During the first day	18
	The following afternoon	18
	At first	20
	after a while	19
Others	covered with	15
	was told	14
	at all	16

Table Four shows the distribution in the original story of the formulaic sequences listed in Table Three. The story began with the background information and with the explanation of the problem by the hotel manager. It used the simple present tense for the background part. Many students, however, inconsistently used past and present tenses or did not add the third person singular morpheme "-s" when using the simple present. The explanation of the problem by the hotel manager took the form of indirect quotations, whose structure was rather complicated to the students. The difficulties with tense and structure resulted in errors in language and in the relatively infrequent use (6) of formulaic sequences in the first one third of the text. The second one third in the middle of the story was mostly narration of the story's development. While retelling this part, the students were able to use many sequences that had occurred in the original script (10). When the story reached the climax in the last part, the language became complicated and hard to predict, and it was also possible the "fatigue effect" began to happen, so the students used very few sequences from the original text (3).

TABLE 4: Distribution of High Use Frequency Sequences (Used by More than 12 Students)(No. = 19)

	1st 1/3 (120 words)	2nd 1/3 (117 words)	Last 1/3 (117 words)
No. of sequences	6	10	3

Opposite to high use frequency sequences, there were also formulaic sequences few people used or used correctly. Of the 79 sequences, 28 were each correctly used by fewer than 3 students or were not used at all.

Several factors could have contributed to their under-use:
1) Errors in tense and number resulted in incorrect use of verbal phrases. Past tense morpheme was often dropped in speech, and only one student in the whole sample used "he stays" as having occurred in the original script while most used "he stayed" or even "he stay." As for the past perfect, only one student correctly used "had decided."

2) Many students lacked the knowledge of formulaic sequences, especially the knowledge of sequences that were different from their Chinese equivalent. For instance, most students did not use or did not use correctly such sequences as "situated at the far end of...," "so great was the demand," "no sooner had he...than...," "it looked as though," "landed on the floor," and "forced its way." In most cases, presumably, they could understand these sequences but could not use them when expressing the same ideas. When there were several sequences of a similar meaning, they chose the one that was close to Chinese. For instance, only one person used "went immediately" while most used "immediately went," only two used "all of a sudden" while many used "suddenly," and no one used "they both" while some used "the two of them."

3) Sometimes the students may not understand the formulaic sequences while listening to the story on tape. No one used "added apologetically" (some used "apologized" instead) or "bits of plaster," and only two used "a sharp metal tool."

Discussions and Conclusions

1) This study reveals the discrepancy between L2 learners in their command of formulaic sequences. This does prove that L2 learners are seriously lacking in the knowledge of formulaic language. However, it also shows that some learners are able to attend to formulaic sequences in the course of learning and have achieved a degree of success. In other words, if using correct learning strategies, adult classroom L2 learners are able to develop a good command of formulaic sequences even in foreign language settings. This is contrary to the rather pessimistic view of some researchers that adult L2 learners are unable to learn formulaic language and therefore will not be able to reach a high degree of fluency and idiomaticity. Wray (2002) has offered insightful explanations for the lack of idiomaticity on the part of L2 users. However, as to how these learners should learn formulaic sequences and improve the idiomaticity of their L2, she only mentions that they should reside and interact for some time in the L2 environment (Wray 2002: 210). No one would deny the usefulness of residing in an environment, but as the same time, this is obviously not a practical solution to most L2 learners. By offering this as a solution, it shows in itself that she sees no pedagogical solutions.

2) The effects of the knowledge of formulaic language on the quality of L2 oral performance, as found in this study, indicates that in addition to learning to use grammar rules, learning to use formulaic sequences should become an important component of learning a second language. This in fact should be a logical conclusion of the dual nature view of language, the view of language as having both analyticity and formulaicity (Skehan 1998). The native speaker's repertoire includes not only single words but also frequently used clauses and sentence patterns. These formulaic sequences are repeatedly and separately stored in the head according to their functions and situations (Wray 2002). Such a mode of storage enables the native speaker to avoid on-line assembling but fetch the ready-made language to cope with the real time pressure, achieving fluency and idiomaticity. This dual nature view of language points at the direction in which we should improve our language teaching practice.

3) Research is called for on how to help learners come to a good command of English formulaic sequences. A language contains numerous formulaic sequences, many of which even native speakers have not consciously noticed. It would not be practical to teach these sequences during the limited number of classroom teaching hours. In fact, students may not study these chunks the way some of them memorize vocabulary lists. One study (Ding & Qi 2001) found that learning texts by heart could help increase the learner's knowledge of formulaic sequences. Good learners pay attention to learning formulaic sequences with function and context. When they read a text aloud and learn it by heart, they attend to the way words collocate with other words and make these sequences into their own, to be used in their own speech and writing, resulting in rather fluent and idiomatic language.

References

Cowie, A. P. & Howarth, P. 1996. Phraseological competence and writing proficiency. In G. M. Blue & R. Mitchell (Eds.), *Language and education: Papers from the Annual Meeting of the British Association for Applied Linguistics held at the University of Southampton, September 1995* (pp.80-93). Clevedon, UK: Multilingual Matters.

Howarth, P. 1998. Phraseology and second language proficiency. *Applied Linguistics*, 19, 24-44.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Myles, F., Hooper, J., & Mitchell, R. 1998. Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, 48, 323-363.

Nattinger, J. R. & DeCarrico, J. S. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Pawley, A. & Syder, F. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-225). London: Longman.

Peters, A. 1983. *The units of language acquisition*. Cambridge: Cambridge University Press.

Sinclair, J. M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Skehan, P. 1998. *A cognitive approach to language learning*. Oxford, U.K.: Oxford University Press.

Weinert, R. 1995. The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, 16, 180-205.

Widdowson, H. 1989. Knowledge of language and ability for use. *Applied Linguistics*, 10, 128-137.

Wray, A. 2000. Formulaic sequences in second language teaching: principles and practice. *Applied Linguistics* 21, 463-89.

------. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

丁言仁、戚焱，2001，背诵课文在英语学习中的作用，《外国语》第5期，58-65页。

文秋芳、赵学熙、王文宇，2001，全国英语专业本科四级口试，上海外语教育出版社。

Attachment

Script of Task One (Retelling a Story) of the National Spoken English Test for English Majors (Band 4) in 2002

Whenever Mr. Smith goes to Westgate, he stays at the Grand Hotel. In spite of its name, it is really not very "grand," but it is cheap, clean and comfortable. Since he knows the manager well, he never has to go to the trouble of reserving a room. The fact is that he always gets the same room. It is situated at the far end of the building and overlooks a beautiful bay.

On his last visit, Mr. Smith was told that he could have his usual room, but the manager added apologetically that it might be a little noisy. So great was the demand for rooms, the manager said, that the hotel had decided to build a new wing. Mr. Smith said he did not mind. It amused him to think that the dear old Grand Hotel was making an effort to live up to its name.

During the first day, Mr. Smith hardly noticed the noise at all. The room was a little dusty, but that was natural. The following afternoon, he borrowed a book from the hotel library and went upstairs to read. No sooner had he sat down than he heard someone hammering loudly at the wall. At first he paid no attention, but after a while he began to feel very uncomfortable. His clothes were slowly covered with fine white power. Soon there was so much dust in the room that he began to cough. The hammering was now louder than ever and bits of plaster were coming away from the walls. It looked as though the whole building was going to fall. Mr. Smith went immediately to complain to the manager. They both returned to the room, but everything was very quiet. As they stood there looking at each other, Mr. Smith felt rather embarrassed for having dragged the manager all the way up the stairs for nothing. All of a sudden, the hammering began again and a large brick landed on the floor. Looking up, they saw a sharp metal tool had forced its way through the wall, making a very large hole right above the bed!

RESEARCH ON PARALLEL CORPUS BASED CHINESE-ENGLISH WORD TRANSLATION MINER[1]

Yang Muyun*, Wang Lixin+, Zhao Tiejun*, Liu Xiaoyue*
*School of Computer Science and Technology
+Foreign Languages Department
Harbin Institute of Technology

Abstract: This paper proposes a semi-automatic tool, Chinese-English Word Translation Miner, to assist professional translators or lexicographers to determine the word translations from Chinese-English parallel corpus. The tool adopts statistical natural language processing techniques to resolve co-occurrence measure, multi-word unit translation and indirection association problem. Its algorithm focuses on extracting as many as candidate translations while preserving a high precision. Although it is far from perfect, experiment results indicate the effectiveness of our method and the potentiality of the lexicon builder system.

Key words: Chinese-English parallel corpus, word translation miner, statistical approach

1. Introduction

The statistical based natural language processing has produced a number of helpful technologies for corpus linguistics, which greatly benefits professionals in language teaching, dictionary compilation and other language industries. One of the non-trivial progresses during this process is the computer-aided compiling tool for lexicographers. Even cross-language lexicographers have already been armed with software providing word translations, term candidates and bilingual concordancer.

In fact, many efforts have been made on building the translation lexicons automatically from bilingual corpus, such as BICORD, Champollion, Termight etc. Though these projects are based on different corpora and use a variety of methods, a common strategy can be summarized as:

- Choose a function F to measure the correspondence between words in source language (Ls) and those in target language (Lt);
- Compute F(s, t) for word pairs, in which $s \in Ls$ and $t \in Lt$;
- Choose a threshold δ and output all entries whose $F(s, t) > \delta$.

It should be pointed out that most of these researches were carried between western languages. Though being two of the most widely spoken languages, Chinese and English are less touched by word translation auto-extraction researchers, let alone commercial software tool. This paper focuses on auto-construction of Chinese-English dictionary form the parallel corpus by statistical approach. A Chinese English Translation Miner is further designed and implemented to assist lexicographers in compiling bilinguai dictionary as well as relevant tasks like MT, CLIR etc.
The rest of this paper is arranged as follows. Section2 sketches the principles and statistical techniques adopted in Chinese-English dictionary auto-extraction, and discussed the problem and our solutions. Section 3 presents an overview of the Miner, focusing on modules specially designed for lexicographers. And, finally, section4 briefly demonstrates the performance of the Miner system with experiments.

2. Statistical Approaches To Chinese-English Lexicon Extraction

The basic hypothesis of statistical word translation extraction is that a word is more often than not to co-occur with its translation(s) in the beads[2]. Therefore co-occurrence is a clear indicator of the word correspondence cross language. It should be noted that different way of employing co-occurrence statistics will turn out different

[1]Supported by the High Technology Research and Development Program of China (2002AA117010-09) and the Scientific Research Foundation of Harbin Institute of Technology (HIT.2001.ARQQ18000104)
[2] Bead is the minimal aligned sentence pair, e.g. a Chinese sentence with its English counterpart.

results. So choosing a proper model is crucial to the final accuracy of the acquired translation candidates.

In practice, there are four models that are most frequently applied with good results in estimating word associations, i.e. Dice coefficient, mutual information, contingency table and log likelihood ratio (see Appendix for detail of the formulas). Being statistical model in essence, they are faced with some common problems: high frequency words, inflection form and frequency discrepancy[1].

High frequency words are those words appear in nearly each sentence, e.g. "the, a, of" in English and "的, 在"in Chinese. The trouble of these words lies in their high co-occurrence with most of the words and thus becoming "universal" translation candidates. The simplest way of solution is to create a "stop word" list for such words and removed them from calculation.

Inflection form of an English word would divide the count of its co-occurrence with its Chinese translation. To prevent this, English words have to be lemmatized before calculation. Tools available for this purpose usually bears the accuracy of 98% or so.

Frequency discrepancy refers to that different co-occurrence may produce similar calculation results. Generally speaking, more statistical data means greater reliability. So we adopt an iteration strategy to solve this problem: a greedy strategy that deletes the most reliable bilingual entry (with highest score) in each iteration and re-calculate. The iteration algorithm can be described as:

1) Input: Chinese-English parallel corpus
2) Preprocess;
3) Calculate bilingual entries by the chosen model;
4) Select the top *n* bilingual entries;
5) Delete the chosen entries from the parallel corpus;
6) If not meet the stop condition, repeat step 3.

3 System Overview

Basically, the Chinese-English Word Translation Miner consists of three components, i.e. pre-processing, statistical word translation extraction and output (see Fig. 1).

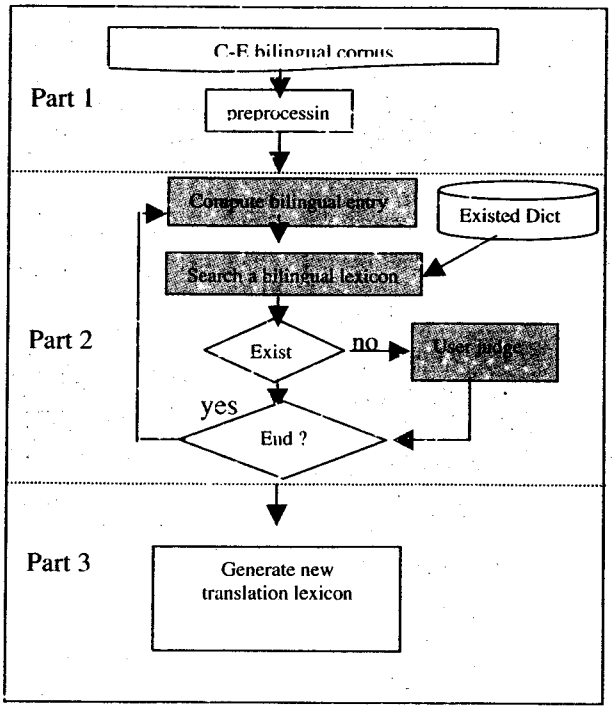


Figure 1: Frame Chinese-English Word Translation Miner

[1] These problems are also termed as "indirect association" by some researchers.

The input of the Miner is sentence aligned Chinese English parallel corpus. The pre-processing module performs lemmatization for English and word segmentation for Chinese This step facilitates direct estimation of word to word associations.

The key component in the builder is the statistical word translation extraction. Beside the algorithm described in above section, feedback from expert (if any) is fully utilized. The Miner could consult the user for the correctness of the chosen entry. Any word pair would never bother the user again once it is specified as "wrong". Also user could specify an existed Chinese-English dictionary as the reference for the Miner, which would save great time for new translation detection purpose.

The last part of Miner is the output module, which is responsible for generating the auto-acquired Chinese-English dictionary form the parallel corpus (with or without the interference of lexicographers).

4 Experiments And Discussions

A parallel Chinese-English corpus with 30094 beads is constructed to evaluate the Miner system. Table 1 lists the details of the corpus. It should be noted that stop-words (appeared more than 1000 times in the corpus) has been removed in the following experiment.

Table 1: Chinese English Bilingual Corpus Statistics

	Tokens	Types	Types in word pairs occurred twice or more
Chinese	380.524	17.711	10195
English	324.302	10.688	6115

The first experiment is to choose among Dice coefficient, mutual information, contingency table and log likelihood ratio for the Miner system. Top 10 translation candidates output by these models are considered for all Chinese words. For a given Chinese word (Wc), the evaluation function is so designed as the weighed sum of the correctness and the position of its translation candidates:

$$\text{Score}(Wc) = \sum_{i=1}^{k} g_i * f(i)$$

in which $k \in \{1,2,3,4,5,6,7,8,9,10\}$, $f(i) = 1.1 - 0.1 * i$, $g=1$ if the translation is correct or 0.5 if partly correct. The total score of a model would be the sum of all the Chinese words' score. And the accuracy would be the model score divide by the model score if all candidates were correct translations.

Table 2 lists the evaluation results of the 4 models when applied to the test corpus. It is clear from the table that log-likelihood model performs best in each index. And thus it is decided as the kernel model of the Miner.

Table 2: Performances of 4 Models

	Dice	MI	CT	
Correct translations	8557	8325	8560	**8596**
Part correct trans	2030	1983	2047	**2061**
Model score	8571.65	8038.65	8591.05	**8742.70**
Accuracy	29.83%	27.97%	29.89%	**30.42%**

The second experiment is designed to test contribution of the adopted iteration strategy. Top 5000 entries got by no iteration, 5 iterations with top 1000 candidates selected each time, 10 iterations with top 500 candidates selected each time and 5000 iterations with only top 1 candidates selected are obtained and evaluated respectively. Above mentioned evaluation function are also applied (see Table 6).

Table 3: Performances of Different Iteration

	Direct result	n=1000	n=500	n=1
Correct translations	3412	3949	4049	4193
Part correct trans	397	453	435	506
Total Score	**3520.9**	**4061.7**	**4148.5**	**4324.8**

Table 3 indicates that the greedy strategy is effective according to the higher total score. And the best result is achieved when only top 1 candidate is selected during each iteration. Table 7 further indicated that the best result of log-model is more than doubled compared with no iteration performance. And the nearly 77% of the Chinese word in the bilingual corpus are found with correct English translations.

Table 7: Performance of Log-model with and without Iteration

	Direct Method	Iteration (n=1)
Chinese word number	8587	7855
Correct translations	8596	8437
Part correct trans	2061	1608
Total Score	8742.7	8693.050
Accuracy	**30.42%**	**65.10%**

It should be pointed out the 65% accuracy and 77% recall would be the expected performance of the Miner system in application. Though this performance is far from satisfaction, it is feasible to be applied to semi-automatically construct Chinese-English dictionary from the bilingual corpus. A research project finished in our lab only just proved that such dictionary for Chinese-English bilingual corpus is substantial to MT quality improvement.

To sum up, this paper presents a Chinese-English Word Translation Miner that can help lexicographers and other professionals to auto-construct a translation lexicon from a parallel corpus. It can also be applied in cross-language processing e.g. MT, CLIR etc. Further research lies in the following directions:

- Identify low frequency translation correspondence using context information (like word alignment);
- Translation unit identification before conducting co-occurrence count;
- Term identification for some practical application;

Acknowledgements

My thank goes to Kang Zhenguo, Hansong Dan, Ma Ye who manually checked all the bilingual entries in the experiment.

References

J. Klavans and E. Tzoukerman. The BICORD System: Combining Lexical Information from Bilingual Corpora and Machine-readable Dictionaries. Proc. of COLING'90. 1990: 174~179

Ido Dagan and Kenneth W. Church. Termight: Identifying and Translating Technical Terminology. In Proceeding of the 4th Conference on Applied Natural Language Processing, Stuttgart/Germany,1994. ACL

F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translation Collocations for Bilingual lexicons: A Statistical Approach. Computational Linguistics. 1996, 22(1): 1~38

I. D. Melamed. Empirical Methods for Exploiting Parallel Texts. Ph.D. Dissertation of University of Pennsylvania. 1998

L. R. Dice. Measures of the Amount of Ecologic Associations Between Species. Journal of Ecology. 1945(26).

K. W. Church, P. Hanks. Word Association Norms, Mutual Information and Lexicography. Computational Linguistics 1991, 16(1)

W. A. Gale, K. W. Church. Identifying Word Correspondences in Parallel Texts. Proceedings of the 4th DARPA Workshop on Speech and Natural Language. 1991: 152-157

T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics. 1993, 19:61-74

Yang Muyun. Chinese English Sentence Alignment and Acquisition of translation dictionary and translation rule. Ph.D. Dissertation of Harbin Institute of Technology. 2003

Corpus-based Dictionary Illustrative Citation System: A Resource and a Methodology

Li Dejun

School of Foreign Languages, Nanjing University

Abstract: This paper introduces a new system for dictionary compilation or lexicographic research based on a parallel corpus. It focuses on how the system is designed or employed to facilitate drawing illustrative citations automatically and at the same time avoid some of the known problems arising from corpus means.

Key Words: Illustrative citation; monolingual or parallel corpus; CpsDict system

1. Introduction

Illustrative citations are an integral part of a dictionary. The importance of citations has been observed by many lexicographers (Zgusta 1983: 360-361; Kasimi 1983: 91). Generally, illustrative citations in a dictionary are either coined by the compilers or selected from publications. If dictionary compilers don't want to make up their own illustrative citations, they have to take to authentic material for illustrative examples, which is usually a hard job and very time-consuming. With the emergence of corpus and corpus tools, dictionary compilers have found a new way to do the old job. But there are at least two problems that make the new method unfriendly to users, especially to bilingual dictionary makers. One is that the large data of the commercial corpus such as the BNC cannot be used as a resource for illustrative citations. The copyright law forbids you to do so. So researchers have to design and build their own specific corpus if they don't want to give up this new method. But after their specific corpus is built, they will face another problem. The corpus tools, both TACT and Wordsmith, can only be used to process the English language only. They can do nothing about bilingual or parallel corpus, which carries special importance for bilingual dictionary compilation.

In the following, we will describe briefly a new system CpsDict that has been developed recently by the author himself, which is meant to be used mainly for bilingual dictionary compilation and research.

2. CpsDict as a resource

In this system, there is a built-in parallel corpus[1], which provides real world documented evidence of how utterances or samples of text in Chinese can be rendered into English, or vice versa. Parallel corpus is thus a valuable repository of data on cross-language usage and can be exploited by a number of practical and theoretical applications.

Take English corpora BNC, Brown, LLELC and Modern Chinese Corpus as references, we have the overali structure and its various constituent genre subject fields as outlined in Table 1.

Text Type	Proportion of the sub-corpus (%)
Press Reportage Editorial Reviews	14
Fiction General fiction novels short stories Historical fiction Science fiction Adventure Folklore	40

Social science Sociology Geography Anthropology Law Education Linguistics	12
Commerce and finance Business Economics Finance Industry	8
History and Belief History Religion Philosophy Mythology Occult	10
Politics and Military Affairs Politics Military Government	4
Arts Visual arts Architecture Performing Media Design	7
Leisure Food Travel Fashion Sports, etc.	5

From the above table, we can see that natural science and applied science are not included. Poetry is not included either. Since the language of pure science is a bit too technical and that of poetry is out of the ordinary, illustrative citations selected from them may not be proper for a general-purpose dictionary. So we don't want to have them built into a two-million-word corpus.

3. CpsDict as a Methodology

Corpus can facilitate dictionary making, especially citations. But at the same time there arise some problems: unscientific and atypical citations; re-selection of the same sentences as citations; overmany examples which make the selection of citations difficult. CpsDict system makes proper disposal of most of the problems found.

3.1 It embraces both an open and a close system

In the system, the built-in parallel corpus is a close databank. The users can only do their different kinds of searches in this databank. They are not allowed to change or modify it (except "Selected" marks, see 3.5). A close system can ensure the integrity of databank.

In order to facilitate users to use their own specific corpus, the system permits them to add text materials under a different window. Users can first of all edit English, Chinese or bilingual texts and then save them as text format or Rich Text Format. The system can open these data in the window through an ordinary open file command. Once the material text has been opened, different kinds of searches can be done in the system.

3.2 It can process monolingual or bilingual texts

The system can process not only monolingual (English or Chinese) texts but also bilingual (English and Chinese) texts that are especially important for bilingual dictionary compilers.

Key words search based on SQL searching language has been simplified for ordinary users. What the users have to do is to key in a key word (headword in a dictionary) in the textbox, and the system will search automatically in this databank and display all results on the screen. For example:

Sample one (the first five searching results for the key word "Chinese culture" from a user-designed monolingual corpus)

a. Chinese culture is rich and profound.
b. Another area of the profundity of Chinese culture is her pre-industrial revolution science and technology.
c. The richness of Chinese culture also finds expression in its diversity and pluralism.
d. The diversity and pluralism of Chinese culture is a tremendous asset.
e. Chinese culture is a complete system, including its own philosophy, literature and arts, medicine, technology and science as well as language and festivals.

Sample two (the first three searching results for the key word "小康" from a user-designed bilingual corpus)

The theme of the congress is to...build a well-off society in an all-round way... 大会的主题是：……全面建设小康社会……

On the whole, the people have reached a well-off standard of living. 人民生活总体上达到小康水平。

As human society entered the 21st century, we started a new phase of development for building a well-off society in an all-round way and speeding up socialist modernization. 当人类社会跨入二十一世纪的时候，我国进入全面建设小康社会、加快推进社会主义现代化的新的发展阶段。

3.3 It has the function of saving records selectively or wholly

The system can search automatically for all examples based on a key word. After a key word is keyed in into the search box, all the illustrative examples will be shown by pressing "Search for All" button. Users can now save the result as a text file or a RTF file.

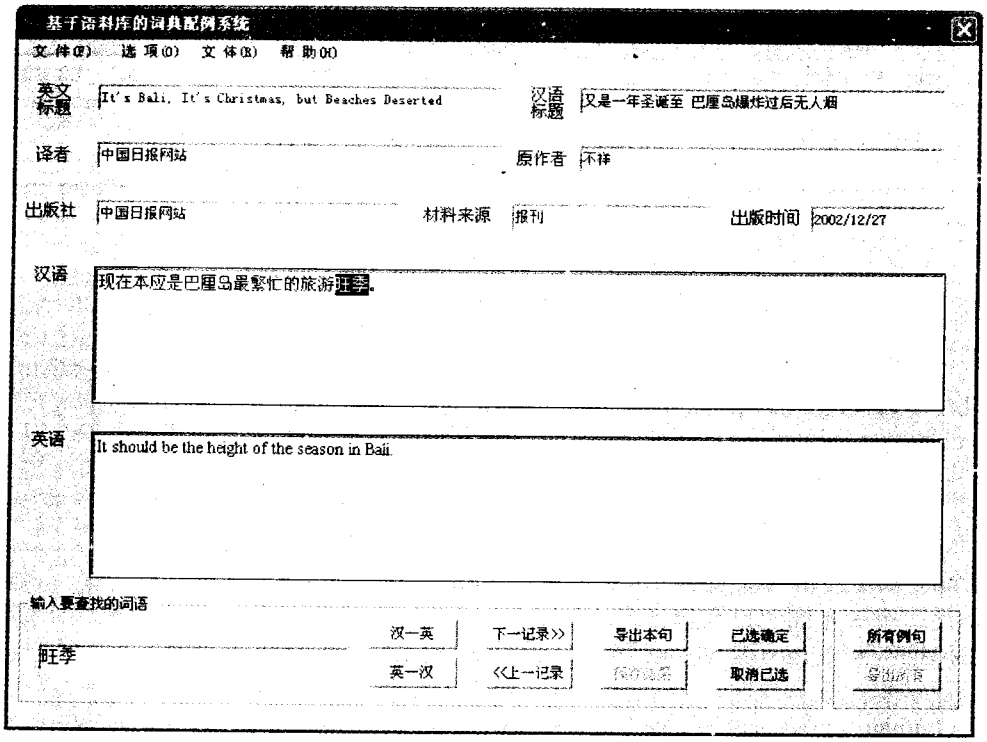
The system also support sentence-by-sentence search. You can press "Search Begins" button to get the first illustrative example from the databank.

If you want to go on for the next example, you may press "Search for Next" button and the result will be displayed. If you find the example unsuitable for your dictionary, you can skip it by pressing "Search for Next" button again. The current example will be dropped and next example will be displayed before you. In this way, users can make their choice and get the best examples. When the search ends (Users can stop a search any time he likes), the result will be saved excluding the skipped examples. The following are the first four searching results for the key word "language", you may choose the ones which are most suitable for you.

a. Chinese culture is a complete system, including its own philosophy, literature and arts, medicine, technology and science as well as language and festivals.
b. At that time India was divided into many states with different traditions and languages.
c. He ordered his Prime Minister Li Si to sort out all the different systems of writing hitherto prevalent in different parts of the country so as to unify the written language under one system.
d. Ming Dynasty artisans used the succinct language of art to express their inner feelings.

3.4 The system can provide detailed additional information for illustrative examples and their translated versions.

In the system, if we make a key word search with "旺季", we can get one of the search results as follows:



From the above diagram, we can see that additional information as source, author, translator, time are shown on the screen together with the bilingual text. The search result can be saved as a text file. The output text for the above example is like the following:

> It should be the height of the season in Bali.现在本应是巴厘岛最繁忙的旅游旺季。(标题：It's Bali, It's Christmas, but Beaches Deserted又是一年圣诞至 巴厘岛爆炸过后无人烟 作者：不详；译者：中国日报网站；出版社：中国日报网站；出版时间：2002/12/27；材料来源：报刊。)

Typicality of citations is one of the most important principles put forward by lexicographers (Fox 1987: 138, 143). Chen Chuxiang, a scholar in China, said it is one of the ten standards for a good dictionary (Chen 1994: 18). The additional information provided can help dictionary compilers to choose the most typical illustrative examples because semantic, cultural and contextual information are helpful in the choice of examples.

3.5 The system is on the alert to re-selection of same sentences as citations

Re-selection of same sentences is a big headache for dictionary compilers. Dictionary compilation is usually teamwork, so it is very hard and almost impossible to avoid re-selection, which is especially true when the dictionary compilers draw examples from the same corpus.

In the CpsDict system, if a sentence is finally selected as a citation example, the sentence will be marked conspicuously. When this sentence is selected again, compilers will get to know at a glance that this is a re-selection. If an example is marked by mistake or else, users of the system can press "Clear Selection" button to clear the mark.

Re-selection information is very important for dictionary compilers. In this system users can backup this important information at any moment. All data can be restored in case of an unexpected computer breakdown.

In order to make the built-in corpus possible for future or other uses (when the current work is finally ended), the built-in corpus of the system has been specially designed. It can be reset by users to its original state.

4. Conclusion

The objective of this article is to demonstrate that CpsDict can provide valuable sources of data for dictionary compilers. And we have described procedures designed to provide user-friendly access to different types of corpora. The procedures have been tested on samples of English, Chinese and English-Chinese. The construction of high quality monolingual or bilingual corpora is of cardinal importance. If users find the source of the built-in corpus not rich enough for their research, they may design and build their own specific corpus, and thus having a more effective use of the system.

Notes

① CpsDict is based on NG Parallel Corpus(NanDa-Guoguan Parallel Corpus) that is developed by teachers mainly from Nanjing University. The corpus embraces several sub-corpora and it is planned to cover 2 million words by the end of 2004. So far, the first stage of work, creation of a sub-corpus (language of the press) has been finished.

References

Al-Kasimi, Ali M. *Lingusitics and Bilingual Dictionaries*[M]. E.J. Brill: Leiden, 1983.

Bejoint, H. *Modern Lexicography: An Introduction*, Beijing: Foreign Language Teaching and Research Press, 2002.

Biber, D. Representativeness in Corpus Design[J]. *Literary and Linguistic Computing*, 1993, 8(4): 243-257.

Bogaards, Paul. Dictionaries for Learners of English[J]. *International Journal of Lexicography*, 1996, Vol. 9, No. 4.

Fox, Gwenyth. The Case for Examples[A], J. M. Sinclair (Ed.), *Looking Up*[C], London: Collins ELT, 1987.

Halverson, S. Translation Studies and Representative Corpora: Establishing Links Between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study[J]. 1998, *Meta* XLIII, 4.

Johansson, Stig. Times Change, and so do Corpus[A]. In Karin Aijmer & Bengt Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*[C], London and New York: Longman, 1991.

Kennedy, G. *An Introduction to Corpus Linguistics* [M]. Beijing: Foreign Language Teaching and Research Press, 2000.

Ming, Li & Jinghua, Zhou. *An Introduction to Bilingual Lexicography*, Shanghai: Shanghai Foreign Language Education Press, 2001.

Summers, D. Computer lexicography--the importance of representativness in relation to frequency[A], Jenny Thomas & Mick Short (Ed.), *Using Corpora for Language Research*[C], Beijing: Foreign Language Teaching and Research Press, 2001.

陈楚祥. 词典评价标准十题[J]，辞书研究，1994，(1).

郭启新. 论语料库与英汉词典配例[A]，张柏然，魏向清. 双语词典学论集[C]. 南京：江苏教育出版社，2001.

刘连元. 现代汉语语料库研制[A]，陈原. 汉语语言文字信息处理[C]. 上海：上海教育出版社，1997.

兹古斯塔. 词典学概论[M](林书武等译). 北京：商务印书馆, 1983.

A Genre Analysis of Research Article Abstracts across Disciplines

Ge Dongmei & Yan Xiaoqiang
Xi'an Jiao Tong University

Abstract Due to the particular role abstracts play in research article reading and publication, it is essential for Chinese researchers to understand and master the genre knowledge of English research article abstracts. But heretofore there has been little empirical work done in analyzing the genre of research article abstracts, still less on the comparison of research article abstracts across disciplines. In this study a stratified random sampling was conducted, and totally 150 abstracts in three disciplines were obtained. The moves and linguistic features of these abstracts were identified and compared in relation to disciplines. The research comes up with fruitful results in exploring a relatively ignored genre and thus greatly enriches the ESP Genre Analyses and corpus researches. Meanwhile, it can provide help for Chinese researchers majoring in these three disciplines to read and write research article abstracts in their own fields more efficiently.

Key words: genre analysis, RA abstract

1. Introduction

With the introduction of the concept "genre" in linguistics, genre analysis, which relates the linguistic features of a genre to the actions they perform, has aroused increasing interests among applied linguists and ESP teachers, who have found genre theories particular useful in analyzing the technical genre (Crookes, 1986) and business genre (Berkenkotter & Huckin, 1995) as well as in ESP teaching (Paltridge, 1996).

As a form of technical writing, the research article abstract (RA abstract) is very important to researchers in that it enables them not only to be informed of the latest development in their fields but also to make their own researches known by their counterparts all over the world in the shortest time. As is frequently read and written by researchers, RA abstracts deserve special attention of genre analysts and ESP teachers.

Unfortunately, however, it has been neglected (Swales, 1990: 181). There are not a small number of genre studies on research articles and its conventional parts---Introduction, Method, Results and Discussion (Brett, 1994; Holmes, 1997; Nowgu, 1997; Posteguillo, 1999). In contrast, genre analyses to RA abstracts are scant. What's more, it is noticeable that regarding the rhetorical structure and the language features of RA abstracts, there are much more prescriptive statements than descriptive analyses. There are still fewer studies devoted to the variation of discourse structures of RA abstracts across disciplines. Therefore, structural patterns as well as linguistic features of RA abstracts in different disciplines should be under cautious analysis.

The objective of this study is twofold: first, to give an accurate description of what actually constitutes the features of an RA abstract and a clear idea of the disciplinary effect on these features; second, to shed some light on ESP teaching and future corpus researches.

2. Method

2.1 The corpus

In this study, the corpus was carefully developed by scientific sampling. First, with reference to the table of enrolment issued by Xi'an JiaoTong University for master's degree in the year 2002, the disciplines which admitted the largest number of postgraduates in the three main fields of this university – engineering, economics and medicine – were chosen. They are Electrical & Electronic Engineering (E), Financial Economics (F), and Surgery (S). Next, the electronic journals in the Elsevier Science (http://www.elsevier.com) were browsed and 26 core journals were found accessible, among which there were ten in E (E1-E10) and F (F1-F10) and six in S (S1-S6) (*The list of our chose journals will be available*

on request). At last, with the help of the table of random numbers, a stratified random sampling was conducted, and totally 150 abstracts (50 in each discipline) were obtained.

2.2 The identification of the Moves

In identifying the schematic structure of these abstracts, we followed the procedure suggested by Nwogu (1997). The four-move pattern proposed by Bhatia (1993) was adopted at the beginning and 30 abstracts were analyzed. Based on the results got in the preliminary analysis, we modified Bhatia's model and identified five moves, the working definitions of which and the examples are listed as follows (The italicized words are the lexical signals. E2-4 refers to the forth abstract in the second Engineering journal):

Move1 Providing Background Information (B) – To make a theoretical or situational preparation for the present research.

(1) The memory intensive nature of object-oriented language...*has created the need* of... Thus, high-performance memory manager *is needed to cope with* such applications. (E2-4)
(2) Personal savings as a percentage of disposable income *have dropped steadily since the early 1980s.* Savings *have continued to decline in 1999*... (F1-3)
(3) Hypertension *is a known risk factor* in heart disease. (S3-5)

Move2 Announcing Present Work (A) – To give a precise indication of what forms the basis of the present research. Either the goals of the research are reported or a general statement about the research is made.

(1) *Our aim was to* develop a risk score for prediction of ... (S2-6)
(2) *This is a study of* the transmission pattern of inflation under alternative exchange rate fegimes, fixed and flexible, among the G-7 countries and their subsets... (F2-3)
(3) *In this paper* a method for ... *is presented.* (E2-1)

Move3 Describing Methodology (M) – To describe the methodology adopted. It may include the subjects under investigation or the equipment and materials used or the steps taken in the research.

(1) *By varying the sequential order*, target descriptions were collected in four contexts. (E10-2)
(2) *Samples of future prices for gold, silver, and copper, and the realized cash or delivery settle prices, based on fixed maturities for a cross-section of contracts are used...* (F10-2)
(3) After left thorocotomy, we *performed* ...and *followed...was performed...followed with...before prior to* the 3-hours...*were determined*...and *we recorded... We performed...at the end of the ... We also assessed the* ... (S5-3)

Move 4 Presenting Findings (F) - The results may be explicit and detailed or they may be general.

(1) PET-FDG imaging correctly *identified* nodal stage (N0-N1 vs. N2) *in 50 out of 61 patients (82%),* overstaging occurred in *eight patients (13%),* and understaging in *three patients (4.9%).* (S4-2)
(2) The growth rate ... *is shown to be* less uniform and symmetric... (E4-4)

Move5 Drawing Conclusions (C) – To summarize the research as a whole. In this move, researchers may give explanations to the result, draw inferences from it, make recommendation for future study or draw implications in the light of practical application.

(1) *This approach offers great potential* for adaptation... (E10-3)
(2) This *is probably due to* the entrenchment of managers and... (F5-1)
(3) These data and those from three other centers *support the conclusion* that... (S3-1)
(4) *A realistic application* of the proposed technique...*will be explored.* (E7-3)

It can be seen from these examples that in this study, a move was identified based on partly by inference from context, but mostly by reference to linguistic clues in the discourse. These linguistic signals were of great importance in deciding the boundary of the moves. A sentence (or sentences) in an abstract was categorized as a move based on its (their) salient function.
Although facing readers whose interests are in different fields, these abstracts have a same purpose, which

is to give enough accurate information to let readers know the important data contained within the articles. Therefore, the schematic structures of all the abstracts were analyzed in terms of the above-mentioned moves no matter which discipline they fall into.

2.3 Data collection

Since a move is "a text segment made up of a bundle of linguistic features which give the segment a uniform orientation and signal the context of discourse in it" (Nwogu, 1991: 114), it is important to study the linguistic features in each move to have a good understanding of this particular genre. This study was limited in those linguistic features which can easily be observed in abstracts and which are often too much taken for granted rather than carefully studied, including modal verbs, the tense and voice of verbs and the use of first person pronouns.

A sample is given in appendix to show the work done on each abstract.

Every abstract was marked this way and totally 150 tables were obtained. Then data in these tables were grouped and processed according to the discipline, followed by a comparison across disciplines. Chi-square analysis was performed to help us to find out the probable disciplinary effects on the generic structure and linguistic features of abstracts and the influence of different moves on the linguistic features.

3. Results and discussion

3.1 The generic structure

3.1.1 The order of moves

As one of the aims of genre analysis, to identify the allowable order of moves in these RA abstracts is necessary. The results are as follows:

1. Moves are sequential in order in 129 (E: 42; F: 40; S: 47) out of 150 abstracts, that is B-A-M-F-C.
2. In 5 (E: 1; F: 3; S: 1) abstracts, the first two moves are in reverse order, that is A-B. Here are some examples:

   (1) *In this paper, we develop* different mathematical models in the framework of ...*and discuss* their...*Previous ASR tests have shown* that... (E10-3)
   (2) *This paper examines* the takeover charter amendments made by 128 firms...*By December 31, 1995, firms were to have adopted* one of three charter amendments that... (F6-4)
   (3) *The purpose* of this article is to...The SFPV *has proven to be* resistant to...and *has shown* no signs of degeneration over the long term. (S3-1)

3. The rest 16 abstracts contain cyclical patterns. For example:

   (1) The algorithm ... eventually reaches a near-optimal or optimal solution. The proposed method is practical as it can handle many practical constraints such as... Experiment results show that... (E1-4: F-C-F)
   (2) Several violations of these properties are found... The hedging performance of the American options is evaluated by constructing delta-neutral and delta-vega-neutral portfolios. The empirical performance of these strategies is sometimes bad. (F9-1: F-M-F)
   (3) Nine patients were divided into two groups according to the criteria... In the group with SIRS... occurred in three of the four patients (75%). All aneurysms were resected with a small part as a remnant...Three patients died after surgery. (S3-2: M-F-M-F)

3.1.2 The frequencies of moves

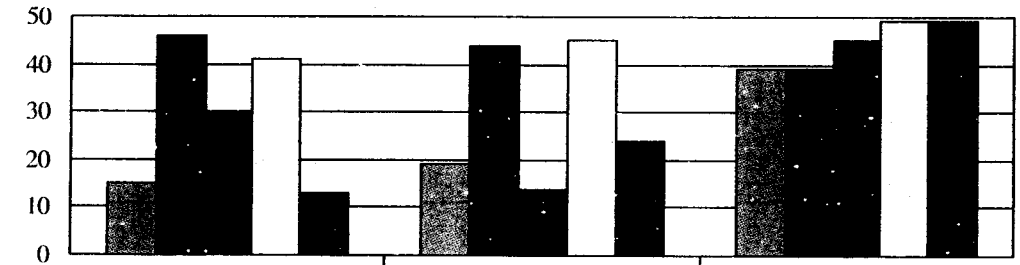
As is shown in Table 1 and Figure 1, there are significant differences in the generic structure of abstracts across disciplines (χ^2 =31.183, *df*=8, $p<0.05$). From Table 1 we can see that abstracts of different interest fields seem to be of different complexity. In E, only one abstract contains all the five moves, and almost

half of them have only three moves. In F, things are a little different – no abstract has only one move or all the five moves. Abstracts in Surgery, however, seem to be most complicated, with half of them having five moves and 22 having four moves.

Table 1: Number of abstracts having different number of moves

	5 moves	4 moves	3 moves	2 moves	1 move
E	1	11	23	11	4
F	-	12	22	16	-
S	25	22	2	1	-

Figure 1: Comparison of moves in abstracts across disciplines



	E	F	S
B	15	19	39
A	46	44	39
M	30	14	45
F	41	45	49
C	13	24	49

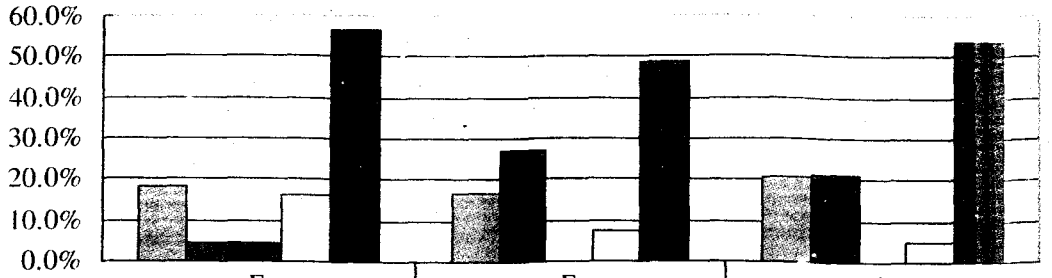
Figure 1 is a clearer illustration of the disciplinary differences. Although no move is completely obligatory in abstracts in any of the disciplines, the numbers of stable moves (those can be observed in more than half of the abstracts) in each discipline are different. Three moves are found stable in abstracts in E, which are Move2, Move3, and Move4. Since it has been stated in 3.1.1 that moves are sequential in most abstracts, it may be concluded that the schematic structure of abstracts in E can be presented as A-M-F. Similarly, with only two stable moves – Move2 and Move4 and a move that occurs in almost half of the abstracts, the structure of abstracts from F1-F10 can be expressed as A-F-(C). The fact that studies conducted in this field do not often involve precisely described procedures and relatively clear criteria of acceptability may account for the lack of Move3. On the other hand, all the five moves are found stable in abstracts in S, which makes the generic structure of these abstracts as B-A-M-F-C. Move4 and Move5 can be regarded as obligatory moves – they occur in all but one abstract. It seems that abstracts in Surgery are more "standardized" than those in the other two fields. One of the reasons may be that participants in this field have been accustomed to the conventional form of the RA abstract since we found that four out of these six journals have their own policies, and these policies, with slight differences, all require the authors to present abstracts with all five moves.

3.2 Linguistic features

3.2.1 Modal verbs

To make the frequency of modal verbs comparable, the number of modal verbs in each move was divided by the total number of the verbs in this move. Our results show that modal verbs in these abstracts are more move-determined ($p<0.05$, $df=4$) than discipline-determined ($\chi^2=0.157$, $p>0.05$, $df=8$).

Figure 2: The distribution of modal verbs



	E	F	S
■B	18.3%	16.8%	20.6%
■A	4.5%	27.0%	21.1%
■M	4.5%	0.0%	0.0%
□F	16.1%	7.3%	4.7%
■C	56.7%	48.9%	53.6%

A clear map of the distribution of modal verbs can be seen in Figure 2, where the total number of modal verbs in each move is transformed into the percentage it accounts for. It shows clearly that the difference of distribution of modal verbs among the three disciplines is not statistically significant, and that modal verbs are very unevenly distributed across moves. The last move is related to modal verbs most closely, which is in accordance with the communicative function of this move (to draw conclusions from the result) and the functions of modal verbs (to express tentativeness and possibility). The third move, which is primarily devoted to describing methodology, needs the least degree of being tentative. Accordingly, the density of modal verbs in this move is lowest.

3.2.2 Verb Voice

To testify whether there are really more passive verbs than active verbs in RA abstracts, the frequency counts of voices in each move of each abstract were recorded. By calculating the percentage of active verbs and passive verbs in each move, we got Table 2.

Table 2: Verb voices

		B	A	M	F	C
E	A(%)	**67.2**	**52.1**	29.8	**70.6**	**56.7**
	P(%)	32.8	47.9	**70.2**	29.4	43.3
F	A(%)	**88.1**	**94.3**	**61.8**	**83.8**	**88.1**
	P(%)	11.9	5.7	38.2	16.2	11.9
S	A(%)	**72.2**	**75.0**	40.9	**78.8**	**79.1**
	P(%)	27.8	25.0	**59.1**	21.2	20.9

*A: active verbs P: passive verbs bold forms indicate the major voice.

Table 2 shows that in all the five moves of abstracts in F, there are more active verbs than passive verbs. There is, however, significant difference among the moves (χ^2=12.226, *df*=4, p<0.05), which is caused mainly by Move3. While there are far more active verbs than passive ones in other moves, verbs in this move do not show so strong a voice preference. In other two disciplines, Move3 is also particular in that it is the only move where there are more passive verbs than active ones. A probable explanation for this is that when the method of the research is reported, to eliminate the subjectivity that may be conveyed by an active sentence with an animate subject, abstract writers tend to use passive sentences. Given the disciplinary differences shown in Table 2 and the particularity of Move3, it can be deduced that the voice of verbs has much to do with the communicative function of different moves as well as the interest field an abstract is in.

3.2.3 Verb tense

Although it is suggested in many books that present simple tense be used predominantly in scientific writing, there is still a need to confirm whether it is true to all scientific genres. In the present corpus, totally six verb tenses were discovered, which are present, past, present perfect, present progressive, past perfect, and future. But since there is only one verb in future tense and past perfect and only three in progressive, these tenses are neglected here.

Table 3: Verb tenses

		B	A	M	F	C
E	**pre.** (%)	**88.7**	**83.5**	**66.3**	**82.8**	**86.4**
	p. (%)	1.9	7.7	28.3	12.7	9.1
	p.p (%)	9.4	8.8	5.4	4.5	4.5
F	**Pre.** (%)	**63.2**	**100**	**100**	**96.4**	**97.8**
	p. (%)	10.5	0	0	2.7	2.2
	p.p (%)	26.3	0	0	0.9	0
S	pre. (%)	**68.3**	43.5	4.1	2.2	**78.5**
	p. (%)	4.9	54.3	**95.9**	**97.1**	20.7
	p.p (%)	26.8	2.2	0	0.7	0.8

*pre: present p: past p.p: present perfect bold forms indicate the major tense in the moves

As can be seen from Table 3, in terms of verb tense, there are significant disciplinary differences ($p<0.05$). Excluding verbs in conditional sentences and those in meta-discourse expressions, we can draw the following conclusion:

E: In general, the present tense is the most frequently chosen tense. When abstract writers provide background information, they may also use the present perfect. If they are indicating the sequence of procedures in the actual research that is being reported, often the past tense is preferred.

F: With a few exceptions, when writing RA abstracts, financial researchers do not choose the past tense. The lack of the past tense may be that there is no specific experiment mentioned in this discipline.

S: 1. Verbs in Move1 are all in the present tense or the present perfect, depending on whether it is a problem presented or previous researches reviewed. 2. It seems that author's preference has much influence on the tenses of verbs in Move2, for there are as many past verbs as present ones. 3. In Move3 and Move4, where the specific experiments or the specific outcomes are reported, the past tense is likely to be the "correct" tense. 4. Since the generalizations of the results are presented in the last move, the verbs are in most cases in the present tense.

To sum up, our findings reveal that "an adequate theory of tense usage in EST discourse need to account not only for obligatory constraints on tense usage, but also for strategic choices that provide authors with the capability of manipulating temporal references of their own rhetorical purposes" (in Swales, 1990:153) .

3.2.4 First person pronouns

Recent years, there has been a change on attitude towards the use of first person pronouns in technical writing. What we got in the present corpus is listed below:

Table 4: First person pronoun

Personal pronouns	E	F	S
we	31	58	57
our	12	10	22
us	1	-	1
I	-	5	-
my	-	1	-
T	44	74	80
Wn	6,581	5,848	11,483
P per 10,000	66.859	126.539	69.668

*T: total number of first personal pronouns Wn: words number in each discipline

P per 10,000: first person pronouns per 10,000 words

Table 4 mainly conveys two points:

1. There are few first person singular pronouns.
Only in two abstracts can first person singular pronouns be found despite the fact that there are 23 single-authored abstracts. Eight single-handed authors chose to use plural pronouns instead of singular ones. In some cases, "we", "us", and "our" can shorten the distance between writers and readers to assume shared knowledge, goals and beliefs and thus stress solidarity. In other cases, these pronouns can be used to refer to the whole discipline or other members who participate in the research to reveal the collaborative nature of the research activity.

2. There are far more first person pronouns in abstracts in F than those in the other two disciplines.
A possible explanation may be that Finance falls into the so-called "soft" science while Engineering and Surgery are regarded as "hard" science.

Generally speaking, issues in the soft disciplines tend to be relatively diverse and range over a wide academic territory. Researchers in these fields have various ways to conduct their research and to make readers accept their results. Therefore, in a soft discipline such as Finance, "establishing an appropriately authorial persona and maintaining an effective degree of personal engagement with one's audience are valuable strategies" (Hyland, 2001: 216).

Knowledge in hard science, on the other hand, tends to be universalistic. Researchers are required to establish uniformities through research activities with precise measurement and systematic examination of a limited number of controlled variables. In hard disciplines such as Engineering or Surgery, researches are in most cases conducted to solve specific disciplinary issues. Consequently, writers in these fields tend to adopt a less personal style in order to strengthen the objectivity of their interpretations of the phenomena under study.

4. Conclusion

As was stated earlier, genre analysis has been carried out in the academic domain and also in ESP contexts. Few researches, however, have explored the RA abstract genre across disciplines. Still fewer studies have inquired into the distribution of the linguistic features in relation to the macrostructure of RA abstracts and subject matter. Because of the scant research carried out in literature, we hope that this study will contribute to a better understanding of the RA abstract genre and the disciplinary effect on this particular genre.

It is widely accepted that understanding the genres of written communication in one's field is essential to professional success (Berkenkotter & Huckin, 1995), so the result of the present study can be helpful for ESP teachers and students in abstract writing in these three disciplines. With a good understanding of the rhetorical structure of the RA abstracts and of the appropriate use of the linguistic devices in each move, students will compose RA abstracts that are more likely to be accepted by the discourse community.

Another important contribution of this study might be that it will contribute to future genre analysts and corpus-linguistic researchers by giving them a reference point and data with which to compare their own data.

One limitation of this study is that access was not available to the authors of these texts, which makes it impossible to capture the writers' thoughts when they are in the process of writing. Future studies can also include an analysis of the RA abstracts composed by authors with different cultural or national background to see the possible differences.

References:

Berkenkotter, C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognitive, culture, power.* Hillsdale, N.J.: L. Erlbaum Associates.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings.* London & NY: Longman.

Brett, P. (1994). A genre analysis of the Results section of Sociology Articles. *English for Specific Purposes*, 13 (1): 47-59.

Crookes, G. (1986). Towards a validated analysis of scientific text structure. *Applied linguistics*, 7 (1): 57-70.

Holmes, R. (1997). Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes*, 16 (4): 321-337.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes*, 20: 207-226.

Nwogu, K. N. (1991). Structure of Science Popularized Medical Texts. *English for Specific Purposes*, 10: 111-123.

Nwogu, K. N. (1997). The medical research papers: Structure and function. *English for Specific Purposes*, 16 (2): 119-138.

Posteguillo, S. (1999) The schematic structure of computer science research articles. *English for Specific Purposes*, 18 (2): 139-160.

Stein, W. (1997). *A genre Analysis of the TESOL Conference Abstract.* PhD Thesis, Oklahoma State University.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Appendix

This is a sample for the work we did on each abstract (The italicized words are the linguistic features counted and processed).

S2-6 – the sixth abstract in the second Surgery journal

(1) Laparoscopic cholecystectomy *has become* the standard operative procedure for cholelithiasis, but there *are* still some patients requiring conversion to open cholecystectomy mainly because of technical difficulty. (2) *Our* aim *was* to develop a risk score for prediction of conversion from laparoscopic to open cholecystectomy. (3) Preoperative clinical, laboratory, and radiologic parameters of 1,000 patients who *underwent* laparoscopic cholecystectomy *were analyzed* for their effect on conversion rates. (4) Six parameters (male sex, abdominal tenderness, previous upper abdominal operation, sonographically thickened gallbladder wall, age over 60 years, preoperative diagnosis of acute cholecystitis) *were found* to have significant effect in multivariate analysis. (5) A constant and coefficients for these variables *were calculated* and *formed* the risk score. (6) Overall 48 patients *required* conversion to open cholecystectomy (4.8%). (7) These patients *had* significantly higher scores (mean 6.9 versus -7.2, P <0.001). (8) Increasing scores *resulted* with significant increases in conversion rates and probabilities (P <0.001). (9) Ideal cut-off point for this score *was* -3; conversion rate *was* 1.6% under -3, but 11.4% over this value (P <0.001). (10) Conversion risk *can be predicted* easily by this score. (11) Patients having high risk *may be informed and scheduled* appropriately. (12) An experienced surgeon *has* to operate on these patients, and he or she *has* to make an early decision to convert in case of difficulty.

Table 1A sample

Move		B	A	M	F	C
Sentence(s)		1	2	3-5	6-9	10-12
Modal verbs	can					1
	may					1
Voice	A	2	1	2	5	2
	P			3		3
Tense	pre.	1				2
	p.		1	5	5	
	p.p	1				
Pron.	our		1			

*A: active voice P: passive voice pre.: present tense
p.: past tense p.p: present perfect Pron.: first person pronoun

Investigating Chinese English Learners' Use of Linking Adverbials: a Corpus-based Approach

Chensong Yan
PLA Institute of Foreign Languages

Abstract: A comparison of the use of linking adverbials in academic prose between EFL learners and native speakers of English produced findings that include: (1) they are similar in terms of order of proportions in which different semantic categories are used, that of result/inference being the largest while that of transition the smallest; (2) native speakers use appositional adverbials more commonly than EFL learners, whereas the latter use contrastive/concessive ones more often; (3) In marking result/inference, academic prose of native speakers show a clear preference for *thus, therefore* and *hence*, while the English learners prefer to use two items: *so* and *then*; and (4) They differ in the most common linking adverbs used.

Key words: linking adverbials; EFL learners; corpus; comparison

1. Introduction

1.1 Linking adverbials

Linking adverbials are one of the important devices for signaling the connections between clauses or beyond in textual communication, oral or written. The primary function of linking adverbials is to state the speaker or writer's perception of the relationship between two units of discourse and thus to help to build textual cohesion.

Cohesion and its associate coherence are recurring concepts in any discussion of textual communication. It is now generally accepted that coherence is the connectivity of underlying content whereas cohesion refers to connectivity on the surface. Logically speaking, coherence precedes cohesion. A text must have coherence for it to be comprehensible, but it may not necessarily, yet most probably involve cohesion, which is realized by cohesive ties like linking adverbials. (Halliday & Hasan, 1976; de Beaugrande & Dressler, 1981; Schiffrin, 1987)

The distribution of linking adverbials differs across registers of conversation, news, fiction, academic prose, etc. A speaker usually does not use as many linking adverbials or as often as a writer does. Formal writings like academic prose seem to depend more on explicit signals to indicate connections than casual speeches or conversations. It is necessary for a writer of academic prose to signify the relationships overtly and succinctly between his ideas, for, unlike a speaker, he is not likely around to clarify what he means when it is ambiguous or unclear. Linking adverbials are therefore an essential part of academic prose.

According to *Longman Grammar of Spoken and Written English* (Biber *et al* 1999, henceforth abbreviated as LGSWE), linking adverbials in English are realized by single adverbs, adverb phrases, prepositional phrases, finite clauses and nonfinite clauses, and they fall into six semantic categories: enumeration (*first, similarly, for one thing*), summation (*in sum, to conclude, all in all*), apposition (*i.e., for example, that is*), result/inference (*thus, hence, as a result*), contrast/concession (*however, though, on the other hand*), and transition (*incidentally, by the way, by the by*).

LGSWE is presumably the first single comprehensive grammar of English drawing extensively on research findings derived from a large-scale corpus. The corpus they used is the Longman Spoken and Written English Corpus (henceforth abbreviated as LSWEC). With figures, tables and authentic examples, the work presents a clear and true picture of how contemporary English is used.

1.2 Learning linking adverbials for EFL learners

It takes time and effort for learners of English as a foreign language to acquire linking adverbials. We assume that the ability to use linking adverbials is one of the indicators of an English learner's proficiency in the language, and that, for various reasons, there will be systematic differences in the use of linking adverbials between learners and native speakers, with regard to semantic categories as well as syntactic forms.

1.3 Approach of this study

This study is intended to examine the use of linking adverbials by EFL learners at their advanced stage. It is cross-sectional in that it does not address the question of order in which the English linking adverbials have been acquired, but investigates the ways they are used by the learners and makes comparisons of use between the learners and native speakers of English, using a corpus-based approach. The research questions addressed in this study are as follows:

(1) How do the English majors use linking adverbials?

(2) How does the English majors' use of linking adverbials compare to that of native speakers?

(3) Why are some of the adverbials under- or overused by the English learners compared with native speakers?

(4) What account for the variation of use among individual learners?

The corpus used comprises 113 theses written as required for the BA degree by the English graduating majors of the class of 2000 at the PLA Institute of Foreign Languages at Luoyang, China. These papers belong to the register of academic prose, covering various topics mainly of four fields: language study, reviews of literary works, translation and American studies. The total number of running words (tokens) of the theses is approximately 417600, with an average length of 3695 words each. The total number of different words (types) used is 21034, showing that the papers cover a considerable range of subject areas.

Using *Wordsmith*, a popular concordancing software, we concordanced 77 linking adverbials that are described or mentioned in LGSWE. As some of the adverbs (e.g. *so, then, yet, rather*) have usages other than linking adverbs, their concordances were edited to delete the irrelevant items. The numbers of their occurrences were entered into a table.

As LSWEC contains 40 million running words and is constructed to provide the basis for systematic analyses of grammatical patterns for LGSWE, we assume that the findings derived from the corpus and described in LGSWE could be considered as representative of the general competence of the native speakers of English. Thus we compared our data with those of LGSWE and made comparative analyses. What we examined include: distribution across semantic categories in terms of frequency normalized to number per million words, and distribution across syntactic realizations in terms of percentage per million words; the most common linking adverbials (listed in descending order) used. The results were summarized in tables and graphs. The analyses were conducted with SPSS 10.0.

2. Results

2.1 Tables and graphs

Table 1 summarizes the frequencies of linking adverbials in both BA papers and LSWEC, classified into the five semantic categories. The frequencies were normalized to numbers per million and the ratios were obtained with LSWEC as the base.

Table 2 compares the 21 most common linking adverbials used in LSWEC and BA papers.

Table 3 compares the distributions of linking adverbials across semantic categories between the two corpora. An observed chi-square test value is recorded.

Graph 1 is a graphic representation of Table 3.

Graph 2 compares the distributions across syntactic forms.

Table 1: Frequencies per million of linking adverbials

Linking adv.	Semantic categ.	LSWEC.	BA paper.	Ratio.
third	enumeration		59.87	
first	enumeration	100.00	138.89	1.39
in addition	enumeration	100.00	95.79	.96
finally	enumeration	100.00	81.42	.81
furthermore	enumeration	100.00	28.74	.29
second	enumeration		112.55	
moreover	enumeration		74.23	
secondly	enumeration		50.29	
firstly	enumeration		45.50	
also	enumeration		33.52	
similarly	enumeration		31.13	
thirdly	enumeration		31.13	
first of all	enumeration		16.76	
For another (thing)	enumeration		14.37	
in the first place	enumeration		11.97	
further	enumeration		9.58	
fourthly	enumeration		7.18	
likewise	enumeration		7.18	
next	enumeration		7.18	
by the same token	enumeration		4.79	
fifthly	enumeration		4.79	
lastly	enumeration		4.79	
to begin with	enumeration		4.79	
added to that	enumeration			
first and foremost	enumeration			
for one (thing)	enumeration			
in the second place	enumeration			
to sum up	summation		31.13	
all in all	summation		11.97	
in sum	summation		11.97	
in brief	summation		4.79	
in conclusion	summation			
overall	summation			
to conclude	summation			
to summarize	summation			
for example	apposition	600.00	680.08	1.13
e.g.	apposition	200.00	409.48	2.05
i.e.	apposition	200.00	102.97	.51
for instance	apposition	100.00	153.26	1.53

that is	apposition	100.00	86.21	.86
in other words	apposition		47.89	
specifically	apposition		21.55	
namely	apposition		19.16	
which is to say	apposition		2.39	
more precisely	apposition			
thus	result/inference	700.00	316.09	.45
therefore	result/inference	600.00	368.77	.61
then	result/inference	400.00	730.36	1.83
so	result/inference	200.00	878.83	4.39
hence	result/inference	100.00	67.05	.67
as a result	result/inference		129.31	
consequently	result/inference		47.89	
in consequence	result/inference			
however	contrast/concession	1100.00	589.08	.54
yet	contrast/concession	100.00	239.46	2.39
on the other hand	contrast/concession	100.00	170.02	1.70
nevertheless	contrast/concession	100.00	74.23	.74
rather	contrast/concession	100.00	47.89	.48
though	contrast/concession	50.00	433.43	8.67
anyway	contrast/concession	50.00	11.97	.24
besides	contrast/concession		143.68	
on the contrary	contrast/concession		59.87	
in contrast	contrast/concession		47.89	
instead	contrast/concession		40.71	
after all	contrast/concession		21.55	
still	contrast/concession		14.37	
even so	contrast/concession		4.79	
alternatively	contrast/concession		2.39	
by comparison	contrast/concession		2.39	
conversely	contrast/concession		2.39	
anyhow	contrast/concession			
at any rate	contrast/concession			
in any case	contrast/concession			
in spite of that	contrast/concession			
by the way	transition		4.79	
incidentally	transition		4.79	
by the by	transition			

Table 2: 21 most common linking adverbials compared between LSWEC and BA papers

	LSWEC	Occurrences per million words	BA papers	Occurrences per million words
1	however	1100.00	so	878.83
2	thus	700.00	then	730.36
3	for example	600.00	for example	680.08
4	therefore	600.00	however	589.08
5	then	400.00	though	433.43
6	so	200.00	e.g.	409.48
7	e.g.	200.00	therefore	368.77
8	i.e.	200.00	thus	316.09
9	yet	100.00	yet	239.46
10	on the other hand	100.00	on the other hand	170.02
11	for instance	100.00	for instance	153.26
12	first	100.00	besides	143.68
13	in addition	100.00	first	138.89
14	that is	100.00	as a result	129.31
15	finally	100.00	second	112.55
16	nevertheless	100.00	i.e.	102.97
17	hence	100.00	in addition	95.79

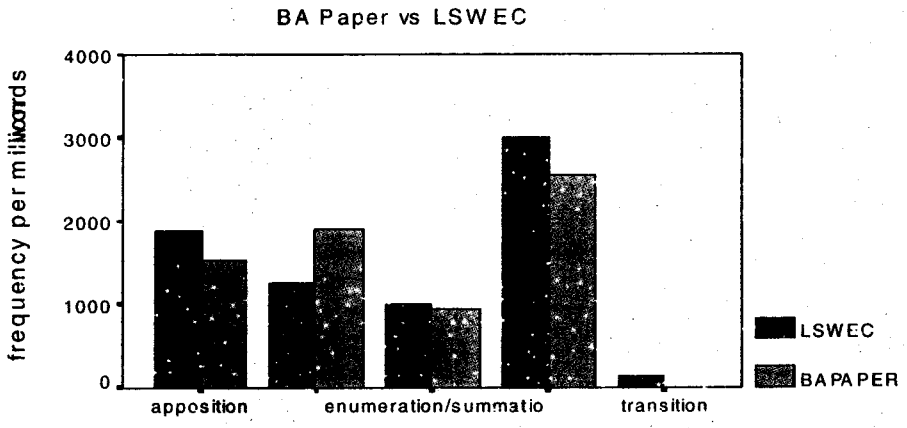
18	rather	100.00	that is	86.21
19	furthermore	100.00	finally	81.42
20	though	50.00	nevertheless	74.23
21	anyway	50.00	moreover	74.23

Table 3: Distributions across semantic categories and chi-square test

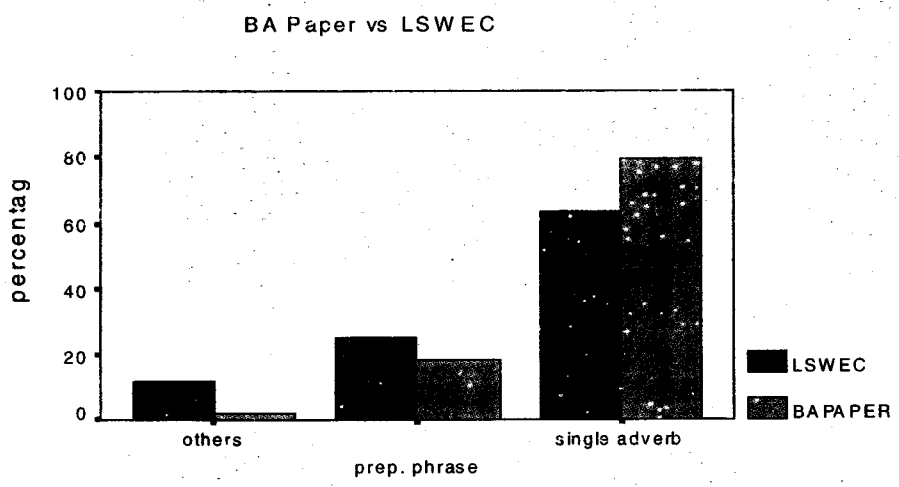
Semantic category	BA papers (per million)	LSWEC (per million)
enumeration/summation	936.3	1000
apposition	1522.99	1875
contrast/concession	1906.13	1250
result/inference	2538.31	3000
transition	9.58	125
Total	6913.31	7250

$X^2 = 304.61 > 18.47$ at .001 level

Graph 1: Distribution Across Semantic Categories



Graph 2: Distribution Across Syntactic Forms



2.1 Findings

Our findings include:

(1) LSWEC and BA papers are similar in having the largest proportion in the semantic category of result/inference, the smallest proportion in that of transition and the next smallest in that of enumeration. The two corpora are almost identical at the level of frequency in the category of enumeration. (Graph 1)

(2) However, the chi-square test conducted indicates that the two corpora are different in the distribution of linking adverbials across the five semantic categories. The highly significant X^2 value observed (304.61) gives us almost a hundred percent confidence in their difference. (Table 3)

(3) LSWEC uses appositional adverbials more commonly than BA papers, whereas the latter use contrastive/concessive ones more often. Marked differences in frequency are also observed in the category of result/ inference and that of transition. (Graph 1)

(4) If we examine Graph 1, Tables 1 and 2 together, we find that the higher proportion of contrastive/concessive linking adverbials in BA papers is most probably due to a high frequency of one item: ***though***, which occurs eight times as often as its counterpart in LSWEC. Many of the other items in this category like *however, nevertheless, rather, anyway*, have lower frequencies in BA papers than in LSWEC.

(5) In marking result/inference, academic prose of native speakers shows a clear preference for *thus, therefore* and *hence*. Together they occur approximately 1400 times per million words in LSWEC, compared with about only 750 (almost half as many) in BA papers. The English learners, in contrast, prefer to use two items: *so* and ***then***, which in BA papers occur respectively 4.39 times and 1.83 times as often as in LSWEC. (Table 1)

(6) In both BA papers and LSWEC, the majority of linking adverbials are realized by single adverbs. The English learners, however, seem to rely more on them, but less on forms like adverb phrases, finite or non-finite clauses (all classified as "Others" in this study), e.g. *first and foremost, to conclude, that is to say*, than native speakers do. (Graph 2)

(7) Though both are of academic prose, BA papers differ from LSWEC in the most common linking adverbs used. The two lists in Table 2 share most of the items but differ in four, thus involving eight items: *hence, rather, furthermore, anyway* on the part of LSWEC and *besides, as a result, second, moreover* on the part of BA papers. Among these, the items of *besides* and *as a result* show frequencies of about 143 and 129 per million respectively. By comparison, they occur less than 50 times in LSWEC and hence fail to make the list. (Table 2)

(8) Table 2 also shows that the shared items do not match in the order of frequencies in which they are used. The most notable differences are exhibited between LSWEC and BA papers with *though* (50.00 vs. 433.43), *so* (200.00 vs. 878.83), *yet* (100.00 vs. 239.46), *however* (1100.00 vs. 589.08), *thus* (700.00 vs. 316.09), *therefore* (600.00 vs. 368.77). (Table 2)

(9) The most common linking adverbial in LSWEC is *however*, which marks contrast/ concession, whereas in BA papers it is *so*, which marks result/inference. (Table 1 and Table 2).

3. Discussion of findings

According to LGSWE, linking adverbials like *so, though* and *then* are most commonly used in conversation rather than in academic prose, whereas *thus, therefore* and *hence* most often occur in academic prose. As we noted above, the English learners tend to use *so* and *then*, words characteristic of conversation, to mark

result or inference, where they should use more formal ones like *thus, therefore* and *hence*. Likewise, they use *though* instead of *nevertheless*. The learners' preferences for informal forms may be because: (1) some learners are unaware of register differences as regards the use of linking adverbials, and they have not acquired the conventions for writing academic papers; or (2) some learners have not yet learnt to use words like *thus, therefore* and *hence*, words which are more difficult to use and are only mastered at a later stage.

Evidence from the corpus of BA papers indicates that the English learners vary in their ability of using some of the "more complex" adverbials. In BA papers, we do see *thus, therefore* and *hence* used. Listed in the following are some concordances of *hence*. We obtained altogether 28 examples of *hence*, which are distributed in only 16 of a total of 113 papers.

1 Chinese are rather certain and strict. Hence, the translation of English long
2 reversing Chinese expression practice. Hence, rearranging the components is nec
3 ic acceptance of the nation's policies. Hence, what are often viewed as Congress
4 They are prey to disease. Their future hence is menaced. In which school do you
5 ld not save its inevitable degradation. Hence it is degraded to mean a boor, whi
6 " in Latin stood for a farm or a house. Hence it entered old French as "vilein"

The learners' problems of using linking adverbials are not confined to stylistic inappropriateness or vocabulary inadequacy only. More problems could be discovered only through a close examination of individual cases. The description and discussion above have only dealt with the general tendencies. We have noted the problem of over- or underuse, but should not ignore the problem of misuse.

Looking at the concordances we obtained, we found many cases of misuse, where the students failed either to distinguish between the subtle nuances of meaning of some of the linking adverbials, or figure out the correct logical relationships between units of discourse.

In the following are two examples of misuse. *Nevertheless* in the first should be replaced by *however,* and *Thus* in the second should be replaced by *for example.*

Actually this is the very feature of an oxymoron. By combining contradictories, writers produce a startling effect And if their oxymora are fresh and apt, they may win for themselves a reputation for wit. Nevertheless, Oxymoron is not alone to convey self-contradicting information. There is another kind of figure of speech: paradox.

Living in the historical period of the transition from feudalism to capitalism, Shakespeare faithfully and vividly reflects, through a host of full-blooded characters in his plays, the major social contradictions of his time. Thus his Romeo and Juliet faithfully depicts the decaying old feudal families and the catastrophe that the generations of feud has brought about.

4. Conclusions

(1) This study has shown that, in terms of general tendencies, the advanced Chinese English learners' use of linking adverbials is roughly in accordance with native speakers in terms of proportions of semantic categories. But distributions of frequencies across semantic categories differ considerably.

(2) Many learners' use of linking adverbials shows stylistic inappropriateness. They tend to overuse those characteristic of conversation whereas underuse those that normally occur in academic writings.

(3) The learners' ability varies in their use of linking adverbials. The inappropriateness and misuse indicate that they are still lacking in vocabulary power and general proficiency in English, being unable to tell nuances of meaning and correct logical relationships between units of discourse.

The limitations of this study include: (1) The corpus of BA papers is not big enough, and confined to only the graduating students of one college. (2) Some of the BA papers, we suspect, may have quoted from articles by native speakers without giving credit. This, of course, could have contaminated the result. (3) The use of some of the adverbials like *though* could be over-rated. It is sometimes hard to distinguish between its usages. These inadequacies will limit validity as well as the generalizability of this study.

References

Beaugrande, R. de, & Dressler, W. (1981) Introduction to text linguistics. New York: Longman.

Biber, D., Conrad, S., & Reppen, R. (1998) Corpus Linguistics. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999) Longman Grammar of Spoken and Written English. Pearson Education Limited.

Halliday, M.A.K., & Hasan, R. (1976) Cohesion in English. London: Longman.

Schiffrin, D. (1987) Discourse markers. Cambridge: Cambridge University Press.

Sinclair, J. (1991) Corpus, Concordance, Collocation. New York: Oxford University Press.

Noticing, Learning and Acquiring the Central Uses of Common English Words

Pu Jianzhong
PLA University of Foreign Languages

Abstract: This paper intends to discuss the problem of mastering the central uses of common English words. It puts forward two new methods of reading which might prove to be effective in learning and acquiring the central uses of common words. These methods of reading are largely based on such notions as patterns (or colligations), collocations, lexical chunks, which have gained currency in the field of applied corpus linguistics. Two interdependent reading methods of prime importance in recognizing, observing and learning the central uses of common words are discussed in the present paper: 1) reading normal texts horizontally and retrospectively for potentially useful patterns, collocations, or chunks of language; 2) reading "abnormal" texts (i.e. KWIC concordance lines) both horizontally and vertically for central patterns, collocations or chucks of language.

Both ways of reading and their interdependence are illustrated in the paper. As for the first one, the critical part is how to decide on the potentially useful patterns, collocations, or chunks of language concerning the uses of common words. One method to be recommended is Hunston and Francis' (2000) way of identifying patterns, which can be reinforced by further considering collocations or chunks of language in some cases as well. In this normal way of reading, the reader or learner has to make full and best use of his/her previous knowledge about the uses of the common words.

With corpora and concordancers widely available, the second way of reading is now feasible. The KWIC concordance lines can be both read horizontally and read vertically. The most importance function of such reading is that it can provide the "observers" with means of observing the central or typical uses of words. More experiences in using concordancers will facilitate this way of reading and facilitate the learning as well.

Key words: central uses, patterns, collocations, chunks, concordance

1. Introduction:

Previous research on Chinese learners' interlanguage shows that their mastery of the central uses of common English words is still far from adequate. This points to the inadequacy of both ESL (or EFL) learning and teaching. Although it has been pointed out as early as the 1980s that the main focus of learning English should be on: a) the *commonest* word forms in the language; b) their *central* patterns of usages; c) the combinations which they *typically* form (Sinclair & Renouf, 1988: 148), ESL teaching has not yet solved the problem of helping students learn and acquire the central uses of common words. One important reason for the above phenomenon is that effective methods of facilitating such learning and acquisition is still not widely available to the teachers, let alone to the learners.

This paper intends to discuss the above problem with the purpose of putting forward a new method of reading which might prove to be effective in learning and acquiring the central use of common words. This method of reading is largely based on such notions as patterns (or colligations), collocations, lexical chunks, which have gained currency in the field of applied corpus linguistics. Two interdependent reading methods

of prime importance in recognizing, observing and learning the central uses of common words are discussed in the present paper: 1) reading normal texts horizontally and "retrospectively" for potentially useful patterns, collocations, or chunks of language; 2) reading "abnormal" texts (i.e. KWIC concordance lines) both horizontally and vertically for central patterns, collocations or chucks of language. In addition, the feasibility of combining these two reading methods is explored.

2. Critical notions

In order to demonstrate and promote this new reading method, some critical notions, which are not familiar to all, have to be first introduced. These notions include patterns (colligations), collocations, lexical chunks(henceforth 'chunks'). They have gained currency in the field of applied corpus linguistics for some time, although they are not first developed by corpus linguists. One important reason why they are introduced here is that they are critical to the use of a word.

The notion of "pattern" is perhaps first used by Hornby in his *A Guide to Patterns and Usage in English* (see Hunston & Francis, 2000: 3) to denote the syntactic features of verbs and words of other classes, or rather their usage. Hornby advocates that learners be told "which words enter into which pattern" (ibid: 5). In his book, Hornby described 25 verb patterns, 4 noun patterns and 3 adjective patterns, as well as usage of words of other classes.

This notion of "pattern" is further developed by Hunston & Francis (1998, 2000). They use this notion to denote superficial syntactic behavior of words, particularly verbs, nouns, and adjectives. The definition they give is "the patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning." (ibid: 37) The patterns of words, in their view, are critical in language learning. With such a belief, they compiled two dictionaries elaborating on the use of common words of these three classes. They are now known as pattern grammar: *Grammar Patterns 1: Verbs, Grammar Patterns 2: Nouns and Adjectives.* The important contributions of these works include: 1) they encode the syntactic behavior of word in a simple and transparent way; 2) they elaborate on the typical uses of all common words; 3) they bring pattern, structure and meaning together.

What Hunston & Francis mean by "pattern" is quite similar to Firth's notion of "colligation". By colligation, Firth meant to refer to "the syntactical characteristics of the text" (Firth, 1957: 95). To be exact, it refers to "the inter-relation of grammatical categories in syntactical structure" (Firth, 1957: 99). Compared with Hunston & Francis's notion of pattern, colligation is a more general term. In some cases, however, what some researchers call colliagation is in fact what Hunston and Francis call pattern. In this paper, we shall only use the notion "pattern" for it is more clearly defined than colligation.

The notion of collocation is also first discussed by Firth to refer to "actual words in habitual company" (Firth, 1957: 99). However there is still no widely agreed definition for collocation. Different scholars seem to have slightly different ideas on what is to be accounted for by collocation and accordingly their definitions for it differ from each other, notwithstanding the fact they have much in common (see Kjellmer, 1991, Krishnamurthy, 1987, Sinclair 1991, etc.). In recent years, the study of collocation also attracted the attention of many Chinese scholars. For example, Li (1999) and Wei (1999), in their doctoral dissertations, discussed and studied collocation. Wei (1999) also managed to work out a deliberate definition of collocation. In this paper, collocation is understood to be *the recurrent co-occurrence of two or more words (particularly, nouns, verbs, adjectives and adverbs) in a certain grammatical pattern* (see Pu 2001).

Besides, another important notion – chunk, has to be introduced here. Different from a pattern or collocation, which often has an individual word as a focus, a chunk can either have or have not a focus. Put it another way, when one talks about a pattern or collocation, it is always the pattern of a certain word or a class of words, or the collocation between one word and other words. This, however, is not necessarily the case with a chunk. For example, expressions like *by and large, as it were, whether or not, boys and girls, day and night, such...as, so as to*, etc. (some of them are traditionally known as phrases) do not have obvious focuses. It is hard to say which word is the focus and which word is subsidiary. In these expressions, it is the whole word combinations but not individual words that is really important in conveying meaning. To give a comprehensive account of these expressions as well as patterns and collocations, we see fit to introduce the notion of chunk. The notion of chunk can be applied to all kinds of word combinations or the patterns they form. In Pu (2001), chunks is defined as "a frequently used, prefabricated multi-word unit which has an identifiable structure, a relatively determinate meaning and allows different degrees of abstraction". This notion of chunk will include both patterns and collocations, and other forms of word combinations. In addition, this notion may facilitate the teaching and learning task for the simple reason that fewer notions are involved and is easily applicable.

3. Corpora and concordancer

In this study, corpora of small sizes are used. They include: 1) a corpus of texts in a college English teaching material, to be exact, NHCE (New Horizon College English); 2) the Frown corpus (a new version of the Brown corpus); 3) CLEC (Chinese Learner English Corpus). The first two corpora consist of texts written by native speakers of English; the third one is a learner corpus used here to represent Chinese learners' interlanguage.

The concordancer used in the present study is Scott's (1998) Wordsmith (Version 3.00). The important functions of the software tool include: 1) concordance; 2) word list; 3) key word. To most learners of English, the most useful function is to make concordances.

The above corpora can serve various purposes and can be used in a flexible way in accordance with researchers' needs.

4. Reading method 1

As noted above, the first reading method is concerned with reading normal texts horizontally and "retrospectively" for potentially useful patterns, collocations, or chunks. This way of reading does not need a corpus or a concordancer, and it appears quite familiar to the learners. What is critical in this way of reading lies in the fact that the reader has to decide for himself on the potentially useful patterns, collocations, or chunks concerning the uses of common words. One method to be recommended here is Hunston and Francis' (2000) way of identifying patterns. They not only demonstrate the usefulness of mastering patterns of words, but also show how they are identified. According to their model, patterns of verbs, nouns and adjectives are especially useful to the learners. Besides the patterns, what is sometimes also important is to the collocations between words and the chunks.

In this normal way of reading, the reader or learner has to make full and best use of his/her previous knowledge about the uses of the common words now in focus. As far as the Chinese learners of English are concerned, this knowledge will inevitably include the knowledge about the uses of the so-called Chinese equivalents of the English words. It will prove to be useful in deciding what are the potentially useful uses. It is important to note that since different students have different knowledge about the uses of words, their decisions on the potentially useful patterns, collocations, or chunks are different.

An example may make the above points clearer. Suppose the reader is a second year Chinese college student learning English as a foreign language. Suppose the teaching material he/she is using is *New Horizon College English* and is now learning text A of unit three in book three. The text is as follows (the title of the text is *Where Principles Come First* and only the beginning two paragraphs are presented here):

> The Hyde School operates on the principle that if you teach students the merit of such values as truth, courage, integrity, leadership, curiosity and concern, then academic achievement naturally follows. Hyde School founder Joseph Gauld claims success with the program at the $18,000-a-year high school in Bath, Maine, which has received considerable publicity for its work with troubled youngsters.
>
> "We don't see ourselves as a school for a type of kid," says Malcolm Gauld, Joseph's son, who graduated from Hyde and is now headmaster. "We see ourselves as preparing kids for a way of life - by cultivating a comprehensive set of principles that can affect all kids."
>
> ...

If the reader follows Hunston and Francis' (2000) model, he/she might notice the uses of some common verbs, nouns, and adjectives with respect to their patterns. They may include the following[1]:

V: *operate* (n V), *teach* (V n n), *follow* (n V), *claim* (V n), *receive* (V n), *see* (V n *as* n), *say* (V with quote), *graduate* (V *from* n), *prepare* (V n *for* n), *cultivate* (V n), *affect* (V n)[2]

N: *principle* (N that), *merit* (N *of* n), *work* (N *with* n), *type* (N *of* n), *way* (N *of* n), *set* (N *of* n)

ADJ: *academic* (ADJ n), *high* (ADJ n), *troubled* (ADJ n), *considerable* (ADJ n), *comprehensive* (ADJ n)[3]

As to which patterns are worth noticing and learning, it depends on the previous experiences of the individual language learners. For those very familiar patterns, only a little attention is needed. But for those less familiar yet useful ones, due attention is still required. One effective way of judging whether a certain pattern is worth paying attention to is to decide whether this pattern is often used and whether the Chinese equivalent one tends to establish for it is often used in different pattern(s) in Chinese to convey similar meaning(s). There is evidence to show that such patterns are more difficult to learn and remember. For example, among the verbs listed above, it is reasonable to predict that most learners will have more difficulty in acquiring the pattern "n V" for *operate* and *follow*, "V n *as* n" for *see*, "V n *for* n" for *prepare*.

As for the simpler and commoner patterns such as "V n" for *claim*, *receive*, *cultivate*, and *affect*, "ADJ n" for *academic*, *high*, *troubled*, *considerable* and *comprehensive*, it is often rewarding to draw attention to the collocations. In these cases, the collocations *claim success*, *receive publicity*, *cultivate principle*, *affect kids*, *academic achievement*, *high school*, *troubled youngsters*, *considerable publicity*, *comprehensive set* are all potentially useful and thus deserving being noticed and learned. This of course does mean that other collocations can be neglected. For instance, collocations such as *school operates*, *way of life*, etc. are

---

[1] Major codes used to encode patterns include: v (verb group), n (noun group), adj (adjective group), adv (adverb group), that (clause introduced by that, realized or not), -ing (clause introduced by an '-ing' form), to-inf (clause introduced by a to-infinitive form), wh- (clause introduced by a wh-word, including *how*), with quote (used with direct speech). Where a preposition, adverb, or other lexical item is part of a pattern, it is given in italics to indicate that it is a lexical item rather than a code. (See also Hunston & Francis 2000: 45)

[2] Note that auxiliary verbs and modal auxiliaries are not included. Note also that the word-class in focus is in upper-case. Besides, the pattern "V" is marked as "n V" for the simple reason that the verb used in this pattern often has restrictions on the lexical realization of the noun group.

[3] Note that patterns which may apply to almost all members of a word class are not worth special attention, and therefore they are not listed. For example, *school*, *students*, *values*, etc. are not included.

certainly as useful as above ones.

Apart from the above patterns and collocations, there are other uses of words which might attract the reader's attention. For example, expressions like *if...then...*, *such...as*, *a...of*,[1] etc. are certainly worth learning. In order to take them into account, we see fit to introduce the notion chunk. If chunk is defined as above, it is appropriate to incorporate patterns, collocations and other ways of word combinations all into the notion of chunk. One important advantage of the notion of chunk is that it regards a word and its immediately relevant environment as a unit in its own right. For a chunk, it is no longer important whether there is a focus or not; it is the whole that convey the intended meaning(s). Besides, the notion of chunk captures a level of abstraction between collocation and pattern. For instance, in between *affect kids* and "V n" (*affect*), there is a level of abstraction which may be represented by such chunks as *affect someone/people*.

One inevitable difficulty in horizontal and "retrospective" (in the sense that the reader has to think of his past experiences with the uses of words) reading, however, is that it is often hard to decide which are common words, which patterns, collocations or chunks are common enough to deserve attention. Such being the case, other reading methods or facilities are highly desirable if these problems are to be resolved.

5. Reading method 2

Now that we have corpora and concordancers, texts can be read both horizontally and vertically. In order to do this, the uses of individual words are shown in KWIC concordances. This format of representing the texts lend them well to both horizontal and vertical reading, and thus to the observation of typical uses of words.

For the sake of space, only one verb, i.e. *operate*, is examined to illustrate this method of reading. If the reader is interested to know the typical uses of this verb, he/she may consult NHCE, which he/she has been learning text by text. And if there is not enough instances, he can consult other native speaker corpora, in this case, Frown. By use of concordancer, he/she will get the following KWIC concordances[2]:

1 ing government is a scalpel, are hot to operate B20 3 on the body politic
2 n A37 172 American line could buy and operate in local competition with A37
3 44 143 and unconscious levels, and that operate binomially, amassing and G44 1
4 ircraft: It had to be able to E16 45 operate from an exceedingly confined spa
5 9 124 But the United States doesn't operate in a vacuum. European E09 125
6 mmunity, some AIDS educational programs operate out of a beauty shop. The ow
7 D13 134 Neopagan Witches also operate with an ethic that forbids them
8 e 500 E28 219 members. Fifty-one operate in Canada; 49 currently do busin
9 the F27 24 commodity programs really operate. F27 25 Imagery F27
10 E18 17 how an unfettered press might operate in wartime. The experience of
11 sing relay ladder logic and designed to operate E32 101 sequentially to emulat
12 smoke-belching monsters of yore and can operate within B16 71 stringent envir
13 inavian group of companies to build and operate what it says will be one of the
14 lize how many of these models F21 143 operate, think about a grid of points ov
15 brains of boys with mathematical talent operate in a way that is physically uniq
16 Tenderloin, the Belvidere continued to operate for years, a protean F19 175 u
17 Acknowledge that you can no longer operate in A42 65 old ways in a new e
18 ay not be on their side. "It's hard to operate a business on someone else's dis
19 arp scalpel, and who can hardly wait to operate B20 25 on the body politic. O
20 he whole movement, largely because they operate D13 60 according to a fairly [3]

[1] These expressions are not usually described as collocations. Besides, they do not have obvious focuses.
[2] For the sake of convenience, only the base form of *operate* is examined. Altogether 43 instances are found in NHCE and Frown, but only 20 of them are retained, keeping all the 4 instances from NHCE, and the rest randomly chosen from Frown.
[3] In the above lines, those numbers beginning with an English letter indicate the source of the lines, therefore, they are not relevant to the uses of words.

To read horizontally, one can read the concordance line by line. For instance, one finds that in line 1, the verb *operate* is followed by the preposition *on* and then a noun group, that is, it is used in the pattern "V *on* n". In line 2, *operate* is followed by a preposition phrase but it does not seem to be closely related to the use of this word with regard to meaning, and therefore it is used in the pattern "n V". But, to read vertically, it is only partially obvious that *operate* is often preceded by "to", indicating the verb use of *operate*, and that it is often followed by prepositions and adverbs, indicating its intransitive use. As for other characteristics of the uses of this verb, they are not very obvious and it is difficult to find the typical uses.

The reason why typical uses of this word are not immediately observable to the readers is partly due to the fact that these lines are not properly sorted. If they are first sorted according to the first right-hand word and then the first left-hand word, the concordance lines will look as follows:

1 ay not be on their side. "It's hard to operate a business on someone else's dis
2 ing government is a scalpel, are hot to operate B20 3 on the body politic
3 arp scalpel, and who can hardly wait to operate B20 25 on the body politic. O
4 44 143 and unconscious levels, and that operate binomially, amassing and G44 1
5 he whole movement, largely because they operate D13 60 according to a fairly
6 sing relay ladder logic and designed to operate E32 101 sequentially to emulat
7 the F27 24 commodity programs really operate. F27 25 Imagery F27
8 Tenderloin, the Belvidere continued to operate for years, a protean F19 175 u
9 ircraft: It had to be able to E16 45 operate from an exceedingly confined spa
10 n A37 172 American line could buy and operate in local competition with A37
11 9 124 But the United States doesn't operate in a vacuum. European E09 125
12 Acknowledge that you can no longer operate in A42 65 old ways in a new e
13 E18 17 how an unfettered press might operate in wartime. The experience of
14 e 500 E28 219 members. Fifty-one operate in Canada; 49 currently do busin
15 brains of boys with mathematical talent operate in a way that is physically uniq
16 mmunity, some AIDS educational programs operate out of a beauty shop. The ow
17 lize how many of these models F21 143 operate, think about a grid of points ov
18 inavian group of companies to build and operate what it says will be one of the
19 D13 134 Neopagan Witches also operate with an ethic that forbids them
20 smoke-belching monsters of yore and can operate within B16 71 stringent envir

This time, one can more conveniently observe that *operate* is used intransitively in most cases, for it is often followed by prepositions, particularly *in*. If one examines the lines more closely, one will find that the verb is used as transitive only in lines 1 and 19[1], forming such collocations as *operate...business*, *operate ... systems*. As to whether such collocations are typical or not, these concordance lines are not enough; big corpus may be needed.

However, in some case, due to the intervening line indicators and other reasons, it still appears difficult to find the typical uses of the word in terms of patterns. In such cases, a further manipulation of the lines is needed. Another useful sorting function provided in the software tool Wordsmith is to sort according to "set". This "set" has to be marked manually first. In our case, we mark the "set" according to patterns: two different patterns of *operate* are marked, i.e. "n V", "V n". After marking them, the lines can be resorted first according to "set" then by the first right-hand word (for reasons of space, the concordance lines are not given here).

Suppose the reader wants to know how the verb is often used by Chinese learners of English, one effective method is to resort to a relevant learner corpus. In this case, one can resort to CLEC. Their uses are represented in the following lines (the search results have already been resorted and manipulated in the way described above):

[1] In the concordancer, the lines can be expanded and thus reveal the use of *operate* in line 19 as "operate ... systems".

1 9-5) No longer was it possible to operate on the assumption that most wome
2 5,-1] knowleage [fm1,-], but also often operate. Only through constant operating
3 teachers. Parents and school should operate." Then my father made a
4 than man's [wd3, 1-] brains. It can operate very correct [wd2,1-] . It can a
5 e. We have to study it and learn how to operate [vp1,1-] , or we can not make
6 ir studies. The computers are easy.to operate, you [pr1,4-] can be easy to in
7 e are asked to know more about how to operate and repair a machine than about
8 eapons and strategies. They are able to operate conventional facilities as well
9 ts on with [vp2,2-] naturally. We can't operate it with our own mind, but only i
10 e in [pp1,-2] study computer. Now I can operate it freely and make programs. Com
11 own content [wd3,s-]. [sn2,s] We should operate it according to its steps. If on
12 nces about it [pr1,s-] and then you can operate machines skillfully. More practi
13 d learnt in school. Now I am able to operate [np7,2-1] computer and type a
14 my cousin to [vp4,6-3] learning how to operate the computer. [wd4,0-2] First d
15 factory only needs some engineers to operate the computers which control the
16 price of different pastry and how to operate the Dolden [fm1,-] Dragon Car
17 ther samples. For instance, learning to operate the computer is very difficult f
18 d easily. Some people don't know how to operate their computer under the DOS. Bu
19 along with personnels [np5,-] that can operate them skillfully is more likely t
20 for a stranger. However, as long as he operate [vp3,1-] it some day, he can gai
21 merely do the things rapidly. The quick operate [wd2,s-] is often careless and s

Compared with native speakers' uses of the verb, one finds that for the learners the transitive use (from line 7 to line 20) but not the intransitive use (from line 1 to line 6) is predominant. Besides, the collocation *operate...computer* is quite frequently used by the learners, but this is not the case with the native speakers. Such findings are quite useful for language teachers to predict how his/her students tend to use a certain word.

What is of critical importance in this way of reading is to make full use of the concordancers as well as of the learners' curiosity about the uses of words. First of all, in order to observe some typical uses of a word, the sorting function provided by the concordancers is often useful. Secondly, marking the patterns is sometimes necessary when sorting does not reveal typical uses. Thirdly, byproducts of searching is often significant, for it may provoke other kinds of searching.

6. Two reading methods combined

Although one can apply the above two reading methods separately, in most cases they can be combined to achieve better reading results. One effective way of reading is to use both methods flexibly, i.e. switch from one to another as the case may be. In normal cases, one will follow the first reading method. But when he/she feels it necessary to stop at a certain word and want to check in some detail the typical use of the word, he/she can apply the second reading method.

Whenever one is not sure whether a certain pattern is typical for a certain word, he/she can resort to native speaker corpora and a concordancer to read both horizontally and vertically just as shown in the previous section. Whenever he/she is not sure a certain collocation or chunk is a typical one, he/she can simply search the collocation or chunk in the native speaker corpora and see how many can be found as to decide it is typical or not. For reasons of space, such searching are not exemplified here.

All in all, if a text is to be effectively read, these two methods have to be combined and a balance has to be achieved. One principle of using such methods of reading is that we need to ensure that one can learn more by reading this way and their interest in reading is enhanced by exploring language themselves.

7. Conclusion

It is held in this paper that notions such as pattern, collocation and chunk are critical to the use of words. To learn the typical uses of common words, one need first to be aware of the immediate linguistic context which defines patterns, collocations or chunks. Though it is possible to notice patterns, collocation or chunks in the normal way of reading, i.e. reading method 1, it is not reliable to consult one's intuition, introspection or retrospection about the typicality of the uses of words. Therefore, the second way of reading introduced in the paper is a necessary supplement to the first one. Keeping a balance between these two reading methods will not only enhance the effective learning and acquisition of the central uses of common words but may also enhance the learners' interest in and enthusiasm for learning English well by way of exploring and experiencing natural, authentic language themselves.

References:

Firth, J. R. 1957. 'A synopsis of linguistic theory, 1930-1955', in Liu, R. Q. *et al.* (eds.) 1988. *Readings in Linguistics: Seventy-five years since Saussure*. Beijing: CeHui Press.

Hunston, S. & Francis, G. 1998. 'Verbs observed: a corpus-driven pedagogic grammar', *Applied Linguistics* 19/1: 45-72.

Hunston, S. & Francis, G. 2000. Pattern Grammar: A corpus-driven approach to the lexical grammar of English. Amsterdam: John Benjamins Publishing Company.

Kjellmer, G. 1991. 'A mint of phrases', in Aijmer, K. & Altenberg, B. (eds.). *English Corpus Linguistics: Studies in honor of Jan Svartvik*. London: Longman.

Krishnamurthy, R. 1987. 'The process of compilation', in Sinclair, J. M. (ed.) *Looking Up: An account of the COBUILD Project in lexical computing*. London: Collins.

Li, W. Z. 1999. *An Analysis of the Lexical Words & Words Combinations in the College Learner English Corpus*. Unpublished Ph.D. thesis. Shanghai Jiao Tong University.

Pu, J. Z. 2001. *Learner Behaviour of Verbs: A corpus-based research on Chinese college students' use of English verbs*. Unpublished Ph.D. thesis. Shanghai Jiao Tong University.

Scott, Mike. 1998. *Wordsmith Tools* (Version 3.00). Oxford: Oxford University Press.

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

Sinclair, J. M. & Renouf, A. 1988. 'A lexical syllabus for language learning', in R. Carter & McCarthy, M. (eds.) *Vocabulary and language teaching*. London: Longman.

Wei N. X. 1999. *Towards Defining Collocations: A practical scheme for study of collocations in EAP texts*. Unpublished Ph. D. thesis. Shanghai Jiao Tong University.

A Study of Chinese English learners' Chunk competence

Diao Linlin
PLA Foreign Language University, Luoyang, 471003

Abstract: Although chunks have aroused growing interest of linguists for its unique function in L2 acquisition and large stacks of theoretical studies have been conducted about it, there is still inadequate empirical study concerned about it, especially when Chinese English learners are involved. This paper is therefore devoted to investigate chunk in an empirical way, applying both quantitative and qualitative methods so as to get a comprehensive picture of chunk competence for Chinese English majors and investigate the relationship between chunk competence and language proficiency. We adopt the form of survey consisting of multiple choice and translation to test chunk receptive knowledge and productive competence respectively in two grades which are assumed to be at two different levels of language proficiency. Since we assume that frequently used words are the very basic lexical items and consequently chunks associated with them should also be fundamental to language use, we concord authoritative corpora such as BROWN, FROWN, LOB and FLOB for the most frequent words and their chunks in native speakers' use, from which we select testing items to examine their acquisition by Chinese English majors. It turns out that the results correspond perfectly with the previous theoretical hypothesis that the more advanced learners are, the more likely they are to achieve high chunk competence and this further leads to the conclusion that chunk competence is an important component, or rather indicator, of language proficiency. In addition, according to statistical analysis, Chinese students' chunk competence as a whole is far from satisfactory, highlighting the fact that they simply ignore chunks in their L2 acquisition, no matter how advanced they are, and indicating an urgent need to introduce chunk into the teaching practice.

Key words: chunk competence, language proficiency, Chinese English learners, survey

1. Introduction

Up to now, a large body of research has consistently shown that chunk plays an essential and fundamental role in SLA. It enhances the level of idiomaticity and fluency and helps learners acquire grammatical rules. But it should be said that this field still fails to receive adequate attention. Most researchers, teachers and learners remain accustomed to glue on the acquisition of grammar and discrete words and consequently are haunted by the same problem of non-idiomaticity and dysinfluency. What's more, due to the lack of consistent empirical studies, there are still a lot of theoretical conclusions waiting to be testified, especially ones of Chinese English learners' background. Therefore, this paper attempts to present a comprehensive picture of the actual chunk competence of Chinese English majors and investigate the relationship between chunk and language proficiency, which is supposed to be of great theoretical and practical significance.

1.1 The definition of chunk

In recent years, more and more people began to observe the fact that much of what we say is formulaic---prestored in multiword units for quick retrieval (eg. Becker 1975; Bolinger 1976; Sinclair 1991; Nattinger

and DeCarrico 1992; Ellis 1996; Wray 1999; Wray and Perkins 2000). They put forward a huge set of definitional terms, such as *formulaic sequence, sentence stem, prefabricated routine, multiword units, phraseological unit, ready-made utterance and chunk*, to describe similar language phenomenon, but fail to reach an agreement in both naming and definition. In this paper, we define chunk as *a sequence of continuous or discontinuous words (below the sentence-level), occurring more frequently than probability would predict and enjoying certain restrictions in semantic, syntactic and pragmatic aspects so that it can be stored in and retrieved from people's mind as a whole meaningful unit.*

Typical chunk may present such characteristics as irregularity and flexibility. The former involves both semantic and syntactic requirement. Semantic irregularity refers to the feature of non-compositionality which means that the meaning of a word combination cannot be worked out from its constituent words plus syntactic relation. For example, when someone *kicks the bucket*, he is not taking a physical action concerned about his foot. Instead, it simply means he passes away. Syntactic irregularity means words may be used in a unusual way not predicted by grammatical rules, such as *by and large, of course* and *as well as.*

The latter means most chunks are capable of taking different morphological forms, that is, adapting to different persons, tenses and so on, and accepting closed class variation such as pronouns. Into this group fall chunks like 'NP TENSE apologize to NP for NP', which may give us varied forms such as *I apologize to her for being late* or *he apologized to the boss for having been late.*

1.2 The previous study of chunk

Based on the previous studies, chunk is supposed to speed processing and production and consequently save precious processing resources to spare more efforts in other demanding concurrent tasks (Wray. 1990, 1992). At the same time it may facilitate the social interaction through providing conventional and preferable ways of expressing ideas. As far as pedagogy is concerned, chunk may help L2 learners internalize and establish grammatical rules (Wong-Fillmore,1976:300) as well as facilitate the acquisition of the depth and width of lexical knowledge since chunk may involve various information about its frequency, pragmatic constraints, grammatical properties, collocation, and semantic features.

Most important of all, lots of researchers believe that native speaker's linguistic competence has a large and significant chunk component (Howarth:1998). Wray (1999:227) points out chunks seem to be in the repertoire of all types of speaker and all types of speaker use them to achieve specific interactional goals, such as greeting and chastising. Pawley and Syder (1983) suggest the formulaic sequences used by native speakers are not easy for learners to identify and master, and that their absence greatly contributes to learners' not sounding idiomatic. Besides, teachers and material writers (e.g. Flower and Berman 1989; Harmer and Rossner 1991; McCarthy and O'Dell 1994; Redman and Ellis 1991) also show increasing attention to the necessity of learnrs to acquire knowledge of chunks and are aware that this component of competence should be addressed explicitly.

Although such theoretical assumption has already been proposed, little empirical attempt is made to prove its validity. Cowie and Howarth (1996a), looking in great detail at a small amount of data, found out a measurable overlap in collocational use between less proficient native speakers and more advanced non-native writers. Therefore, they came to a conclusion that phraseology is a significant component of native and non-native proficiency. Other researchers explored from the opposite direction which could also lead to the similar conclusion. Granger examined non-native written academic English and revealed that

"learners use fewer prefabs than their native-speaker counterparts" (1998: 151) and have much less sensitivity to collocational relationships. All in all, there is still far from enough empirical study in exploring the relationship between chunk competence and language proficiency, especially as far as Chinese English learners are concerned.

2. Methodology

This paper is therefore devoted to investigate chunk in an empirical way, applying both quantitative and qualitative methods so as to provide practical evidence for the hypothesis which can be concluded from the previous theoretical studies: there exists correlation between Chinese English students' chunk competence and their language proficiency, and the former should be regarded as one of the important component of the latter.

2.1 Subject

The subjects in this testing were composed of two grades, the first-year and fourth-year, from the English Department of PLAUFL, altogether 280 students. Among them there are 212 male students and 68 female students.

All these subjects use exclusively Mandarin Chinese as their mother tongue. Before recruited into this university, they had learned English for at least six years in junior and senior middle school. Undergoing similar studying experience within the same teaching system, they can be regarded as representative of two sections along a continuum of development as far as learners' chunk competence is concerned so that our study is possible to give a more comprehensive idea. All the seniors, except for 5 students, have passed the Test for English Majors-Grade 8 (TEM-8).

The main reason for selecting particularly English majors in this research is that previous studies, such as Qian (1999) and Bahns and Eldaw (1993) in their experiments on learners' collocational competence, proved that advanced learners have reached certain level of English proficiency at which stage problems in chunk competence have emerged and turned out to affect their further study.

2.2 Instrument

Since it is widely acknowledged that vocabulary proficiency is composed of two parts: receptive knowledge and productive competence. The chunk competence test is composed of two parts: 30 multiple choices to test testees' receptive chunk knowledge and 20 items of translation to test their chunk productive competence. In the process of making testing paper, we resort to many corpora such as FROWN, BROWN, FLOB, LOB and CLEC for lexical information and rely heavily on professional software WORDSMITH for concordance. All the data were entered the computer and processed by the statistical software SPSS 11.5.

2.3 Preparation for the testing paper

2.3.1 Preparation for multiple choice

As for this part, we focus mainly on chunks of frequent words based on the assumption that frequently used words should be the very basic lexical items and chunks associated with them should also be fundamental to language use. After using WORDLIST, one of the tools in WORDSMITH, we made a wordlist of 1000

frequent words after examining the four authoritative corpora: FROWN, BROWN, FLOB and LOB, and set out to selecting proper items for testing. Since this test specifically meant to test learners' ability of recognizing chunks, we would like to choose especially those lexical words which showed strong tendency to associate with other words and had concrete meaning of its own. We also attended to the degree of word's difficulty so that testees' attention will not be diverted and the validity was guaranteed. All the target words are supposed to have been mastered by all the testees', that is, they should have certain knowledge about these basic words. Adhered to the above principles, we carefully picked out 30 items, including 21 nouns, 6 verbs, 2 adjectives and 1 preposition.

In the second step, we turned to CONCORD, another tool in WORDSMITH for chunks related to these target words. We set up one node and concorded the whole four corpora for the most frequent as well as meaningful co-occurred words. Since the target words are all rich in chunks as we have mentioned earlier, we designed the multiple choice with more than one correct answer, hoping to reflect the true picture of testees' chunk competence and avoid the interference from guessing. As for the selection of distractors, we tried to find resources from CLEC for chunk errors that can be detected. At the same time, we also filled in distractors with synonyms, semantically matching words or typical errors come across in teaching practice.

In terms of scoring, each option correctly chosen would be rewarded one point and the total marks will be 70. A wrong answer would result in one point deducted for punishment in order to prevent testees' from guessing. At the same time, in order to further guarantee the reliability of the testing and downsize the negative effect of guessing, we also employed the weight method to introduce into scoring the parameter of confidence so as to reflect testees' competence in an indirect way. The testing question would be presented in the following form:

We need to _____ (our) attention to the air pollution.

A. pay B. arise C. draw D. arouse

你答对本题的信心是：

① 完全有信心 ② 比较有信心 ③ 一般 ④ 信心不足 ⑤ 完全没信心

Every item was followed by a scale of confidence among which five scales, from number 1 to number 5, were given 5, 4, 3, 2 and 1 points respectively. Once the testee chose the right option, for example option A, and □ in the confidence scale, the score of the confidence, 5 points in this case, will be added to the total score. On the contrary, if he picked out the wrong option, for example B, the score of the confidence for this one will be deducted for punishment. Following this method, testees are bound to get three marks for part one: the primitive score, the score of confidence, and the weighted score.

2.3.2 Preparation for chunk translation

The second part made use of translation to test testees' productive competence of chunk. In consideration of both creativity and fixedness, we chose particularly two-word and three-word chunks occurring most frequently and supposed to be essential to language use. Again with the help of WORDLIST, we made a list of two-word and three-word word clusters after concording the four corpora. Removing meaningless clusters such as the combination of grammatical words *of the, in the,* and *of a,* we got chunks most frequently used by native speakers as testing items, trying to make a comparison between native speakers and L2 learners. To be specific, we extracted examples containing these chunks from Collins COBUILD Dictionary and the above four corpora. Of course, they underwent certain change if necessary so that

testees would not be distracted by difficulties in comprehension or unfamiliar words. We translated these idiomatic examples into Chinese to make up and paid particular attention to highlighting chunk in its Chinese version, trying to avoid triggering individual word with similar meaning and elicit chunks as much as possible.

Far different from the traditional translation testing, we adopted a special way in scoring thanks to our special purpose. Since we simply wanted to have an idea of learners' competence in producing target chunks, we put the answers into altogether 4 ranks according to their performance. When testee could give the very target chunk, the answer belonged to the first grade and 3 points were given; chunk expressing similar meaning but correctly given fell in the second grade, earning 2 points; individual words with the similar meaning would be put to the third grade and won 1 point. As for those incorrect chunks or words, or when testee simply ignored it, of course, no point would be given.

3. Results and discussion

Altogether 280 students, 117 from Grade 4 and 163 from Grade 1, took part in the survey, and 273 testing papers turned out to be valid. 6 were excluded because the testees did not follow the requirement of the survey, failing to provide the confidence index or providing it in a wrong way. Another one was excluded because he neglected the last page of the paper apparently due to carelessness.

Since students in grade 1 and grade 4 are supposed to be at two different levels of second language proficiency, we assume that the relations between their performance can be used to explore the relations between chunk competence and language proficiency.

Since the receptive knowledge and productive competence we tested belong to different dimension and cannot be added together mechanically, we compared the two respectively. What's more, we assume that if learners really have mastered knowledge about certain chunks, he will be pretty certain about his answers and thus achieve high confidence index, so confidence is also set as an index indirectly reflecting learners' grasp of chunk knowledge. Therefore, it is also included as one aspect of comparison. We select Independent-Samples T Test to measure their relations. Below is the table of the product from SPSS.

Table 1: Independent-Sample T Test between two grades

	Primitive multiple-choice	Weighted multiple-choice	Translation	Confidence index
T	-7.412	-6.482	-3.909	-6.189
df	271	271	271	271
Sig (2-tailed)	.000	.000	.000	.000

The table shows that significant differences exist between the two grades in all the four categories ($p<.05$). It is clear that in terms of both receptive knowledge and productive competence of chunk, advanced learners tend to achieve higher marks and wide gap between two levels of proficiency leads to fairly significant differences in their performance, which is manifested by the value of p, 0.000. The higher confidence index also demonstrates that more advanced learners are far more certain about their performance, freer from the negative effect from guessing and thus indicates more versed command of such knowledge. These statistics provide us with hard evidence that there does exist close correlation between chunk competence and second language proficiency.

After the comparison between two levels of proficiency, we examined the testees as a whole. Since the full mark for multiple choice is 70, 420 for weighted multiple choice and 60 for translation, if we follow the common practice and set the passing score as the 60% of the total, student would not pass until he achieved 42 for primitive score of multiple choice, 258 for weighted score and 36 for translation. According to this criterion, we conduct One-Sample T test to the three items and get the following results.

Table 2: One-Sample T Test of all the testees

Category/test value	T	Df	Sig. (2-tailed)	Mean difference
Primitive/ 42	-44.800	272	.000	-19.48
Weighted/ 258	-56.911	272	.000	-137.37
Translation/ 36	-6.482	272	.000	-2.533

The table showed clearly a significant difference between practical scores and theoretically assumed ones. Besides, the mean differences were -19.48, -137.37 and -2.533, which meant that practical scores were all far lower than expected ones. As a result, we can claim that no matter how different the two grades are in levels of language proficiency, their performance on the whole does not satisfy the requirement they are supposed to achieve. In other words, all of them are incompetent as far as chunk use is concerned.

The above analysis can lead us to such conclusions. In the first place, this result corresponds perfectly with the previous theoretical and empirical studies which believe that chunk competence is an important, or rather essential, component of language proficiency. Chunk competence is of great significance to second language proficiency and should be regarded as one of its important criterion. Sinclair (1991:110), as a result of his experience in directing the COBUILD project, the largest lexicographic analysis of the English language to date, proposes the principle of idiom to claim that native speakers usually give priority to a large number of semi-preconstructed phrases as their single choices for the sake of the economy of effort or due to the pressure of conversation, which reveals the fact that chunk contributes a lot to native speaker's language proficiency. As for L2 learners, it follows that the more advanced L2 learners are, the closer they are to native speakers as far as language proficiency is concerned, and consequently the more competent they are in either recognizing chunks or producing them. Besides, this further reminds us that words, traditionally assumed as the very base of language use, are over and over again shown not to operate as independent and interchangeable parts of language, but constituents of chunks which should be brought to prominence.

In the second place, the result of One-Sample T Test indicates that Chinese English learners' chunk competence is far from satisfactory. Traditional teaching practice holds vocabulary is composed of single words and fixed phrases, giving no notice of chunks. To make matters worse, it always prefers rote learning to expand vocabulary. The severe consequence is that students usually rely on semantic matching to remember single word while neglecting lots of information provided by related chunk or sacrificing idiomaticity and fluency. This is typically illustrated by the testing item of 'strong tea'. According to analysis, 20.5% grade-one students and 35.1% grade-four students select 'heave tea' or 'dense tea' as acceptable expression for the same meaning. Apparently, they simply find English equivalent for Chinese word '浓' and apply grammar to create awkward expressions by themselves, unaware of the fact that native speakers already have a preferable and conventional one. In addition, when we try to test learners' grasp of semantic prosody, for example that of 'commit', it turns out that 26.7% of grade-one students have no idea that 'commit' features a negative semantic prosody and therefore regard 'commit the matter' or 'commit the responsibility' as correct. All these reflect the negative influence that traditional teaching exerts and explain the fact why vocabulary remains to be a headache for the majority of Chinese English learners. It is clear that our vocabulary learning calls for the introduction of chunk so as to improve itself more efficient and productive.

4. Conclusion

This research adopts the form of survey to test Chinese English learners' chunk competence from the aspects of both receptive knowledge and productive competence. The results on the one hand prove the previous theoretical hypothesis that chunk competence is highly correlated with language proficiency and should be attached importance to as an important indicator or component of the latter. On the other hand, the survey describes the Chinese English learners' chunk competence as inadequate due to, to a great extent, the traditional heritage of ignoring chunk. Thus, it turns out to be an urgent need to expand the concept of vocabulary as consisting of single word as well as chunk and introduce chunk into L2 teaching to make up for the defect of traditional teaching. Further study should penetrate into this field for more significant findings, such as the relationship between chunk competence and discrete language skills, learners' strategy in using chunks and how to introduce chunk into teaching practice due to its great difference from single words, which are all supposed to be of great theoretical and pedagogical significance.

References:

Bahns. J. & M. Eldaw. 1993. 'Should we teach EFL students collocations?' *System* 21:101-14.

Becker, J. 1975. *The phrasal lexicon.* Bolt Beranek & Newman Report no. 3081, AI Report no. 28.

Bolinger, D. 1976. 'Meaning and memory.' *Forum Linguisticum* 1/1:1-14.

Cowie, A. P. and P. Howarth. 1996a. 'Phraseological competence and written proficiency' in G. M. Blue and R. Mitchell (eds.): *Language and Education* (British Studies in Applied Linguistics 11). Clevedon: Multilingual Matters.

Ellis, N. C. 1996. 'Sequencing in SLA: phonological memory, chunking and points of order.' *Studies in Second Language Acquisition.* 18: 91-126.

Flower, J. and M. Berman. 1989. *Build Your Vocabulary.* Hove: Language Teaching Publications.

Granger, S. 'Prefabricated Patterns in advanced EFL writing: Collocational and Formulae.' In Cowei. A. P (eds.): *Phraseology: Theory, Analysis and Applications.* Oxford: Oxford University Press.

Harmer. J. and R. Rossner. 1991. *More Than Words.* London: Longman.

Howarth, P. 1998. 'Phraseology and second language proficiency.' *Applied Linguistics* 19: 24-44.

McCarthy, M and F. O'Dell. 1994. *English Vocabulary in Use.* Cambridge: Cambridge University Press.

Nattinger, J. R. and J. S. DeCarrico. 1992. *Lexical Phrase and Language Teaching.* Oxford: Oxford University Press.

Pawley. A. and F. H. Syder. 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency' in J. C. Richards and R. W. Schmidt (eds.): *Language and Communication.* London: Longman.

Qian, D. 1999. 'Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension.' *The Canadian Modern Language Review.* 52: 282-307.

Redman, S and R. Ellis. 1991. *A Way with Words.* Cambridge: Cambridge University Press.

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

Wong-Fillmore, L. 1976. *The Second Time Around: cognitive and Social Strategies in Second Language Acquisition.* Doctoral dissertation, Stanford University.

Wray, A. 1990. 'The dual systems hypothesis: a right hemisphere account for left hemisphere language.' *Speculations in Science and Technology* 13/1:3-12.

Wray, A. 1992. *The Focusing Hypothesis: The Theory of Left Hemisphere Lateralized Language Re-examined.* Amsterdam: John Benjamins.

Wray, A. 1999. Formulaic language in learners and native speakers. *Language Teaching.* 32:213-231.

Wray, A. and M. R. Perkins. 2000. 'The functions of formulaic language: an integrated model'. *Language and Communication.* 20:1-28.

A Corpus-based Study on Adjective Intensification in Chinese EFL Learners' Writing

Chen Jiansheng
Tianjin University of Science and Technology

Abstract: Adjective intensifiers are a category of adverbs which can modify gradable adjectives. A peripheral category as they belong to, adjective intensifiers play an extremely important role in both oral and written communication. Through intensifying or weakening the meaning of an adjective it modifies, the language user can express such modalities as assertion, emphasis, uncertainty or doubt, and his attitude and stance. It is this communication function that reveals the linguistic skill of a language user.

This paper, based on corpus linguistics, compared and analyzed the uses of adjective intensifiers in both Chinese EFL undergraduate English essay corpus and native English student essay corpus. It is found that there are differences in the use of adjective intensifiers in the English argumentative essays written by native speakers of English and the Chinese EFL learners, not only in quantity but also in type. The statistical data drawn from the EFL learner and native speaker essay corpora indicate that the Chinese EFL learners overuse adjective intensifiers which fall into only a few types. The underuse of downtoners of different degrees also constitutes a major problem for EFL learners. The analysis of both the data and examples in the essays proper shows that these problems are mainly caused by the fact that EFL learners excessively rely on closed-class intensifiers for adjective intensification while at the same time fail to use open-class intensifiers actively. Open-class intensifiers are believed to be more creative and expressive than the closed-class ones. As far as the information structure is concerned, no significant problem has been found in the learners' corpus that is caused by the misuse of adjective intensification by the EFL learners at the upper intermediate level, but the suggestion is made that attention should be paid to its development tendency.

Key words: adjective intensifier, EFL learner corpus, English language teaching

I. Introduction

When we read an essay written by a learner of English, we can often have the feeling that this essay is not written by a native speaker of English because there are great differences between the essays written by an English learner and by a native speaker of English. Our experience tells us that these differences exist in many aspects, both in grammar and in vocabulary. It is these differences that make the essays written by learners of English foreign sounding and non-idiomatic. Some of these differences, for example, grammatical mistakes in the uses of tenses, non-finite verbs and sub-clauses, are quite obvious; others such as lexical and stylistic mistakes or misuse, are not so obvious, because these are subtle differences which can only be found through analysis and comparison of a large number of texts written by both the learners and native speakers. With the development of corpus linguistics and its wide application in English language teaching and research, many researchers of interlanguage have begun using learner corpora in their studies of these differences, for example, the study of English modal verbs in the interlanguage of

Swedish learners of English by Aijmer (2002), the study of the discourse marker *so* in Chinese learners' written English by He (2002), the study of adjective intensification in German learners' English argumentative essays by Lorenz (1998, 1999). These learner-corpus based studies compare and analyze the data collected from the corpora, describe the minute differences which cause foreign-soundingness of learner's English and draw conclusions which can improve EFL teaching.

This paper intends to study the uses of English adjective intensifiers in the essays written by the Chinese non-English major EFL learners in order to find the differences between the Chinese learners of English and native speakers of English in this aspect and the causes of these differences.

II. Adjective intensifiers

Adjective intensifiers are a category of adverbs which can modify gradable adjectives, for example: *very good, extremely difficult*. The term is used to differentiate between adverbs of this kind and those adverbs modifying verbs at the sentence level, for example: *greatly appreciate*. Quirk made a general distinction between amplifiers and downtoners: the former refer to the adjective intensifiers which amplify or strengthen the meanings of the adjectives they modify, eg. *a very funny film*, the latter refer to those which tone down the degree of the meanings of the adjectives they modify, eg. *barely intelligible* (Quirk et al., 1985: 445)。

A peripheral category in syntax as they belong to, adjective intensifiers play an extremely important role in both oral and written communication. Through intensifying or toning down the meaning of an adjective it modifies, the language user can express such modalities as assertion, emphasis, uncertainty or doubt. It is this communication function that reveals the linguistic skill of a language user.

III. Study on adjective intensifiers in Chinese non-English major EFL learners' essays

As has been stated, the use of adjective intensifiers can reveal the writing ability of the learners of English. Then what is the situation in this respect for the Chinese non-English major undergraduates who constitute the major part of learners of college English in China? This paper describes how this group of learners use adjective intensifiers in their English writing with the data drawn from the Non-English Major EFL Undergraduate English Essay Corpus[1]. References are made to the data from the native English essay corpora and British National Corpus. The data are compared and analyzed in order to find problems in the use of adjective intensifiers by the Chinese undergraduates to improve college English teaching in China.

IV. Non-English Major Undergraduate English Essay Corpus and the reference corpora

The data used in this study come from Non-English Major EFL Undergraduate English Essay Corpus (referred to as the learner corpus hereafter), which is composed of argumentative essays written by the non-English major undergraduates from 13 universities of Tianjin who have finished three semesters of college English course..The total word number of the corpus is 165,133. The essays, each about 130 to 210 words in length and written in a 30- to 40-minute test, cover a wide range of topics and have got grades between 9 to 13 (15 being the full mark). Therefore, the data from this corpus can be used to study the written production of the Chinese college English learners at or above the upper intermediate level.

The learner corpus has been POS tagged automatically with C7 tag set[2] and the tagged version has been manually checked so that over 98% of the words in the corpus are correctly tagged. The spelling mistakes in the original essays are marked up with angle brackets and the correct spelling is given immediately after the right bracket for the convenience of computer automatic tagging and information retrieval.

The data used for the comparative study come from GCE and LOCNESS (referred to as the native corpus hereafter), which are the two corpora used by Lorenz (1999). These corpora are composed of argumentative essays written by native British teenagers and undergraduates respectively with the total word number of 160,557, about the same size as that of the learner corpus. In the Appendices of Lorenz (1999: 246-321) the detailed information about adjective intensifiers used in these two corpora is given, providing other researchers with convenient references to these corpora. Besides, the present study also consults the 100-million word British National Corpus (BNC) for reference.

V. Data collection and analysis

The total occurrences of adjective intensifiers used in the learner corpus are counted and standardized to x per 100,000 words. Then, the results are compared with the occurrences of adjective intensifiers in the native corpus (Table 1).

Table 1. Comparison between occurrences of adjective intensifiers in the learner corpus and those of the native corpus. (SF stands for standardized frequency; adj-int stands for adjective intensifiers)

Corpora	learner corpus	native corpus
total word no.	165133	160557
adj-int (RAW)	904	751
adj-int (SF)	547	468

The data in Table 1 show that the Chinese EFL undergraduates use more adjective intensifiers than the native speakers. The value of χ^2 test is 6.18, indicating that there is a significant difference between the two. But the count of tokens and types shows that the EFL learners use much fewer types of adjective intensifiers than native speakers, with $\chi^2 = 78.25$, indicating an extremely significant difference between the two (Table 2).

Table 2: Comparison of type-token ratios between the learner corpus and the native corpus

	learner corpus	native corpus
adj-int token	904	751
adj-int type	42	144
type-token ratio	4.65	19.17

As can be seen from Table 2, the adjective intensifiers used by the EFL learners are limited to a few types as compared with those used by the native students. Table 3 lists the top ten adjective intensifiers both in the learner corpus and the native corpus. It can be seen clearly that the top ten adjective intensifiers in the learner corpus take up almost 91% of the total number of adjective intensifiers while in the native corpus, the top ten only take up 47%.

Table 3: Top ten adjective intensifiers in both the learner corpus and the native corpus.

Adj-int (learner corpus)	**Freq. (RAW)**	**Freq. (SF)**	**Adj-int (native corpus)**	**Freq. (RAW)**	**Freq. (SF)**
very	526	318.2	very	190	118.3
more and more	116	70.2	so	40	24.9
so	76	46.0	quite	32	20.0
too	51	30.9	too	23	14.3
not so	23	13.9	extremely	19	11.8
much	18	10.9	totally	17	10.6
not very	16	9.7	rather	16	10.0
quite	14	8.5	particularly	15	9.3
really	13	7.9	really	12	7.5
enough	11	6.7	highly	12	7.5
Total (raw)	**864**		**Total (raw)**	**376**	

Quirk *et al.* (1985: 67) divided English word classes into two general categories, ie. the closed-class and open class categories. The closed-class category includes articles, prepositions, conjunctions, to which no new words can be added. The open class category includes nouns, adjectives, lexical verbs and adverbs, to which new words can be added constantly. Adjective intensifiers are adverbs which belong to the open class category, and we can further divide them into the closed-class and open class categories. The closed-class intensifiers are a finite, non-productive set. The forms of these words will not change. They are limited in number and no new words can be added to this set. For example: *all, almost, enough, indeed, little, most, rather, so, somewhat, too, very, well.* The open class intensifiers are formed by adding suffix *-ly* to an adjective, therefore, new items can be added to this category constantly. For example: *greatly, powerfully, quickly.*

Since the open class intensifiers can express such concepts as degree, modality, evaluation, comparison more accurately and specifically, Lorenz suggests that the types of adjective intensifiers used in the writing can be an indicator of the writer's English level. The more competent the writers, the fewer close-class intensifiers they are likely to use (Lorenz, 1999: 79). The competent writers use open class intensifiers to express their meanings creatively.

As the data in Table 3 indicate, although among the top ten adjective intensifiers, 5 types (*very, so, too, quite, really*) appear in both the learner corpus and native corpus, there are great differences in frequency and in categories. The learners used 9 closed-class intensifiers[3] and only one open class intensifier (*really*), while the native students used 5 open class intensifiers (*extremely, totally, particularly, really, highly*). Of all the 42 adj-int types used by the learners, the ratio of the closed-class to open class intensifiers is 8.9:1.1 . This ratio is 1.6:8.4 for the 144 types in the native corpus.

In both the learner corpus and native corpus, *very* is the most frequently used adjective intensifier. However, *very* takes up only 25% of all the adjective intensifiers in the native corpus while it takes up 58% in the learner corpus. This large percentage is partly responsible for the overuse of adjective intensifiers by the learners.

To further understand the overuse of *very* by the learners, adjectives which are intensified by *very* over ten times are counted[4] and the results are compared with the corresponding data drawn from the native corpus (See Figure 1).

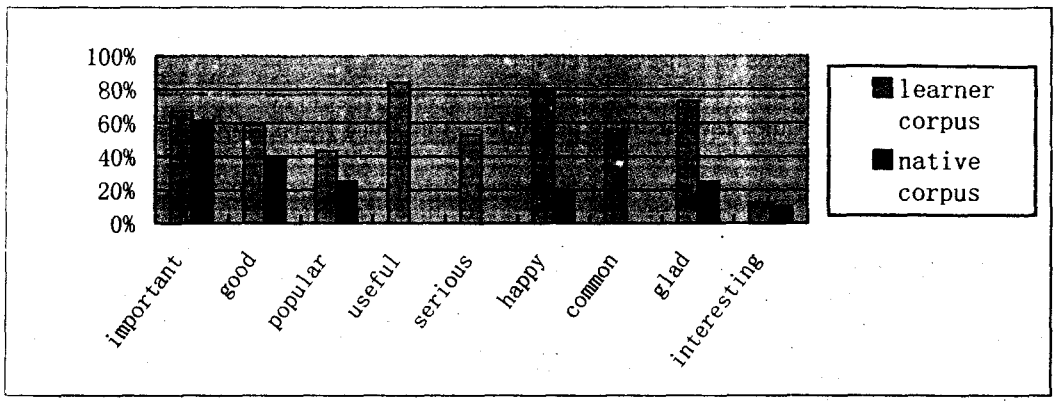


Figure 1: Adjectives modified by *very* over 10 times in the two corpora.

We can find in Figure 1 that learners overuse *very* as an intensifier for all these adjectives. For example, nearly 60% of all the occurrences of the adjective *good* are intensified by *very* in the learner corpus, while in the native corpus it is only 40%. Also, in the native corpus there are no cases where the adjectives *useful*, *serious* and *common* are intensified by *very*. In fact, some adjectives in English are rarely intensified by *very*. In BNC, for example, there are 19,866 occurrences of *common*, of which only 297 occurrences (1.5%) are intensified by *very*. But in the learner corpus, *common* occurs 78 times, 11 of which (14%)[5] are intensified by *very*.

The overuse of *very* by the learners may be accounted for in the following two ways. First, the open class intensifiers which the Chinese undergraduates can use with confidence are limited in number. *Very* is a widely used adverb and learners feel safer to use it as an adjective intensifier than to use other intensifiers. Second, learners may feel that single adjectives are not powerful enough in expressing the meaning or to attract the reader's attention, therefore they use *very* as intensifiers to emphasize whatever they want to express. There are many such cases in the learner corpus:

(1) *On one hand, bicycle is cheaper than other traffic tools, on the other hand it is* ***very*** *easy to learn how to ride bike.*[6]

The intensifier *very* in (1) produces an impression of overstatement since common sense or our own experience tells us that, generally speaking, to learn to ride a bicycle may be easy but it is by no means very easy. Another reason for the learner's overuse of *very* may be that in learning English learners have got into the habit of adding *very* to common adjectives casually. It seems that *very important* (*good*, *useful*, *common*, etc.) have become fixed in the learner's English.

Another adjective intensifier overused by the learners is *more and more*, which, with 116 occurrences distributed in all the topics of essays, is second to *very* in Table 3. In contrast, there are only 6 occurrences of *more and more* in the native corpus. And in BNC, the total occurrences of *more and more* are 2,492 (2.5 per 100,000 words), including those of *more and more* used other than adjective intensifiers. This shows *more and more* is not frequently used in English. The overuse of this adjective intensifier may be caused by the misguidance of some test-oriented English writing instruction books in which students are taught to write an essay on the basis of a few skeleton sentences with gaps to be filled in by the students according to

the topic of the essay. For example,

With the development of _____, _____ is becoming more and more _____.

These skeleton sentences are repeated in the EFL learner's essays, making them stereotyped. In fact, in many cases *more and more* can be replaced by other open class boosters, such as *increasingly*. For example,

(2) ...which means it becomes **increasingly possible** to introduce legislation in the U.K.

Besides, in the learner corpus, there are 95 occurrences of the "link verb + *more and more* + adj." structure, 13 of which contain adjectives which cannot be intensified by *more and more*. For example: **more and more bad (busy, dirty, happy, large, lazy, little, narrow, rich, sharp, weak, wide)*. These adjectives do not take phrasal comparison by using the degree adverb *more*. If we are to express the concept of *more and more* + adj. with these adjectives, we should use the inflectional suffix *–er* or their irregular comparative forms, for example: *larger and larger, worse and worse.*

In Table 3, *much* is the sixth adjective intensifier frequently used by the learners. As a closed-class intensifier, *much* is not used to modify ordinary adjectives but used mainly to intensify V-ed adjectives. We only use *very much* to modify those adjectives used as complement, for example: *afraid, alike, alive, awake* (Sinclair *et al.*, 1990). In the native corpus, there are only 3 occurrences of *much* used as adjective intensifiers, two of which are used to intensify V-ed adjectives (*a much debated one, a much discussed topic*). However, a search in the learner corpus shows that there are 18 occurrences of *much* used as adjective intensifiers and all of them are used to intensify ordinary adjectives (Figure 2).

s_NN2 of_IO the_AT road_NN1 are_VBR **much_RR beautiful_JJ** ._. you_PPY will_VM se_
transportation_NN1 which_DDQ is_VBZ **much_RR cheap_JJ** and_CC convenient_JJ ._. Fo
G 10_MC years_NNT2 before_CS is_VBZ **much_RR different_JJ** from_II it_PPH1 ._. At_
VVD ,_, it_PPH1 make_VV0 them_PPHO2 **much_RR easy_JJ** contact_NN1 with_IW stranger
ecause_CS it_PPH1 can_VM become_VVI **much_RR good_JJ** ._. As_RG follow_VV0 I_PPIS1
to_II some_DD extent_NN1 ,_, is_VBZ **much_RR harmful_JJ** to_II everyone_PN1 ._. Wi
T1 ._. So_RR punctuality_NN1 is_VBZ **much_RR important_JJ** for_IF us_PPIO2 ._. I_P
AT fast_JJ food_NN1 would_VM be_VBI **much_RR popular_JJ** in_II the_AT world_NN1 ._
0 that_CST the_AT books_NN2 is_VBZ **much_RR useful_JJ** of_IO learning_VVG the_AT

Figure 2: Part of the concordance of *much* used as an adjective intensifier in the learner corpus (with a POS tag after each word).

As has been stated in Section II above, there are two kinds of adjective intensifiers: amplifiers and downtoners. The search in the corpora also reveals differences between the Chinese EFL undergraduates and native English students in this aspect. The data from the search in the learner corpus indicate that the ratio between amplifiers and downtoners is 93:7, which is much higher than the corresponding ratio (74:26) in the native corpus. Besides, in all 62 downtoners in the learner corpus, 47 are formed by *not + very / so / really*, for example::

(3) *Generally this marriage is **not very firm** and can result in the divorce.*
(4) *The general gap is **not so good**, we must eliminate it.*
(5) *And today, some books are **not really good**, they have some wrongs....*

Downtoners are composed of approximators (eg. *mainly, largely, nearly, practically, virtuely*), compromizers (eg. *fairly, comparatively, relatively, pretty, rather*), diminishers (eg. *slightly, mildly,*

possibly) and minimizers (eg. *barely, hardly, supposedly*), and the proper use of these downtoners in argumentative essays can make the arguments more persuasive and acceptable. But EFL learners use only a few of them in their writing, for example:

(6) *..., if you want to find a tutor, it's a good way to the Normal University, since the students they introduced are* ***probably good.*** (*probably* is a compromizer.)

The Chinese cultural background is not to blame for the underuse of downtoners by the Chinese learners of English because Chinese is not lacking in such expressions. The main reason is that the learners at this level have not mastered the use of adjective intensifiers of different degrees between the two extremes (maximum and minimum), especially the downtoners. Therefore, they can only use those general ones.

In addition to the comparison of the frequency and type of adjective intensifiers in the two corpora, the information structure can also be analyzed to find problems in the use of adjective intensifiers by the Chinese EFL undergraduates. According to the studies of Halliday (1985), Quirk *et al.*(1985: 1361) and Fries (1994), the theme of an English sentence is often the position where given information is expressed, while the rheme is the position where new information is expressed. Since the major function of adjective intensification is to highlight the interesting, relevant and new information for the readers, they are likely to appear in the position of a sentence where new information is expressed, ie. rheme. Besides, it seems that it is most effective to modify a single adjective (eg. predicative adjective) by an adjective intensifier. Therefore, the most typical use of adjective intensifiers is to intensify predicative adjectives in the rheme position. Look at a sentence taken from the native corpus:

(7) *However, the drop out rate is incredibly high, about 75%.*

In this sentence, the adjective intensifier *incredibly*, which is used to evaluate the concept of the adjective *high* by the writer, is in the rheme position.

In view of this consideration, the adj-int occurrences have been checked for attributive versus predicative position in the two corpora and the results are normalized to x per 100,000 words (Table 4).

Table 4: The adj-int occurrences for attributive versus predicative position in both the learner corpus and native corpus (SF stands for x times per 100,000 words.

	learner corpus	**native corpus**
in pred. position (SF)	485.1	352.5
in attr position (SF)	84.2	122.5
attr position (%)	15%	26%

The data from Table 4 indicate that of all the adjective intensifiers used by the learners, only 15% of them are used to intensify adjectives in the attributives position. This percentage is much lower than that in the native corpus. The close observation of the concordance in the learner corpus shows that there are only five occurrences in which the intensifiers are used to modify attributive adjectives in the theme position, for example,

(8) *Too*_RG large_JJ population_NN1 makes_VVZ living_JJ level_NN1 ...

All the other intensifiers are used either to modify predicative adjectives, for example:

(9) So_RR they_PPSH2 are_VBR *equally*_RR important_JJ ._.

or to modify attributive adjectives not in the theme position, for example:

(10) And_CC it_PPH1 can_VM finish_VVI task_NN1 in_II a_AT1 *very*_RG short_JJ time_NNT1

These search results show that the Chinese EFL learners at this level have no problem in information structure as far as adjective intensification is concerned. However, Lorenz (1999: 206) finds that the more advanced the EFL learners, the more adjective intensifiers they will use in the theme position. On the contrary, the more advanced the native learners, the fewer adjective intensifiers they will use in the theme position. Since the learner corpus used in the present study is not big enough to carry out comparative studies on adjective intensification between learners at different levels, this will not be dealt with further.

VI. Conclusion

In general, the Chinese EFL learners use more adjective intensifiers but much fewer types than native English students do in argumentative writing. The EFL learners tend to overuse *very* and *more and more*, which makes their essays overstating and stereotyped. College English writing courses can benefit from these findings. Teachers should help students understand fully what they realiy want to express in writing and try to avoid using *very*, *more and more* and other adjective intensifiers where they are not necessary or will create an impression of overstatement . In writing classes, students should also be taught to use the open class intensifiers, especially those downtoners which can tone down the meanings of the modified adjectives to certain degrees, so that they can use these more concrete and creative words in their essays consciously. In fact, the Chinese undergraduates at this level have already learned many adjectives which can be turned into open class intensifiers by adding *-ly*, for example: *entire*, *whole*, *full*, *high*, *wide*, *fair*, *general*, *large*, *main*, *near*, *practical*, *hard*, *poor*, but they have not been conscious of doing so actively. Therefore, it is teachers' task to remind students of using the adjective intensifiers of this kind in their essays. As advanced learners have the tendency of overusing adjective intensifiers in the theme position, they should learn something about the information structure of "theme-given information, rheme-new information" so that they will avoid overusing adjective intensifiers in the theme position.

In a word, such details as the use of adjective intensifiers can not be ignored in the EFL teaching and research if we wish to raise the English writing ability of the learners substantially.

Notes

① The corpus used in this study is part of the project which has been approved and financed by the Ministry of Education. The name of the project is "Study and Application of Non-English Major Undergraduate English Essay Corpus in English Language Teaching" (Project Code: 126303223)

② For detailed information about C7 tag set, refer to Appendix III of *Corpus Annotation: Linguistic Information from Computer Text Corpora* edited by Roger Garside, Geoffrey Leech and Tony McEnery (1997), Addison Wesley Longman Inc., or visit the following website: http://www.comp.lancs.ac.uk/ucrel/claws7tags.html

③ *Not so* and *not very* are also adjective intensifiers. They belong to downtoners.

④ The adjectives *convenient* and *tired* are also modified over ten times by *very* in the learner corpus, but because in the native corpus these two adjectives are not modified by any intensifiers, they are not listed here.

⑤ The word *common* modified by *very* has a distribution in the nine topics of the learner corpus.

⑥ The examples quoted here are the sentences or parts of the sentences taken from the original compositions in the learner corpus, and no corrections are made.

References

Aijmer, K. Modality in advanced Swedish learners' written interlanguage [A]. In *Computer Learner Corpora, Second Language Acquistion and Foreign Language Teaching* [C]. ed. S. Granger, J. Hung and S. Petch-Tyson. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002. 55-76.

Fries, P. H. On Theme, Rheme and discourse [A]. In Advance in Written Text Analysis [C] ed. M. Coulthard. London: Routledge, 1994. 228-249.

Halliday, M.A.K. An Introduction to Functional Grammar [M]. London: Edward Arnold, 1985. 38-67

He, Anping. On the discourse marker so [A]. In New Frontiers of Corpus Research [C]. ed. P. Peters, P. Collins and A. Smith. Amsterdam – Atlanta: Rodopi, 2002. 41-52.

Lorenz, G. R. Overstatement in advanced learners' writing: stylistic aspects of adjective intensification [A]. In Learner English on Computer [C]. ed. Granger, S. London: Longman, 1998. 53-66.

Lorenz, G. R. Adjective Intensification – Learners versus Native Speakers, A corpus Study of Argumentative Writing [M]. Amsterdam/Atlanta: Rodopi, 1999.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. A Comprehensive Grammar of the English Language [M]. London: Longman, 1985.

Sinclair, John McH, editor in chief. Collins COBUILD English Usage [M]. London: HarperCollins, 1990. 411.

A Study of Intensifiers in Chinese EFL Learners' Speech Production

Liang Maocheng
Nanjing University

1. Introduction

In English, degree adverbs, such as *very*, *slightly*, etc., are very often used to modify adjectives, adverbs and verbs, indicating the intensity of the meanings expressed by these words, and achieving accuracy in word meaning expression. Intensifiers, as these degree adverbs are traditionally termed (Quirk *et al* 1985, Biber 1999, etc.), do not always indicate the increase in the intensity of word meanings. Instead, they often correspond to the various points on the intensity scale. Therefore, the intensity attached to these intensifiers can vary considerably, from minimum intensity (as denoted by *hardly*) to maximum intensity (as denoted by *absolutely*) (Quirk *et al* 1985). Whether these intensifiers are used properly is an important measure of learners' language proficiency. Though syntactically not fundamentally vital to the structure of learner language, the intensifiers do play an important role in the written and spoken interaction, and convey different attitudes of the speaker or writer. Improper use of these subtle aspects of language can often lead to the non-nativelikeness or lack of idiomaticity, undercutting the effect of learner language, and exerting a negative effect on communication. (Lorenz 1998)

Quirk *et al* (1985:589ff) think that intensifiers constitute a gradable category. According to this widely-held view, intensifiers can be classified into the subcategories of *amplifiers* and *downtoners*, the former often used to increase the meaning conveyed by relevant words, while the latter often used to decrease the meaning conveyed.

Quirk *et al* (1985: 589ff) then further classify amplifiers into maximizers (*completely*, *absolutely*), and boosters (*very*, *highly*), and downtoners into approximators (*nearly*, *virtually*), compromisers (*fairly*, *fairly*), diminishers (*slightly*), and minimizers (*hardly*, *scarcely*).

As EFL learners' use of intensifiers is an important indicator of their language proficiency, studies of EFL learners' use of intensifiers can be of significant value to the understanding of the learners' interlanguage development. However, such studies have not been done extensively. The only study of such kind is that done by Lorenz (1998) on German EFL learners' written language. This paper employs what Granger (1998) terms 'contrastive interlanguage analysis' to study Chinese EFL learners' use of intensifiers in their speech production, and attempts to offer some pedagogical suggestions.

2. Research Questions

Lorenz's (1998) study indicates that German learners tend to overuse intensifiers, and this suggests that German EFL learners are guided by different principles when they try to convey new information and use intensifiers. German EFL learners' overuse of intensifiers contributes to the lack of idiomaticity of their language, and there is a tendency of overstatement in their L2 writing. A relevant question for the present study is:

A. Is there a tendency of overstatement in Chinese EFL learners' speech production? If yes, what causes the tendency?

In addition, one of the functions of intensifiers is to adjust the intensity of word meanings, suggest the

speaker's or writer's attitude, and attract the reader's or listener's attention to the new information to be conveyed (Lorenz 1998). Different intensifiers serve different semantic functions, and too much or too little use of any category of intensifiers will undoubtedly lead to the inaccuracy of the presentation of meaning. So a second question for the present study is:

B. As viewed from the overall frequencies of different categories of intensifiers used by the learners, can Chinese EFL learners make full use of the different categories of intensifiers to accurately convey their intended information in their speech production?

Furthermore, if there is a tendency of overstatement in Chinese EFL learners' speech production, it may be assumed that the tendency can be attributed to the insufficient repertoire of their intensifiers. If they indeed do not have a sufficient repertoire of intensifiers, they may assort to other means to make up for this insufficiency so as to continue their attempt to convey new information. One more question we will try to answer in this study is:

C. Is their an insufficiency of repertoire of intensifiers in Chinese EFL learners' speech production? If yes, what compensatory means do they employ to achieve their communicative goal?

3. Methodology

In an attempt to answer the questions above, the present study employs what Granger (1998) terms 'contrastive interlanguage analysis' to compare the findings from an English native speakers' corpus with those from a Chinese EFL learners' corpus of spoken English.

The native speakers' corpus employed in the study is the British component of the International Corpus of English (ICE-GB) developed by Survey of English Usage (SEU) at University College London. This corpus comprises of about a million tokens, of which about 3/5 is spoken and 2/5 is written. As the present study attempts to analyze some features of learners' spoken English, only the spoken component of ICE-GB (about 600 thousand tokens) is used.

The learner corpus used in the study is the Spoken English Corpus of Chinese Learners (SECCL), a corpus being constructed at Nanjing University. The corpus comprises transcriptions of speech recorded at the annual Test for English Majors Band 4 (TEM-4), a test for Chinese university students who take the English language as their major. In the present study, the transcriptions investigated amount to 42 million tokens, transcribed from TEM-4 testee recordings ranging in time from 1999 to 2002.

In order for the data in the two corpora to be more comparable, the list of intensifiers given in Quirk *et al* (1985) and the list of intensifiers given in Biber *et al* (1999:564ff) based on Longman Spoken and Written English Corpus (LSWEC) have been drawn on to decide on the list of intensifiers to be investigated in this study. Based on this criterion of selection, 22 most frequently used English intensifiers have been recruited for study, including *absolutely, completely, totally, perfectly, entirely, fully, very, so, really, too, extremely, highly, terribly, nearly, almost, pretty, rather, fairly, slightly, somewhat, hardly,* and *scarcely.*

4. Results and Discussions

4.1 Overall frequency of intensifiers

The overall frequency counts and absolute frequency (per 10 thousand tokens) of all categories of intensifiers in the two corpora are listed in Table 1. In every adjacent pair, the number on the left is the

number of the respective category of intensifiers, while the number on the right is the normalized frequency counts of the category in every 10 thousand tokens.

Intensifier categories		ICE-GB		SECCL	
		N	Frequency	N	Frequency
Amplifiers	maximizers	394	6.68	24	0.58
	boosters	3115	51.91	6218	148.06
	Sum	3509	58.59	6242	148.64
Downtoners	approximators	222	3.70	109	2.60
	compromisers	402	6.70	65	1.55
	diminishers	108	1.80	10	0.24
	minimizers	41	0.68	63	1.50
	Sum	773	12.88	247	5.89
Total		4282	71.47	6489	154.53

Table 1: Overall Frequency

As can be seen from Table 1, Chinese learners use a much larger proportion of intensifiers in their speech production than their English counterparts. Further investigation reveals that this is particularly true when it comes to the use of amplifiers. Chinese learners use over 2.5 times as many amplifiers as English adults (148.64:58.59) in their speech production.

Contrary to the case of amplifier use, the overall frequency of Chinese EFL learners' use of downtoners is considerably lower, with the exception of minimizers.

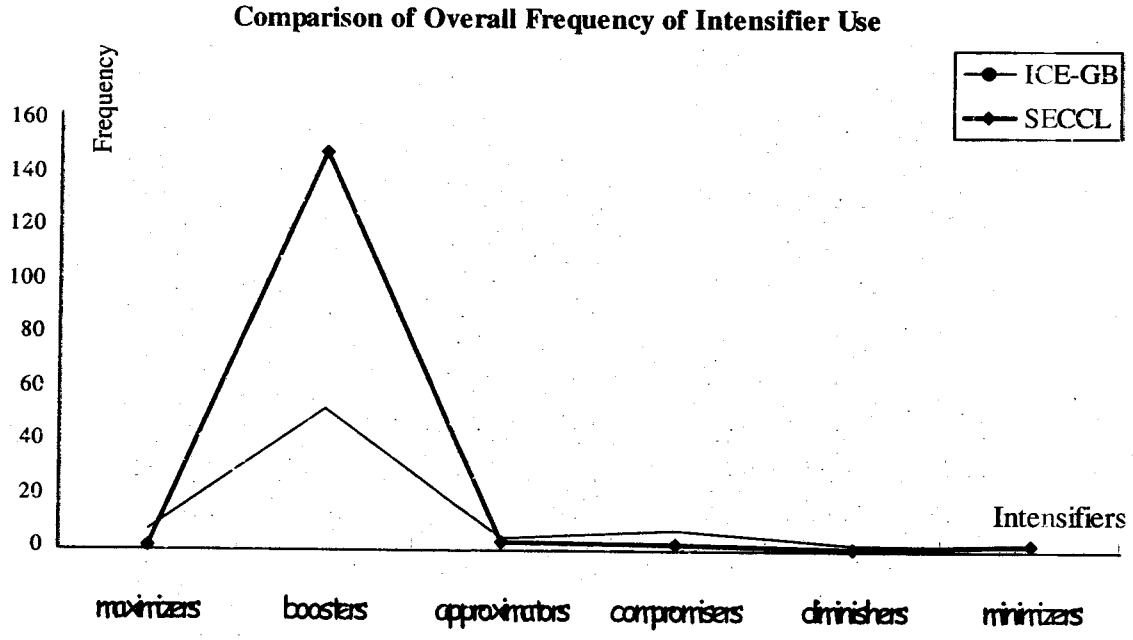


Figure 1: Overall Frequency of Intensifier Use

The difference in the overall frequency of intensifier use in the two corpora (See also Figure 1) clearly shows the tendency of overstatement of Chinese EFL learners in their speech production. The overuse of amplifiers, boosters in particular, greatly reduces the idiomaticity of the learners' spoken language.

Amplifiers		ICE-GB		SECCL	
		N	Frequency	N	Frequency
Maximizers	absolutely	113	1.99	2	0.05
	completely	76	1.27	4	0.10
	totally	78	1.30	13	0.31
	perfectly	48	0.80	0	0
	entirely	51	0.85	1	0.02
	fully	28	0.47	4	0.10
	Sum	394	6.68	24	0.58
Boosters	very	1952	32.53	4816	**114.67**
	so	467	7.78	971	**23.12**
	really	278	4.63	214	5.10
	too	301	5.02	209	4.98
	extremely	63	1.05	8	0.19
	highly	26	0.43	0	0
	terribly	28	0.47	0	0
	Sum	3115	51.91	6218	148.06
Total		3509	58.59	6242	148.64

Table 2. Comparison of Amplifier Use

Table 2 shows the difference in the frequency of amplifier use in the two corpora. Analysis of the statistics in the table reveals that among the amplifiers investigated, as with the case of downtoners, there is an obvious tendency of underuse of maximizers by Chinese EFL learners. In other words, the overuse of Chinese EFL learners' intensifiers is solely attributable to the overuse of boosters. As shown in Table 2, among the 7 boosters investigated, the greatest difference in the frequency of booster use is found in the use of two boosters only, *very* and *so* namely. The following is one of the transcribed texts taken from SECCL.

> I am a college student and my home is *very* far from the college, so I have to live in a dormitory. And one of my dormmates is *very*, a *very* energetic girl. She is always *very* active in daytime. Meanwhile. She is also *very* excited in the night. Almost every day, I sleep with her voice. Oh not voice. It's even noisy because she wants to express her ideas and her excitement in her dreams, it's *very* terrible for me, because I want to sleep *very* well, but she always in the middle night suddenly made a terrible sound. So I always awaked or disturbed by hers dream. I remembered one night I just fall asleep a few moment. Suddenly she made a big voice and then silent and then she sang song she sang a song and it's an English song, although it's *very* beautiful, but you know it's in the middle night midnight, so I feel *very* terrible, so I trembled in the bed and after that I feel *very very* threatened, so I made a noisy a noise too and wanted to wake her up and and then I will go to sleep again. But it's *very* it's *very* difficult for me to wake her up because she is in her dream and *very* deeply. So it makes me *very* angry. Every night I always find that in the morning that my eyes were *very* were *very* poor, so I want I always want to say to her that you are *very* active, too active to let me go sleep. So please made your dream silently.

Within this short text of 269 words, the booster *very* is used 18 times, with an absolutely frequency of 6.69%, more or less the same with that of the English definite article *the* as found in any of the major native English corpora. This should have reminded many of the readers of this paper who have ever taught Chinese EFL learners, whose spoken English is stuffed with numerous cases of *very*. We have every reason to assume that many of these cases, as shown in the above extract, are not necessary. It seems that the word *very*, which has a clearly definable lexical meaning, has been deprived of much of its lexical meaning in

Chinese EFL learners' speech production.

4.2 Accuracy of intensification

A comparison between the two corpora also indicates clearly that Chinese EFL learners often cannot use some of the intensifiers with good accuracy. This is not only seen in the overstatement that results from their overuse of amplifiers, boosters in particular, but also in their inadequate use of downtoners.

Downtoners		ICE-GB		SECCL	
		N	Frequency	N	Frequency
Approximators	nearly	65	1.08	39	0.93
	almost	157	2.62	70	1.67
	Sum	222	3.70	109	2.60
Compromisers	pretty	104	1.73	7	0.17
	rather	199	3.32	54	1.29
	fairly	99	1.65	4	0.10
	Sum	402	6.70	65	1.56
Diminishers	slightly	89	1.48	1	0.02
	somewhat	19	0.32	9	0.21
	Sum	108	1.80	10	0.23
Maximizers	hardly	35	0.58	63	1.50
	scarcely	6	0.1	0	0
	Sum	41	0.68	63	1.50
Total		773	12.88	247	5.89

Table 3: Comparison of downtoner use

Statistics in Table 3 show that Chinese EFL learners' use of downtoners is not even half as much as that of English native adults. We believe we are justified to assume that the significant underuse of downtoners will lead to the inaccuracy of the learners' expressions. It appears that Chinese EFL learners are at their best in their use of approximators. This may be an evidence of L1 transfer, as both words in this sub-category have ready equivalents in Chinese. Another possible reason for this may be that these two words are taught and learned earlier, and the learners have achieved automaticity in their use of these two words. It is also interesting to note that the relatively less used diminisher *somewhat* is also used with roughly equal frequency by the learners. Although this word enters the learners' active memory much later than most of the other downtoners, and automaticity is less likely for many of the learners, it may be L1 transfer that has led to the favorable result.

Apart from the downtoners mentioned above, most other downtoners are used much less commonly in SECCL, and an unavoidable result for this underuse is the semantic inaccuracy of expression.

Another kind of easily recognizable inaccuracy is found in the modification of non-gradable adjectives. It is well known that intensifiers cannot proceed non-gradable adjectives as modifiers (Biber et al, 1999: 521). Compare

1) The test will be fairly *easy*.
2) *The car was very *motionless*.

The adjective *easy* in sentence 1) is a gradable one, rendering it possible to modify it with the intensifier *fairly*, while the adjective *motionless* in sentence 2) is non-gradable, so that it cannot be modified by the intensifier *very*.

The concordancing of some fairly common non-gradable adjectives in SECCL turns out some pedagogically thought-provoking lines. The following lines are such cases extracted from concordance lines of *excellent*.

N	Concordance		
1	with us. They think you refuse the better	excellent	students. B: er errI think the
2	etter education to them. If they are more	excellent	...err...after...the study time, it
3	lity ... specialties (...) females are more	excellent	than ... males. And ... err ... f
4	ents, and...this proved that girl are more	excellent	than boys. At least, the girls
5	t that ... the females are ... Um ... more	excellent	the males, so why not...why d
6	ll I think the students around me are so	excellent	and I'm under great pleasurepr
7	, because some of the students are so	excellent	. B. Oh, you can be excellent,
8	think the students in universities are so	excellent	in their high school. But now I'
9	now ``` I ``` er``` the student here are so	excellent	that I ``` em``` B: I see as En
10	s I know, education abroad is somewhat	excellent	in some major, such as MBA,
11	... his friend all says said he is very very	excellent	absent-reminded...For one hot
12	li>only wants us our <sais> to be a very	excellent	people ...en...I think you shou
13	uipment private ... private for you is very	excellent	. S:A: Yes, I think it is more c
14	many students in the university are very	excellent	. Some of them even gain gain
15	S:B: Yes, I know you are excellent, very	excellent	student in ... the university.
16	... said his idea was very good and very	excellent	. The g ... Thethe guests surro
17	rink that is served by Mr Holm was very	excellent	. They had ... they all had goo
18	education. He must, they must be very	excellent	, umm, to have the, have the c
19	like him a lot because his class is very	excellent	. We like to hear what he say
20	at have given me lessons. They are very	excellent	, while some of them are very
21	here are so many students that are very	excellent	. Yea, You have to work hard .
22	.everyone said...the party was...er...very	excellent	...un...un...When the party was

Similar errors in SECCL are more commonly found with such non-gradables as *perfect*, *equal*, *exhausted*, *acquainted*, *alive*, *asleep*, *unique*, among a couple of others.

It must be noted that the overuse and improper choice of some intensifiers, and the misuse of intensifiers before some non-gradable adjectives have greatly sharpened the non-nativelikeness of many Chinese EFL learners' language, resulting in an easily identifiable lack of idiomaticity in their speech production.

4.3 The insufficiency of intensifier repertoire

It goes without saying that in most cases, each of the English intensifiers has its unique shade of lexical meaning, and that there is always a best choice in a certain semantic and syntactic context. However, as shown in the above tables, many of the fairly common intensifiers are not given their due status, and have been markedly underused. Since it is less likely that the learners have not had the need to use such intensifiers, which otherwise are much more needed by native speakers, the conclusion can be safely drawn that many Chinese learners have not achieved automaticity in the use of these words, and that the insufficiency of intensifier repertoire among the learners renders it necessary for them to struggle for better and comprehensible ways to express their own ideas.

It is interesting to note how Chinese learners manage their communication tasks without the convenience of retrieving the most suitable intensifiers from their memory. As noted earlier in this paper, the booster word *very* is used with an exceptionally high frequency in SECCL. It seems that this word has become an unmarked term for intensification. In cases where a maximizer is optimal, learners tend to go one step

down along the intensifier scale and assort to the booster word *very*. As they are well aware that *very* does not warrant them their intended meaning, this booster word is then used twice, or even three times in succession, to achieve their desired effect. On the other hand, when they feel a need to use a compromiser, which is not readily retrievable, they tend to use the booster word *very* again, but this time they try to downtone the booster slightly by adding the negative adverb *not*, so that *not very* results. Table 4 shows how common the mentioned bi-grams are in Chinese EFL learners' speech production.

Bi-gram	ICE-GB		SECCL	
	N	Frequency	N	Frequency
very very	98	1.63	140	3.33
not very	34	0.57	203	4.83

Table 4. The Use of *very very* and *not very*

5. Conclusion

Chinese EFL learners tend to overuse intensifiers in their speech production, and this overuse is primarily attributed to the overuse of boosters, leading to a tendency of overstatement and a lack of accuracy. Among other boosters, *very* is probably the earliest-learnt one, which is always easily retrievable from memory. For this reason, and in their struggle for better communication, Chinese learners often either go one step up to use *very very* or one step down to use *not very* in their attempt to amplify and downtone their intensification. While this effort does help to achieve communication goals, the overuse of boosters and the underuse of most other intensifiers significantly strengthen the non-nativelikeness and the lack of idiomaticity of their language production.

Learners need to be made aware that overstatement is not always necessary, and that proper intensification can be achieved only by the use of different intensifiers.

References

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

Lorenz, G. (1998) *Overstatement in advanced learners' writing: stylistic aspects of adjective intensification*, in Granger, S. (eds) 1998: 53-66

Granger, S. (1998) *Learner English on Computer.* (eds) Longman, Addison Wesley

Biber, D., Johansson, S., Leech, G., and Finegan, E. (1999) *Longman Grammar of Spoken and Written English.* Beijing, Foreign Language Education and Research Press

Nelson, G. (1998) *The International Corpus of English: mark-up for spoken language*, in Leech, G., Myers, G., and Thomas, J. (1995) *Spoken English on Computer.* (eds) Longman, New York

Problems and Coping Strategies of Speech Data Collection: Insights from a Special-purpose Corpus of Situated Adolescent Speech

XU Jiajin
Beijing Foreign Studies University

Abstract: This paper is concerned with five problems in speech data collection. Drawing on the work with the Corpus of Situated Adolescent Speech, we propose some tentative coping strategies to solve the five problems. Our governing principle is that we should give credit to the most natural and rich language. In the meanwhile, the relationship between data and theory is discussed.

Key words: Speech data collection; problems; coping strategies

1. Preliminary considerations

Linguists have to be keenly aware that single text-based linguistic research is distanced from real life situated discourse, so is quasi- or prepared speech. As Labov (1972) so aptly discussed, linguistic science is rooted in the efforts of the bush linguists and street linguists and only secondarily advanced by those who do most of their work in the library, their offices and their laboratories. Actually to put this differently, there is no such thing as "the ideal speaker/listener in a completely homogeneous speech community" (Chomsky 1965). What corpus linguists do is to find order from heterogeneity of everyday language. The situated speech data provide samples of naturalistic discourse instead of those data under experiment conditions or during interviews. Now with audio recording, we are capable of carrying out research into phonetic and prosodic nature of language. In addition to studies on lexical and syntactic levels, meaning in interaction, viz. pragmatic and discourse analysis can be conducted within social and textual contexts. In this paper, we adopt two general principles regarding the development of special-purpose corpus of situated speech: the Naturalness Principle, and the Richness Principle (Gu forthcoming).

Then we would like to clear the ground by defining what are desirable spoken corpus data. According to Williams (1996), the ideal spoken corpus should include all forms of speech, from diverse speakers and covering various styles and accents. The recordings should be orthographically transcribed, grammatically tagged, and prosodically annotated. Finally, the corpus would be very large. His criteria are actually set for the general-purpose spoken corpus like LLC and the spoken part of BNC. For many linguists who cannot obtain adequate funding, such a corpus is much too utopian. Our understanding towards desirable spoken corpus for small-scale specialized research is that spoken corpus should "mimic" the general composition of the general corpus within the specific registers or genres of speech. A spoken corpus thus complied would be functional and operational for linguistic research in its own right.

2. From Speech Data to Theory

Observable speech data do not advance scientific understanding of discourse (Chafe 1994: 15), but speech data *per se*, speakers and/or linguists, research objectives, personal and situational information and elicitation in collecting data do capture the interplay between speech data and linguistic theorizing. In the current paper, the five problems in speech data collection and their coping strategies are addressed to explore such relationship between data and theory. Certainly they are suggestive and do not claim to be exhaustive.

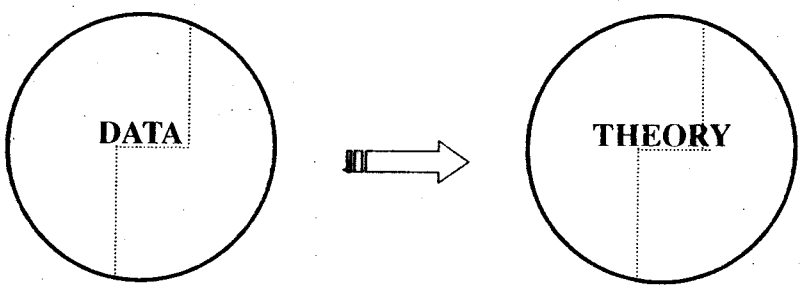


Figure 1: From data to theory

Linguists first of all must ensure what they have collected is the same thing as the speakers' daily conversation. That is the data are as objective as the speakers produce when they are not observed. Anyone who is aware that his behavior is to be accessible to the public, he would be either unhappy or unwilling to speak more. Sometimes the obvious change can be noticed with the increase of the number of people present or potentially present. For instance, students speak rather at will with parents at home and much hesitantly in class and even speak really bad language with their peers. With sophisticated recording equipment we can now have such behavioral observations stored and make them retrievable for linguistic research.

Corpus linguists believe hard evidence from corpora, however, in order to rule out the threats to the objectivity of corpus data, the linguistic fieldworker has to ask 1) what role he plays in the development of linguistic theory? 2) what sort of data do we need? 3) does linguistic expertise have a niche in the creation of a spoken corpus? 4) how does fieldwork methodology affect the data?

3. Problems and Coping Strategies of Speech Data Collection

Once the general construct of a spoken corpus is determined, it is time to get on the fieldwork stage of corpus creation—how to collect speech data. This paper will look at some important variables in collecting speech data that dictate the overall quality of the corpus.

3.1 Problem 1: Recorder's paradox

Following Labov's *observer's paradox*, we coin this *recorder's paradox*. To explain this change of addressing the difficulty in obtaining real life speech data, we are arriving at a critical issue in speech data collection. Namely, the exposure of the intent of recording will inevitably sensitize speaker's awareness in their ways of speaking in varying degrees. The data collected are thus invalid for rigorous theoretical investigation.

Normally in sociolinguistic interviews, it is almost impossible for researchers to be impartial observers of linguistic facts (Schilling-Estes 2000). Even if the researchers do not find themselves self-conscious of their research purposes, informants would switch to a, say, formalized way of speaking. In such sociolinguistic interviews the major discourse types are questions and answers. Speech data thus gained will not represent the everyday language of informants. To ensure most natural recordings possible, we maintain that the revelation of recorder's purpose is to be made known after the conversation.

In the work with the Corpus of Situated Adolescent Speech, different recording personnel are involved. We enter the speech community ourselves to play the role as an onlooker, or a participant sometimes, and most often we recruit junior high school students to record their talk with fellow students before class, after class, in the teacher's office, at the bus stop, and on the way home. We also asked some adults to record family

discourse with their teenage children, like dinner table talk. These recruits are found to be unexpectedly cooperative, recorded talk is not different (as some recorders later claimed) from their natural conversation with teenagers. Sometimes, to extract as much natural speech as possible, we erase the initial section of the conversation recorded (ten minutes are believed to be a good cutting point), because the recorder (especially the teenage recruit) is more or less hesitant to speak more or speak in a controlled way.

One important thing should be borne in mind that ethical issue arises when we do the recording surreptitiously. Although it is fortunate for Chinese linguists that this is not a very big issue at present only if we keep the personal speech data among the academics and use them for research only, we would rather not intrude on others' private spaces. Two suggested solutions in this case are 1) record linguists' family or his close relatives' family talk if they do not mind at all; 2) we ask some potential informants for recording their everyday talk at any time if they are kind enough. Our recording, however, begins at any unknown time. In these two cases, we can get natural speech data. The first method uses himself (the most reliable data supplier) as informant or takes advantage of his solidarity with his relatives to get natural data. This method has been adopted by many linguists on their children for example. The second method is another acceptable compromise to get natural data.

No matter the researcher or a recruit does the recording, his attempt to record the talk should always be kept to himself before and during the talk, or he will not expect any normal conversation any more. In other words, the recorder has to be an "invisible" data collector, and he should always have a ready mind (and ready recording equipment as well) to record those uninformed speakers.

3.2 Problem 2 How and to what extent should situational information be kept?

To assist functional analysis out of the speech data, situational information ought to be kept as much as possible. Therefore a detailed log keeping is required for every piece of recording. As we all know, recording and transcription result in a loss of information that is otherwise available to the actual situations of the discourse. This explains why transcripts of spoken discourse are very often incomprehensible to outsider readers. Moreover, situated discourse, as part and parcel of the ever-resolving social process, goes out of date very quickly, and future users of the corpus will fail to see the social significance if the information is not sufficiently provided (Gu forthcoming).

As I mentioned previously, corpora are complied for future interpretation, and in the meanwhile any interpretation of linguistic data requires a context in time and space. Linguists will make sure against the situational information provided whether the discourse "slice" recorded is affected by other people present (teachers, research, or peers) or truly happens as it is.

To acquire information about the field site, we have to do much preliminary work. In our case, we need to study the floor map. Sometimes we should establish certain rapport with the students (on a practical basis we actually first find teachers to whom we have some connections). These will enable us and "latecomers" more at ease to get adjusted to the field situation.

3.3 Problem 3: How and to what extent should personal information of speaker be kept?

This problem is closely related to the preceding one. In this case, the demographical information and role relationship of participants in the speech interaction should be jotted down as much as possible for future examination.

Speakers in the situated discourse shape the speech data in their particular way. Their identity or role

relationship makes a significant difference of their talk in the small speech community. We know teacher talks like a teacher, and student talks like a student. A boy student talks also different from girls. A mischievous student speaks even more different from others.'

An on-the-spot log keeping of speakers' demographic information is badly needed for future research (if we have access to it at all). Unfortunately in anonymous observations or invisible recordings, personal information is impossible to get. At this time, we need to at least take down our rough estimation of the speakers' age, role relationship with other teens and so forth.

3.4 Problem 4: Is preset linguistic motivation for collecting speech data justifiable?

The fourth problem goes whether the sampling and collecting of the target speech data is to be theoretically motivated. A special-purpose corpus compilation is usually directed to a certain research objective, because it is not economical and practical to make a small corpus all-inclusive and all-embracing.

The speech data from fieldwork will ultimately be shaped by not only the language itself but by the research goals we aim to achieve. For instance, in situated adolescent spoken corpus, we want to investigate the discourse markers from the prosodic perspective. Therefore we need to record more casual talk, instead of formal speech or sociolinguistic interview. If the purpose is on the language of urban adolescent speakers, the sampling is confined to this particular type of population.

Some people would argue that it is myopic to limit the record to the data pertinent to issues of current theoretical interests, but we have to check our recording quantity. We cannot hope to anticipate all future needs (Mithun 2001:53), theory gives us much on methodological issues, helps us find finer things to look at. This problem again points to our discussion of the relationship between data and theory. It is not appropriate to say that we set the theoretical framework for natural data to fit it; it is economical in actual field research to include a general theoretical orientation of data collection.

Linguistics benefits when fieldworkers are doing more than merely gathering data for a theoretician to interpret (Everett forthcoming). We understand Everett as meaning linguistic theory modifies our corpus planning, narrows our categories of samples.

By linguistic motivation, generally we mean given the funding and energy we have, what priority should be given to certain genre or register of discourse. As in the Corpus of Situated Adolescent Speech, if our object of investigation is on phonetic and/or phonological aspects of discourse, we need to find less noisy settings so as to obtain higher quality audio recording.

In a sense, the identity of a corpus is shaped before it actually comes into being. A corpus is by its very nature a purpose-built linguistic databank.

3.5 Problem 5: Does elicitation have a role to play in accumulating data?

Generally, sociolinguistic interview does not present a true picture of natural speech interaction. However, some researchers (like Labov and Schegloff 1989) argue that well-devised interviews can also represent talk in action. But we hold that naturally occurring speech is the sole representation of human speech. Speech data out of interview can only be used for stylistic and/or variational comparison.

We made some recordings of interview for comparative study. Sometimes in all of our data gathered, we can hardly find instances of some intuitively very frequent linguistic facts. In such cases, well-devised elicitation also has a role to play.

4. Conclusion

Most of the problems are revisited here in the work with the Corpus of Situated Adolescent Speech. Here in this paper we just present very briefly some piratical guidelines for linguistic fieldwork especially for speech data collection, which actually requires a book length work to cover.

Actually many other important issues like the overall and sample size, time frame, sociolinguistic variables (e.g. gender, age, literacy etc) should be considered to create a valid corpus. But these issues have already been covered in many corpus linguistics monographs. The problems addressed in the present paper are small but significant for the quality of the corpus data collection and the ensuing theorizing. We hope that what we are presenting here is useful as analytical and practical tools.

References

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago and London: University of Chicago Press.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Everett, Daniel. Forthcoming. Coherent Fieldwork. Paper presented at XVII International Congress of Linguists, Prague.

Gu, Yueguo. Forthcoming. *Segmenting and Annotating Situated Discourse: With Special Reference to Spoken Chinese Corpus of Situated Discourse*. London: Routledge.

Labov, William. 1972. Some Principles of Linguistic Methodology. *Language in Society* 1:97-120.

Mithun, Marianne. 2001. Who Shapes the Record: The Speaker and the Linguist. In Newman, Paul and Martha Ratliff (eds). 2001. *Linguistic Fieldwork*. Cambridge: Cambridge University Press.

Schegloff, Emanuel A. 1989. Survey Interviews as Talk-in-Interaction. In D. W. Maynard, H. Houtkoop, N. C. Schaeffer and H. van der Zouwen (eds.) *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*. New York: John Wiley.

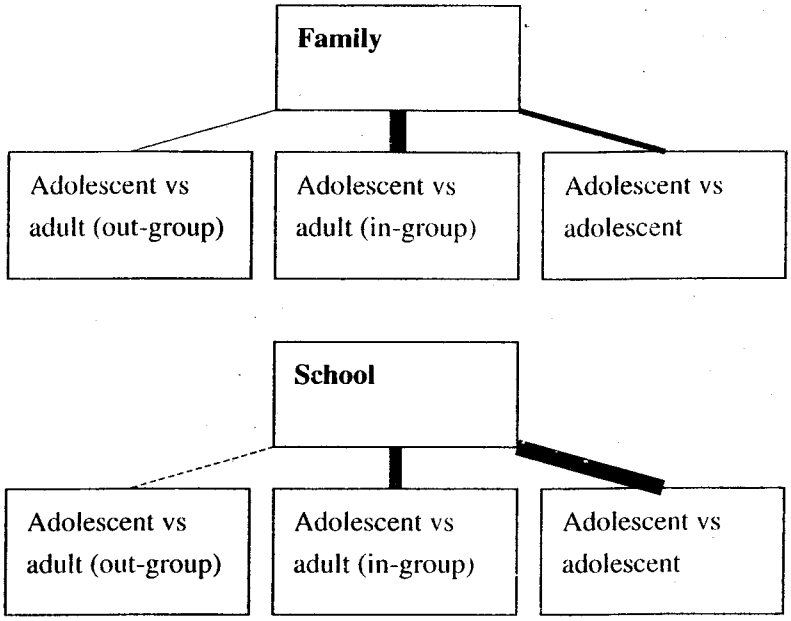
Schilling-Estes, Natalie. 2000. Introduction to "Fieldwork for the New Century: Papers from the SECOL 1999 Panel Presentation". *Southern Journal of Linguistics* 24:83-90.

Williams, Briony. 1996. The Status of Corpora as Linguistic Data. In Knowles, Gerry, Anne Wichmann and Peter Alderson (eds). 1996. *Working with Speech: Perspectives on Research into the Lancaster/IBM Spoken English Corpus*. London and New York: Longman.

Appendix 1

Corpus of Situated Adolescent Speech mentioned in the paper is started in January 2003 and still under construction. The projected size is about 20 hours surreptitiously recorded spontaneous conversation of adolescents. The recordings are made by the researcher himself and several recruits. The corpus will be orthographically transcribed and grammatically, prosodically, and probably pragmatically annotated. And all these annotations are converted into the codes readable by the software -- Codingstar. With the software, we tag the plain text with the codes, and the tagged text is then exported in XML format.

Appendix 2: Preliminary sampling strategies and procedures of Corpus of Situated Adolescent Speech



SCHOOL-BASED/RELATED *PERIPHERAL* TALKING-DOING INSTANCES

Picnic/visiting museum/seeing a movie/voluntary work...

FAMILY-BASED/RELATED *PERIPHERAL* TALKING-DOING INSTANCES

Shopping/visiting relatives...

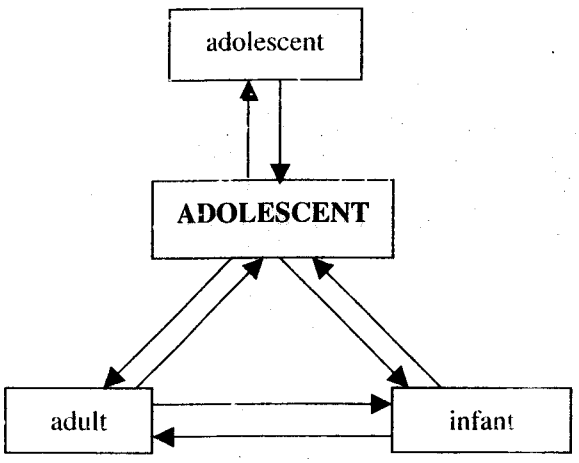
THE CHARACTERISTIC TALKING CONTEXTS OF ADOLESCENCE

1) Families > 5 hrs
2) Peer groups > 5 hrs
3) School: work > 5 hrs
4) School: leisure > 5 hrs

CATEGORIZATION OF ADOLESCENT DISCOURSE IN TERMS OF PARTICIPANTS

Peer groups
- *Among boys*
- *Among girls*
- *Mixed*

Adolescent vs adult (out-group)
Adolescent vs adult (in-group)
Adolescent vs infant (out-group)
Adolescent vs infant (in-group)
Monologue



Adult-initiated adolescent-directed speech
Adolescent-initiated adolescent-directed speech
Adolescent-initiated adult-directed speech

A Survey of Lexicalization of Causative Verb Structures in the CLEC

Zhang Jidong
Donghua University

Liu Ping
Shanghai Jiaotong University

Abstract: This paper is concerned with the extraction and analysis of the causative structures of "make + O + C" format and their lexicalization. The CLEC has provided us with a substantial evidence that the Chinese college students are inclined to over-use the causative verb structure rather than their possible counterparts of the lexicalized verbs. This language phenomenon can be possibly explained from the standpoints of 1) human beings' schemata specific to the two categories of expressions, namely, synthetic and analytic expression; 2) possible learning strategy of simplification (or avoidance). 3) Language transfer from Chinese causative structures; 4) morphological transformation of the causative verb structure in Chinese and English language. From the study, we can also make an inference about how well the learners have mastered the productive vocabulary of lexicalized causative verbs and how the learners access to their habitual patterns of the language learning.

Key Words: CLEC, causative verb, causative 'make + O + C' structure, lexicalization

1. Introduction

College English Test, as a China's most authoritative standardized test, has been successfully administered nationwide in China for seventeen years on ends. And it is a well-accepted fact that with the motivation of the test, the college English learners have improved tremendously in their comprehensive language proficiency. However, we can not afford to neglect the fact that in the CET candidates' writings there still exists the problem concerning the insufficiency of the students' productive vocabulary, specifically the causative verbs. In the Chinese College Learner English Corpus (CLEC), what the writer has found about over-consumption of the 'causative verb structures' on the part of CET candidates are somewhat indicative of the college students' underdevelopment in their language production. In the CLEC, this language phenomenon is significantly evidenced by the overuse of 'make + object + complement' structures (hereinafter referred to as 'make + O + C' structure). Why are the Chinese college students more inclined to use such a structure in their CET compositions? The reason for the underdevelopment in such language use and understanding could be explained in numerous ways. It might be ascribed to the language learning features specific to language learners. Such language phenomena are somewhat suggestive of how Chinese college students develop English language from the norm of the source language to that of the target language. As regards the over-consumption of 'make + O + C' structures on the part of CET candidates, it may reflect the fact that the latent schemata of Chinese thinking are more preferable to them especially in terms of causative expressions. In addition, it has revealed that the problems in vocabulary teaching have not been approached in great depth in our methodology research. In English language learning and teaching, learners are simply urged to build up a large repertory of vocabulary in isolation. Usually these words are mainly notional words, especially nouns. As for verbs, students are quite fresh about the lexicalization, which refers to the most effective and economical way of expressing their ideas. Such a treatment can be

problematic, as a result of which so much is lost concerning the activation of the target language words. By extraction of the causative verb structures and their possible lexicalized verb forms, the writer found that the language structures had frequented the CET compositions and had partially affected the rating. Therefore, what we should do is to discover what is really behind the Chinese EFL learner's language unusualness and to understand their underlying learning strategies in their language development.

2. Research Instrument

Since this paper is aimed at conducting an investigation into the causative verb structures involved in the CLEC, the writer adopted some feasible procedures in carrying out the research. First, s/he tried to take advantage of the College English Learner Corpus (Li Wenzhong & Pu Jianzhong 1998) and some 49 CET composition range-finders for the latest two tests. Second, the writer tried the instrument of MicroConcord software package programmed by Michael Scott et al for extracting all the causative verb structures in the corpus and the collected data. The first two steps actually serve as a preparation for the later coming analysis on the findings from the CET writing corpus. Third, the writer tried to find out all the possible lexicalized causitive verbs with reference to the vocabulary list in College English Syllabus (for students of Arts and Sciences). As for the last step, the author tries to make a semantic analysis about the valuable findings and some remedies and suggestions will be presented accordingly. In the CLEC, we find that the CET candidates were more inclined to use 'make + O + C' structures instead of their counterparts of lexicalized verbs. The findings are somewhat suggestive of the status quo of our Chinese students' language proficiency.

3. Lexicalization of Causative 'Make + O + C' Structure in the CLEC

3.1 Extraction of Causative Verbs and Causative 'Make + O + C' Structure from the CLEC

In the CLEC, it has been searched out that our college students are inclined to overuse the 'make + object + complement' structures rather than their counterparts of 'compressed' causative verbs, say, like verbs with 'en-', '-en' or '-ify' as their prefixes or suffixes. In order to make reliable what the writer has postulated about the shortage of 'compressed' causative verbs involved in the CLEC, the writer tried to make a list of possible causative verbs on the basis of the wordlist of the College English Syllabus for non-English majors (2000). Then the writer managed to run the MicroConcord software for searching all the possible 950 causative verbs in the corpus. According to the statistical findings of both frequency and percentage, most of the causative verbs are not prominent in the distribution of occurrences, except for the verb 'adapt', 'advance', 'benefit', 'change', 'concentrate', 'decrease', 'improve', 'develop', 'decrease' and 'limit'. This number is far from satisfactory in terms of the causative verbs required of the CET candidates.

Table 3-1: Causitive Verbs from the CES & Their Frequency in the CLEC

No.	Causative Verb	Frequency	Percentage	No.	Causative Verb	Frequency	Percentage
1	accelerate	6		36	ensure	6	
2	accustom	6		37	expand	10	
3	activate	19		38	force	15	
4	adapt	90	0.03	39	fulfill	4	
5	adjust	20		40	gain	240	0.07
6	advance	57	0.02	41	impress	5	
7	advise	13		42	improve	444	0.14
8	affect	20		43	increase	423	0.13
9	aiter	18		44	keep	241	0.08

10	allow	5		45	kill	13	
11	appeal	5		46	limit	170	0.05
12	apply	57	0.02	47	lower	45	0.01
13	assure	5		48	occupy	7	
14	benefit	58	0.02	49	pollute	292	0.09
15	change	995	0.31	50	perfect	8	
16	charge	6		51	prohibit	4	
17	concentrate	23		52	promote	13	
18	control	27		53	protect	106	0.03
19	cut	37	0.01	54	punish	8	
20	decrease	213	0.07	55	purify	11	
21	determine	23		56	puzzle	4	
22	develop	1338	0.42	57	qualify	9	
23	devote	73	0.02	58	raise	58	0.02
24	disappoint	6		59	realize	143	0.04
25	discourage	7		60	recycle	17	
26	disturb	4		61	reform	11	
27	divide	5		62	relieve	9	
28	drop	103	0.03	63	renew	7	
29	enable	12		64	satisfy	53	0.02
30	encourage	18		65	stimulate	7	
31	end	140	0.04	66	strengthen	12	
32	engage	51	0.02	67	survive	33	
33	enhance	10		68	threaten	7	
34	enlarge	4		69	transform	5	
35	enrich	4		70	widen	8	

From the statistics based on the concordance lines, the writer found that about 453 cases of 'make + O + C' structures had been involved in the CLEC composition sample. Among the 453 cases, about 260 cases of 'make + O + C' structures had been sorted out, including 96 cases of 'make + n. + adj' and 164 cases of 'make + n. + infinitive / gerund' respectively. The cases of the 'make + O + C' structures make up about 57.3% of the total. This statistical result shows the fact that 'make + O + C' structures had been used prominently in the CLEC. The high frequency of such a structure may indicate the statistical significance, compared with the total number of all the sentences involved in the CLEC. (Refer to the Appendix)

3.2 Synthetic and Analytic Expressions and Lexicalization

In the process of analyzing the lexicalization of the 'make + O + C' structure, the writer found that the CET candidates are more inclined to employ these Chinese-specific structures than their counterparts of compressed causitive verbs. There are 260 cases of the 'make + O + C' structures involved in the CLEC. The number of such language cases are suggestive to some extent of how the CET candidates seek out the linguistic structures most salient or accessible to them in order to ensure the expressiveness of their ideas. But why are such structures more preferable to Chinese college students in their writing production? The writer manages to approach the language phenomena from the viewpoint of synthetic and analytic lexical expression.

Languages are usually characterized by two categories of expressions, namely, synthetic expression and analytic expression. If a complex given concept can be expressed in terms of a lexical item, then it refers to

synthetic expression. If it needs a phrase to realize the same semantic result, it means analytic expression. The synthetic expression can be considered as lexicalized expression, which is usually restricted by the degree of the language morphological development. The synthetic expression is usually achieved by means of derivatives and compounds. However, besides derivatives and compounds, there still exist many exceptional cases to reach the synthetic expression. Language still occupies many morphologically individual words which should also be labeled as the same belongings. For instance, English verb 'stink' is an individual word, which connotes the meaning of 'give a strong bad smell'. So is the case with the expression of 'polluted water' as 'sewage'.

As far as the synthetic and analytic expressions are concerned, almost every language is equipped simultaneously with these two different modes of expressions. The co-existence of the two expressions can be justified by some examples. In English we have 'to foul' and 'to make... dirty', 'to cause... to become dirty', etc. In Chinese, we have "使...生气" (make sb. feel angry with) and "惹恼"(annoy), "使...害怕"(make sb. fear) and "吓唬"(frighten), etc. However, the problems lie in the fact that our Chinese learners can not be encouraged to over-consume 'make + O + C' structures, as a result of which it will result in monotonous formulaic expressions in English.

3.3 Explanations of the Language Phenomenon from Psycho-linguistic perspective

As we know, the meanings embodied in the lexicon or grammar of language determine the thought patterns of the language users. The charge to be semantically expressive is a charge to language from thought. Compared with English language, the schemata of causative relations in Chinese language are supposed to be lexically overt, e.g., "他让孩子安静下来" (he made the children to be quiet), "他所做的一切使大家伙高兴" (what he did made the crowd happy), "他的行为让他的朋友失望" (his deed caused his friends to be disappointed). Such are very typical Chinese structures as "使..." and "让..." structures. Therefore this results can be understood in terms of mother tongue transfer. Second, from the psycho-semantics point of view, the externalization of this structure on the part of Chinese students happens to be identical with the internal schema of native speakers. Therefore, logically the internal schema is the starting point of materializing the causative structure from the deep structure to the surface structure of language form. This means that the native speakers of English would rather resort to the following expressions: 'he quieted the children', 'he pleased the crowd', and 'his deed disappointed his friends'. Third, under the influence of time, peer or test pressure, the CET candidates possibly adopt the strategy of simplification, i.e., the learning strategy of 'playing safe'. As we know, the easier structures do not cause any difficulty on the part of CET candidates in the process of recognition and utilization. Therefore, the test-takers are motivated to use them to achieve their expressiveness in their compositions. But here we should note that 'playing safe' in language production can lead to few errors. Instead, it will result in under-presentation of words or structures in language use. This strategy occurs when a test-taker thinks that certain features of a language likely cause him difficulties, and consequently he tends to avoid these language features. We can not, therefore, conclude much about what the testee knows and does not know simply from the analysis of his or her language performance. This can mean the limitation of the research.

3.4 Explanations from Comparing Chinese and English Derivatives of Word Building

In language research, the most effective research method is usually comparison. Language phenomena can be made clear by means of making comparison between the language to be learnt and the native language. Therefore, in the following section, the author tends to make some distinctions between the causative

structures of L1 and TL. In English, it is quite clear that there are two kinds of synthetic expressions for the identification of 'causative relations'. One is about the derivatives with the affixes of prefix 'en-', of suffixes '-en', '-fy (-fy)', and '-ize (-ise)', etc. Actually it is the derivation of word building that enriches the English language. Some of the synthetic derivatives have their counterparts in Chinese, but some have no equivalents.

Table 3-2: Derivatives of Causative Verbs & Chinese Equivalents

en-	enrich enable	使丰富；丰富（文化生活等） 使能够
-en	frighten soften	使吃惊，使害怕；吓唬（某人） 使变软，使变柔和
-fy	beautify intensify	使变美丽，美化（环境等） 加强
-ize	mobilize modernize	动员 使现代化

Table 3-3: Sampled Causative Verbs & Their Frequency in the CLEC

VERB	Fre.	VERB	Fre.	VERB	Fre.
annoy	2	disappoint	5	modernize	2
assure	3	discourage	5	please	4
attract	9	distress	Φ	purify	6
beautify	Φ	enable	12	puzzle	6
better	Φ	encourage	18	quicken	Φ
bore	6	enhance	10	relax	1
comfort	Φ	enlarge	20	relieve	9
compel	Φ	enrich	14	soften	Φ
confuse	1	fulfil	4	stabilize	Φ
convince	1	heighten	Φ	suffice	Φ
cost	2	improve	444	torture	Φ
damage	3	industrialize	2	trouble	Φ
darken	Φ	legalize	Φ	weaken	Φ
deepen	2	lighten	Φ	worsen	1

From the above table, we can make a generalization about how the Chinese causative verbs are composed. There are three methods of lexicalizing the causative lexical items. First, the '使 + 宾语 + 形容词' (make + O + C) structure is needed as a direct compensation of causative verbs. Second, the method of morphological transformation is asked to achieve the realization of the causative relations. For example, '使丰富' (enrich) is transformed into '丰富'. And some Chinese adjectives and nouns can be attached directly with the suffix '化'. For example, '使变美丽' is transformed into '美化' (beautify). So are the cases with nouns '工业化' (industrialize), '绿化', and '欧化', etc. Third, the word formation of compounding can also fulfil such lexical realizations. We can connect some verbs with adjectives together. For instance, '使变得强一些' (make sth. stronger) can be transformed into a compound word '加强'(strengthen). It is true of the compounding methods with the following words '提高'(raise), '加深' (deepen), '减轻' (relieve, mitigate), etc. In Chinese, these compound words are usually confined in the adjectives with affective meaning like '激怒' (enrage) or which can be measurable with semantic differential*. Besides what we have found about the analytic methods of expressing the causative relations, there are some synthetic expressions both in English and Chinese.

Table 3-4: *Synthetic Expressions of Causative Verbs in English & Possible Chinese Equivalents*

1. annoy	1. 使生气；惹恼
2. bore	2. 使厌烦；ф
3. disappoint	3. 使失望；ф
4. excite	4. 使兴奋；刺激
5. move	5. 使感动；感动
6. shock	6. 使震惊；震惊
7. vex	7. 使烦恼；ф

From the comparison of causative verbs between Chinese and English, we are motivated to conclude that Chinese language learners prefer the structure of 'make + O + C' in expressing the causative relations. And this is evidenced by the statistics of the causative verbs from the CLEC. Therefore, this structure is, to some extent, reflective of the linguistic features of the morphology of Chinese language. Hence we can say that Chinese language is characterized by the under-development of the lexicalization of causative verbs. Therefore such language cases on the part of CET candidates are not suggestive of students' deviations from English language norm, but rather are reflective of the students' underdevelopment in terms of lexicalized causative verbs and the results of mother tongue transfer.

3. Conclusions

3. 1 Some Inferences from the Research Findings

This paper is concerned with the extraction and analysis of the causative structures of "make + O + C" format and their lexicalization. The CLEC has provided us with a substantial evidence that the Chinese college students are inclined to over-use the causative verb structure rather than their possible counterparts of the lexicalized verbs. This language phenomenon can be possibly explained from the standpoints of 1) human beings' schemata specific to the two categories of expressions, namely, synthetic and analytic expression; 2) possible learning strategy of simplification (or avoidance). 3) Language transfer from Chinese causative structures; 4) morphological transformation of the causative verb structure in Chinese and English language. From the study, we can also make an inference about how well the learners have mastered the productive vocabulary of lexicalized causative verbs and how the learners access to their habitual patterns of the language learning.

From the perspective of language learning, the research findings shed light on some areas of problems for the learners in their vocabulary acquisition and learning. Therefore, the language teachers would be more committed to enlarging the learners' calibre of word building especially in terms of the causative verb structures and their lexicalization. As a result of such specialized training, such a word building power can be strengthened on the part of language learners.

Here I should mention two points for our reference. First, the causative 'make + O + C' structures and their lexicalized counterparts share the equivalence of information, however, they are still quite different from each other in terms of stylistic variation. One is comparatively colloquial; while the other is quite 'frozen'. Second, the lexicalized causative verbs are characterized in terms of the degree of difficulty, compared with their counterparts of the causative 'make + O + C' structures. Therefore, such a research finding is also worthy of our attention in the language instruction.

Bibliography

Leech, G. (1981). *Semantics*. Second Edition. England: Penguin Books.

Gui, Shichun (1993). A *Study of the Mental Lexicon of Chinese Learners of English*. In Pemberton and Tsang, 1993

Richards, J., Platt, J. & Weber, H.(1985). *Longman Dictionary of Applied Linguistics*. London: Longman.

Lyon, J (1977). *Semantics*. Cambridge: Cambridge University Press

Katz, J. J. & J. A. Fodor. (1963): The Structure of Semantic Theory. Englewood Cliffs: Prentice Hall.

Li, Wenzhong. (1998). An Analysis of the Lexical Words & Word Combinations in the College learner English Corpus. Unpublished Ph.D. Dissertation.

Liu, Runqing & Wedell, Martin. (1995): *Language Teaching & Learning from Theory to Practice*. Beijing: Higher Education Press.

Scott, M. & Murison-Bowie, S(1995). MicroConcord Manual, *An Introduction to the Practices & Principles of Concordancing Language Teaching*. Oxford: Oxford University Press.

Zhang, Jidong (1998): An Analysis of the Lexical Collocation Errors in CET Compositions. Unpublished M. A. Dissertation.

刘宓庆（1995）：《英汉语言对比的理论问题》：《英汉语言文化对比研究论文集》（李瑞华 主编，1996）。上海：上海外语教育出版社

刘润清、Stephen Magee (1988)：《现代语言学名著选读》。北京：测绘出版社

许云龙（编著）（1989）：《现代语言学概论》。上海：上海外语教育出版社

胡壮麟、刘润清、主编（1988）：《语言学教程》。北京：北京大学出版社

《大学英语教学大纲》词表修订工作组（2000）：《大学英语教学大纲词汇表》。上海：上海外语教育出版社/北京：高等教育出版社。

Appendixes

IV. MicroConcord Search SW: make*\made for Its Causative Structures (1)

1. helpful. Taking early morning walks makes a man healthy and wise. <No 0053>
2. benefits. Firstly, trees and grass make air cleaner. So we can enjoy fresh air
3. d lose big. Let us try our best to make all fake commodities disapeared. <no
4. The substance CH3OH in fake wine will make bright eyes dim. On the other hand,
5. cern people's health. Fake wines may make buyers blind, fake electric apparatus
6. me. So many things are to be cone to make cars popular in China. Even when the
7. can learn from many other examples. Make everything arranged. Don't like a chic
8. lt in death. Fake machines, they may make factory bankrupt. So, Fake commidites
9. work at all; a woman buy some oil to make her skin well and hurt her skin instead
10. interesting on it. or they want to make him outstanding in his job based on
11. The treatment of his disease and even make him dead. Many buildings have
12. s can bring him lots of money, which makes him very rich though it's a real adve
13. r, and the strangerous relationship makes him uncomfortable. But some people li
14. t bought the fake drugs and this can make his disease heavier. As the whole, Our
15. maged from it. Indstrial alcohol can make one dead. Fake commodities are also
16. tively. It can change one's life and make one's life very happy. Everyone, young
17. <no 0045> <score 11> <TITLE Make Our Cities Greener> <Band 4> The
18. y to get rid of fake commodities and make our life better. We should remember,

Note: The materials searched here are concerned with the causative structures of "make + noun + adj." format.

V. MicroConcord Search SW: make*\made for Its Causitive Structures (2)

1. makes waste. Firstly, it will make you become very nervous. Under this others.
2. In this kind of job, it makes him have the sense of safety. Someon
3. must listen a lot of tapes and make many speaking practices to improve our En
4. ep studying the skill. This will make me do the job much better than others.
5. find the best position, but may make one no gain finally. <BAND 6> <SEX ?> <Y
6. etimes you find that you can not make promotion and be annoyed it. you must
7. nothing. On the other hand, haste makes people not know how to do. So people
8. short, "Practice Makes Perfect" makes the baby get the great progress in learni
9. now, they failed. They couldn't make their dreams come true at all because it
10. ege enough. Job-Hopping will make them find the job which is most suitable for te
11. ou facing. Secondly, haste will make you become careless. You can not do or
12. ther of them think that this can make them doing their cause deeply. And they
13. course, there are other factors make them change their job . For example, their
14. is a food banner, because it can make us do this work more and more effciencely
15. benefit from it. First it often make us overestimate ourselves. We perhaps
16. ore practice. Much more practice makes workers produce perfect product with
17. curious and the curiosity will make you have mistakes easily. It can't indicate
18. icates your common ability. It make you lose the chance. For example, when yo
19. the bag. It is the haste that make you have to go back home. It is certain
20. perience and skill. it can also make you learn the sample way to solve problems
21. "Practice Makes Perfect". It will make you obtain many of success. <BAND
22. in firm with familiar people can make you relax. But, on the other hand, q
23. ll show you examples in order to make you understand. Do you hear of the story
24. ce of practices. Practises will make you use you two hands, ten fingers respect
25. s skillfully. More practices can make you do your study or work very well and
26. like you present job. you should make yourself like it, whatever you do, you
27. have been used to these jobs which have made them master some skills. At the
28. g our learning English, practice makes us remember the words that we wrote many
29. do. above all, we should try to make us love the job. Thus, we cannot often
30. basic learning. Hard working can make us success, the "Hastes" is just a kind

Data from the following files: *The Chinese College English Learner Corpus*

Note: The materials searched here are concerned with the causitive structures of "make + noun + infinitive" format.

VI. Tables for "make + O. + C." Structures & Their Possible Lexicalized Verbs

	"Make + N + adj. / adv." Structure	Possible lexicalized Verbs
1.	a lot of troubles, which make us discouraged	discourage
2.	a stable job makes you satisfied and feel safe	satisfy, suffice
3.	a steable work may make our iife comfortable	comfort, *stabilize
4.	a woman buy some oil to make her skin well	improve, protect
5.	CH3OH in fake wine will make bright eyes dim	darken, faint, weaken
6.	Doing things too hurry will make you careless	*slack
7.	Experiment and practical make them favourable	favor
8.	fake commodities also can make us angry	annoy, irritate
9.	Fake commodities make people headache	trouble, bother, *plague
10.	fake commodities will make the society unsteady	*stabilize, firm
11.	fake drugs can make his disease heavier	worsen, *deteriorate
12.	Fake machines may make factory bankrupt	bankrupt

"Make + N + V (infinitive)" Structure	Possible Lexicalized Verbs
1. made our national economic develop slowly	slow (down), *stagnate
2. made them get more salaries or be promoted	promote, raise, award
3. make a person get more	benefit, win
4. make consumers diminish the purchase	lessen, diminish
5. make country to lose a lot of money	damage, injure
6. make every consumer know the danger of it	warn, advise
7. make he show his ability fully	show, demonstrate
8. make him do the job more and more efficiently	improve, *enhance
9. make him feel comfort and safe	comfort, secure
10. make him having new feeling	refresh, *energize
11. make his English level progress fastly	quicken, hasten
12. make his oil go through a very little hole	*dribble, *trickle, drip
13. make me do the job much better than others	*outdo, excel, surpass
14. make me fell pleasure	please, amuse, delight
15. make me live more comfortable	relieve, comfort, ease

Note: Since the causative verb list and the concordance lines of the causative 'make + O + C' structures are so large a bank of data that the writer can not possibly present all the relevant materials here. Instead, they have chosen some of them in the appendixes.

A Comparative Study on the Use of Coordinators between CLEC and LOCNESS

Yang Bei
Guangdong University of Foreign Studies

Abstract: Connectors play a very important role in making communication coherent and clear. A number of studies have shown that the use of connectors is problematic for language users, in particular foreign language learners. In this paper, the author has made comparisons on the use of coordinators of CLEC and LOCNESS. It is observed that, compared with native speakers' writing, Chinese learners' writing not only overuse or underuse coordinators, but also misuse or avoid using coordinators. It is also found that Chinese learners tend to put coordinators at sentence-initial position in academic writing which bespeaks that they lack register awareness. In addition, the use of coordinators of Chinese learners' writing differs from that of native speakers' writing in terms of the number of occurrences of different semantic relations: Chinese English learners overuse some semantic relations and neglect others. In the end of the paper the author discussed the pedagogical implications of this corpus-based study.

Key words: LOCNESS, CLEC, coordinator

1. Introduction

Effective communication requires coherence and clarity. One way of achieving this is to signal logical or semantic relations between units of discourse with connectors. Connectors can be said to function as cohesive 'signposts' in discourse (Leech and Svartvik, 1994, p.177), helping the listener and reader to relate successive units to each other and thus making sense of the text. A number of studies have shown that the use of connectors is problematic for language learners, in particular foreign language learners (e.g. Crewe, 1990; Granger & Tyson, 1996; Altenberg & Tapper, 1998). The following several points account for the difficulty in the correct use of connectors. Firstly, connectors are not always needed since relations that can be inferred from the text do not have to be marked explicitly; on the other hand, underuse and misuse of connectors are likely to make the text less comprehensible. Secondly, the use of connectors is sensitive to register and discourse type. Therefore, connector usage is dependent on the development of the learner's communicative competence and how language is taught. Thirdly, the use of connectors tends to vary from one language and culture to another (Altenberg &Tapper, 1998, p.80-81). Chinese is a distant language from English. The arrangement of most Chinese clauses uses parataxis which emphasizes on covert coherence, while the arrangement of most English clauses uses hypotaxis which emphasizes on overt coherence, so the Chinese speaking learners may demonstrate unique features in their TL production, especially in the use of connectors. The author of this paper carried out a comparative study on the use of coordinators* (and, but and or) between native speaker's writing and Chinese English learner's writing.

* According to Longman Grammar of Spoken and Written English, coordinators include *and, but, or* and *nor*. Nor occurred only 5 times in CLEC and 16 times in ICLE, therefore in this paper the author will not study coordinator *nor*.

2. Data and Methodology

The computer learner corpus used for the study is based on a sub-corpus of CLEC, which includes more than 2000 compositions written by college learners (non-majors in English) who have passed band 4 or band 6 English test. This sub-corpus contains 394,255 words. The native speaker control corpus used is Louvain Corpus of Native English Essays (LOCNESS), a sub-corpus of International Corpus of Learner English (ICLE). It is a collection of texts written by British and American students and contains 181,678 words. All the essays in both corpora are argumentative in character.

The corpus concordance software used in this study is WordSmith Tools by Mike Scott. Since the sizes of the two corpora used are different, all the raw counts are computed into normalized frequencies (occurrences per 100,000 words). SPSS will be used to deal with statistics. All frequency differences across the samples were tested by means of the chi-square test, with 99% as the critical level of confidence ($p<0.01$). An asterisk in the tables marks statistically significant differences between corpora.

3. Findings and Discussions

3.1 Normaiized Frequencies of Coordinators in LOCNESS and CLEC

	LOCNESS	CLEC	P value
and	2637.63	2233.83*	.000
or	337.41	263.03*	.003
But	396.86	497.39*	.001

Table 1: Normalized frequencies of coordinators in LOCNESS and CLEC

The table brings out a significant difference between the compositions by Native English speakers and Chinese English learners, i.e. the latter underuse coordinators *and* and *or*, while overuse *but*.

The reason that Chinese learners underuse *and* and *or* might be L1 interference. The clauses in a Chinese complex sentence are usually connected by parataxis, whereas those in an English complex or compound sentence by hypotaxis (Xiao, 1982; Wang, 1990). One of the main differences between English and Chinese complex sentence lies in the fact that connectives are much less imperatively needed in a Chinese complex sentence than in an English complex or compound sentence. Naturally, in an English text there are usually some obvious conjunctions that will be omitted in its Chinese version. However, it is quite wrong to consider hypotaxis an English pattern and parataxis a Chinese pattern. Both are used in Chinese and English text. The discussion above is just a general tendency.

A second explanation for Chinese learners' underuse of coordinator *and* and *or* might be that Chinese students do not use as many complex sentences as native speakers do. Mean sentence length in Chinese learner corpus is 17.23 words, while that in Native speaker corpus is 27.48 words. Although longer sentences are not necessarily complex sentences, complex sentences do tend to be longer. According to a study done by Ma in 2001, English compositions by Chinese learners are mainly made up of simple sentences, and native speakers mainly use complex sentences in writing. The reason that Chinese students avoid using complex sentences might be that they are not sure how to use these sentences correctly, so they resort to simple sentences in writing.

According to the finding of Longman Grammar of Spoken and Written English (LGSWE), *and* and *or* are considerably more frequent in academic prose than in conversation. LGSWE concludes that the high degree of phrase-level coordination is responsible for the high overall frequency of *and* in academic prose. On the

other hand, the low degree of coordination at the phrase level in conversation, which is consistent with the general simplicity of phrases in this register, accounts for the unexpectedly low frequency of *and* in conversation. The high frequency of *or* in academic prose is probably also to a great extent a reflection of coordination at the phrase level. The low frequency of *and* and *or* in Chinese learners' compositions might have some relation to the fact that writing by Chinese students is more like conversations as coordinators *and* and *or* are mainly used at clause-level.

From table 1, there is a significant difference in the frequency of *but* in CLEC and LOCNESS. Chinese learners overuse *but*. In the above section, we concluded that Chinese English learners tend to use less overt connective devices due to L1 interference. The overuse of *but* does not contradict with the conclusion since they may underuse overt connective devices on the whole and overuse some particular ones at the same time.

In LGSWE it is found that *but* is most frequent in conversation and fiction, and least frequent in academic prose. The high frequency of *but* should be seen in conjunction with the high frequency of negatives in conversation. Negation and contrast are closely related concepts. The high frequency of *but* in negation and contrast is due to the fact that conversation is interactive. The speaker can use *but* to modify a statement, and the addressee can use it to express a contrary opinion, refute a statement by the interlocutor and reject a suggestion. The low frequency in academic prose may be due in part to the fact that contrast is more often expressed by other means in that register: forms such as *however* and *yet* which are more frequent in academic prose than in the other registers. The findings in this study verify the conclusion in LGSWE.

	LOCNESS	CLEC	P value
Negatives	1119.01	1168.79	.296

Table 2: Frequency of negatives in LOCNESS and CLEC

Negatives in this study include *not, no, *n't, never, nor* and *neither.* This table does not bring out a significant difference in the frequency of negatives in ICLE and LOCNESS, but the number of occurrences of negatives in the writing of Chinese learners is higher than that in native speakers' composition.

	LOCNESS	CLEC	P value
however	193.20	58.85*	.000
yet	55.04	3.30*	.000

Table 3: Frequency of *however* and *yet* in LOCNESS and CLEC

Table 3 brings us the fact that Chinese learners overuse *but*, which is informal in style and mainly used in conversation. From table 3 it can be seen that Chinese learners underuse formal conjuncts *however* and *yet*. The fact suggests that Chinese learners lack a stylistic certainty about the use of connectors in argumentative writing, so they tend to avoid formal conjuncts and replace them with more informal equivalents.

3.2 Misuse of coordinators in CLEC

Chinese English learners not only underuse coordinators but also misuse or avoid using them, which could be roughly grouped into the following 4 categories: redundant coordinators, misused coordinators, ungrammatical structures connected by coordinators and missing of coordinators.

(i) redundant coordinators e.g.

1. ening broadcast, reading newspaper **and** etc. In addition, they should take part in
2. Because they...working condition. **And** they do it well. If they change work they
3. but the waste of time, material **or** etc. For instance, a certain man want to bu
4. basic factor. If you don't like to do **or** have no interests, you may not have imput

5.	y know a say "havete mares waste". **But** why we say so? When one does a thing,
6.	etten from a brook, or under ground. **But** in fact fresh water, is scarce nowadays.

In lines 1 and line 3, the deletion of the emphasized *and* and *or* will make these sentences grammatical. In line 2, connector *because* alone could realize cause-effect relation between the first and second sentence, so the *and* at the beginning of the second sentence is redundant. In line 4 the clause following *or*, which is the simple repetition of the previous clause, should be deleted. In line 5 the adversative or concessive relation doesn't exist between the two sentences connected with *but*, so *but* is redundant. Both *but* and *in fact* in line 6 express adversative relation and either one of them must be deleted.

(ii) misused coordinators e.g.

1.	ng in a family, advertising in streets **and** other odd jobs. Yes, we can know the wo
2.	r is that one in a hurry can't do well **and**, it maybe postpone the achievement. The
3.	we often faced a problem: if the lake **or** the river is dry, what the people could dep
4.	at the beginning we can't drive fast, **or** we can't know some skills in driving, but a
5.	expectancy in 1960 is 40 years old. **But** in 1990 it becomes 60 years old. Secondl
6.	will go in for a job through my life, **but** I will still go in for a lot of amateur job, t

According context, the alternative relation is needed in line 1, so the emphasized *and* should be replaced with *or*; while in line 3 the addition relation is the most appropriate, therefore the emphasized *or* need to be replaced with *and*. In line 2, the relation between the first and the second clause should be adversative, therefore *on the contrary* but not *and* is suitable there. To express the causative relation in line 4, *and* need to be replaced with *because*. The addition relation is needed in line 5 and 6, so *but* should be replaced with *and*.

(iii) Ungrammatical structures connected by coordinators

1.	expectancy is 40 years old in 1960 **and** is 60 years old in 1990. Thus, the life expe
2.	d a much better working condition **and** pays. My view on job-lopping is that we s
3.	As a student studying in university **or** college, it is necessary to get to know the
4.	a same work all his life, because he **or** she don't want to change their position, per
5.	eached 200 deaths per 1.000 births, **but** to 1990, it decreased by 50%. What cause
6.	e have 200 deaths per 1,000 births, **but** only have 100 deaths in 1990. Therefore, t

In lines 1 and 6 the second clause after and *lack subject, which lead to the ungrammaticality of the sentences. The structure of the clauses connected by coordinators should be equivalent, therefore in line 2, the adjective* higher *need to be added to modify* pays *so that the structure of the second phrase also conform to the adjective+noun structure. Likewise, the phrases connected by* or *in line 3 should be revised as 'in university and in college' and the preposition* to *after* but *in line 5 should be* in. *According to the principle of proximity, the predicative need to be in agreement with the nearest subject to it, which in line 4 is* she, *so the correct predicative should be* doesn't *but not don't.*

(iv) missing of coordinators

1.	continue to do its practice and practice, # we can do it perfectly. During practicing
2.	ecide the wordes for two or three years, # gradually you can find the more words
3.	er the item and do it, you can be better, # best. As we all know, "practice makes pe
4.	ou can be familiar to it and do it better, # perfect. For example, when you study E

The instances of missing of *but* is quite few in Chinese English learners' writing, so the missing of

coordinators is focused on *and* and *or*. All the above sentences are run-on sentences. For lines 1 and line 2, *and* is needed after the comma to make the sentence grammatical. For lines 3 and 4, *or* needs be added after the punctuation mark in the middle of the figure above.

3.3 Position of coordinators in LOCNESS and CLEC

There is a well-known prescriptive reaction against beginning an orthographic sentence with a coordinator. It will be of significance to study the position of coordinators in LOCNESS and CLEC

	LOCNESS	CLEC	P value
And	19.82	177.80*	.000
or	0.55	6.59	.034
But	51.19	302.34*	.000

Table 4: Frenquency of coordinators in sentence-initial position in LOCNESS and CLEC

The table shows that there is a significant difference between the frequency of *and* and *but* in sentence-initial position in CLEC and LOCNESS. There is no significant difference in the frequency of *or* in sentence-initial position in CLEC and LOCNESS, but the frequency of *or* in sentence-initial position in Chinese learners' composition is almost 12 times more than that in native speakers' writing. According to the research result of LGSWE, coordinators in sentence-initial position is considerably more common in conversation than in the written register. The larger proportion of coordinators in sentence-initial position in CLEC bespeaks that Chinese learners are insensitive to register distinctions in the target language, and that their register awareness need to be strengthened.

3.4 Distribution of different semantic relations of coordinators in LOCNESS and CLEC

Coordinators could express different semantic relations in connecting complex and compound sentences or clauses. To find out the similarities and differences in using these semantic relations between Chinese English learners and native speakers, 6 compositions from LOCNESS (3,357 words) and 20 composition from CLEC (3,346 words) were randomly selected and the distribution of different relations were studied. Since the two samples are comparatively small and the sizes are almost the same, raw counts were not transformed into normalized frequencies.

Category	LOCNESS	Percentage	CLEC	Percentage
Addition	53	69.7%	48	80%
Sequence of of actions	2	2.6%	1	1.7%
Resultive	17	22.4%	7	11.7%
Emphasis	1	1.4%	4	6.7%
Contrastive	1	1.4%	0	0
Adversative	2	2.6%	0	0
Total	76	100%	60	100%

Table 5: Number of occurrences of *and* in 6 semantic relations in LOCNESS and CLEC

And could mainly express the semantic relations of addition, sequence of actions, result, emphasis, contrastive and adversative. The table shows that with the exception of the category of emphasis, both the total number of occurrence and the occurrences of *and* in five other categories in CLEC are less than those in LOCNESS. What's more, Chinese learners don't use *and* to express contrastive and adversative relations. In the writing of Chinese learners, 80% of the coordinator *and* belongs to the category of addition, which is

10% higher than that in Native speaker corpus. This may be related to English teaching and the L1 interference. *And* is learned early as a core word and Chinese students tend to equate it with the Chinese "和" (he: and) which is used to express additional relation, so the great majority of *and* in Chinese students' compositions belong to this category. In Chinese, *he* couldn't be used to express the relation of sequence of actions, result, emphasis, contrastive and adversative, which result in difficulty in L1 transfer, therefore the occurrences of *and* in these categories are less than that in native speaker writing on the whole. The use of *and* to express contrastive and adversative relations may appear completely strange to Chinese learners, since in their mind *he* couldn't be used to express these relations, and they would rather use *however* to express contrastive relation and *but* to express adversative relation. The only category in this table that Chinese learners use more than Native speakers do is the use of *and* to convey emphasis relation. It might because that Chinese learners have realized that *and* sometimes means "又" (you: again) which express emphasis in Chinese, and therefore they could use *and* in this category more freely.

Category	LOCNESS	Percentage	CLEC	Percentage
Adversative	1	9.1%	13	52%
Concession	10	90.9%	12	48%
Total	11	100%	25	100%

Table 6: Number of occurrences of *but* in 2 semantic relations in LOCNESS and CLEC

The semantic relations expressed by *but* are adversative and concession. This table illustrates that both the total number of occurrence and the occurrences of *but* in the two categories in CLEC are greater than those in LOCNESS. This is in accordance with the above finding that Chinese learners overuse *but*. Compared with the Native speakers' writing, the percentage of *but* in the category of adversative is greater, while the percentage in the category of concession is less. English teaching in China may lead to the above problem. Chinese learners may be taught that *but* is used to express adversative relation while *though* and *although* concessive relation.

Category	LOCNESS	Percentage	CLEC	Percent
Alternative	7	58.3%	4	57.1%
Replacement	2	16.7%	1	14.3%
Condition	3	25%	2	28.6%
Total	12	100%	7	100%

Table 7: Number of occurrences of *or* in 3 semantic relations in LOCNESS and CLEC

Or mainly express the semantic relation of alternative, replacement and condition. The total number of occurrence and the number of occurrences of the three categories, i.e. alternative, replacement, and condition, in CLEC are lower than those in the LOCNESS, so Chinese learners should be encouraged to use more *or* in their writing. The proportions of each category in the two corpora are quite similar, which bespeaks that Chinese learners could use different meaning categories of *or* correctly.

4. Conclusion and Major Pedagogical Implications

From the above study, it is found that Chinese Learners underuse *and* and *or* and overuse *but*. Chinese English learners also misuse and avoid using coordinators which could be classified into the following four categories: abundant coordinators, misused coordinators, ungrammatical structures connected by coordinators and missing of coordinators. Generally speaking, Chinese learners tend to place coordinators in sentence-initial position, which suggests that Chinese learners couldn't distinguish the style of academic

writing from that of conversation and their register awareness need to be strengthened. Having studied the three thousand-odd words compostions in LOCNESS and CLEC, the author got to the conclusion that the number of occurrence of different semantic categories of coordinators in Chinese learners' writing differs from that in native speakers' writing. Chinese English learners overuse some semantic relations and neglect some others.

Given the big difference between Chinese and English and the fact that Chinese EFL learners' writing is influenced by their mother tongue, contrastive English-Chinese teaching will shed light on EFL teaching and learning. The study shows that Chinese EFL students lack stylish awareness, therefore they need to be exposed to a greater range of registers and to a more extensive training in expository writing.

Acknowledgement:

For the completion of the current paper, I acknowledge my heart-felt gratitude to my mentor Dr. Li Wenzhong, who has given me valuable instruction and constructive criticism.

References:

Altenberg, B., & Tapper, M.(1998). The use of adverbial connectors in advanced Swedish learners' written English. In Granger, S. (Ed.), *Learner English on Computer.* London and New York: Longman.

Altenberg, B., & Sylviane, G. (2001). The Grammatical and Lexical Patterning of MAKE in Native and Non-native Student Writing. *Applied Linguistics*, 22, 173-193.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Edward, F. (1999). *Longman Grammar of Spoken and Written English.* Pearson Education Limited / Beijing: Foreign Language Teaching and Research Press. 2000.

Crewe, W.J. (1990). The illogic of logical connectors. *ELT Journal*, 44, 316-25.

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

Halliday, M.A.K., & Ruqaiya, H. (1976). *Cohesion in English.* Pearson Education Limited. Beijing: Foreign Language Teaching and Research Press. 2001.

Leech, G., & Svartvik, J. (1994). *A Communicative Grammar of English.* London: Longman.

Li, Wenzhong. (1998). An Analysis of the Lexical Words & Word Combinations in the College Learner English Corpus. PhD dissertation, Shanghai Jiaotong University.

Liu Jiarong, & Jiao Hui. (2002). Engiish and Chinese Textual Connective Devices. In Chen Zhi'an, Liu Jiarong, & Wen Xu. (Eds.), *Contrastive English-Chinese Pragmatics And TEFL.* Beijing: Foreign Language Teaching and Research Press.

Odlin, T. (1989). *Language Transfer.* Cambridge: Cambridge University Press.

Qian Yuan. (1990). A comparison of some cohesive devices in English and Chinese. In 杨自俭．李瑞华（编）．英汉对比研究论文集．上海:上海外语教育出版社.

Richards, J.C., & Sampson, G. P. (1984). The Study of Learner English. In Richards, J.C. (Ed.), *Error Analysis: Perspectives on Second Language Acquisition.* London: Longman Group UK.

Selinker, L.(1972). Interlanguage. In Richards, J.C.(Ed.), *Error Analysis Perspectives on Second Language Acquisition.* London: Longman Group Limited.

Xiao Junshi. (1982). *An Approach to Translation from Chinese into English and Vice Versa.* Beijing: Commercial Press.

胡壮麟. (1994). 《语篇的衔接与连贯》. 上海: 上海外语教育出版社.

连淑能. (1993). 《英汉对比研究》. 北京: 高等教育出版社.

马广惠. (2001). 中美大学生英语语篇对比修辞分析. 《解放军外国语学院学报》, 6, 5-8.

赵永新. (1983). 汉语的 "和" 与英语的 "and". 《语言教学与研究》, 1, 86-97.

A Corpus-based Study of Characteristics of Adjective Collocation in CLEC

Sun Haiyan
Shanghai Jiaotong University

Abstract: Collocation is a ubiquitous phenomenon in language because of the density of its occurrence and is becoming an increasingly important field in language teaching. However, few methodical studies have been made as regards the learners' collocational behavior, especially the characteristics of their use of adjective collocation. The present paper attempts to make an investigation into the collocational characteristics in Chinese learners' use of adjective collocation, by comparing the adjective collocational patterns of Chinese learners with those of native speakers on the basis of corpus evidence.

Two types of corpora are adopted in this paper: a learner corpus – CLEC, and reference corpora – LOB, JDEST, Cobuild. The concordance software used in this study is MicroConcord and WordSmith. The statistical methods of Z-score and MI-value are adopted to measure the significance of the co-occurrence of collocates. The present study focuses on the semantic characteristics of Chinese learners' use of adjective collocation, and summarizes three typical characteristics: semantic imprecision in the selection of collocates, semantic confusion of polysemous adjectives, and semantic inharmony.

The present paper further makes a tentative exploration of the possible causes of the learners' collocational incompetence, namely, the mother tongue interference, and the communication strategy-based errors. The feasible solutions are proposed to improve the learners' collocational competence. Concordance should be introduced into language classroom, and learners should be provided with adequate, high-quality input. Through describing the semantic characteristics in learners' use of adjective collocation, the present paper intends to arouse the teachers' as well as learners' awareness of these weaknesses so that they will try to enhance the learners' collocational competence.

Key words: corpus; adjective; collocation; semantic characteristics

1. Introduction

Collocations are indispensable and ubiquitous elements in English and their significance to language teaching can by no means be ignored. However, the study on learners' use of collocation has been largely neglected by researchers and practitioners. And in the research of learners' language, little attention has been paid to the analysis of the use of adjective collocations. According to Leech (1989), adjectives are the largest open word class in English after nouns and verbs, and grammatically and semantically, they have the same degree of importance as the other content words in the language. Therefore, the present author intends to make a tentative exploration into the collocational patterns in Chinese learners' use of adjective collocations in terms of semantic characteristics. This study tries to make a combination of quantitative measurement and qualitative analysis.

2. Research Question and Method

Second language learners frequently make grammatically well-formed sentences which nevertheless sound awkward or unnatural to the native ears, one reason of which is that they have not internalized enough

knowledge about collocation in English. This study tries to detect the collocational problems in Chinese learner' use of adjective collocations and to offer conventional collocations used by native speakers. The research question is specified as: What is the semantic characteristic in Chinese learners' use of adjective collocation?

The present analysis is based on two kinds of corpora: a learner corpus which is exploited for the study of the learners' collocational errors and infelicity, and reference corpora which are used to extract conventional collocations. The Learner Corpus adopted here is CLEC – Chinese Learner English Corpus, which is constructed by Shanghai Jiaotong University and Guangdong Foreign Studies University, with a total of 1.1 million running words. Three corpora of native speakers are employed as reference: LOB, JDEST and Cobuild. In addition, the present study employs *The American Heritage Dictionary* from Powerword (2002) for some definitions of words.

The concordance software used in this study is MicroConcord and WordSmith. The statistical instruments of Z-score and MI-value are adopted to process the data so that the data extracted from different corpora is comparable. Z-score compares the difference between the observed frequency of a collocate and its expected frequency in standard deviation units. Mutual information (MI) measures the collocational strength between words. As to the specific procedures to calculate Z-score and MI-value, see Wei (2002: 44-50).

3. Semantic Characteristics of Adjective Collocation in CLEC

The first step of the present research is to select some node words for the analysis and generalization of the features in the learners' use of adjective collocations. The selection is largely based on two criteria: the first one is the frequency list generated by WordSmith of the corpus of CLEC; the second one is the error tagging scheme designed by the Institute of Linguistics and Applied Linguistics of Guangdong Foreign Studies University. According to these two criteria, the following eight words are selected as node words: "big (497), large (362), great (1273), average (64), common (202), ordinary (71), rather (209), quite (300)". The number in the bracket refers to their frequency in CLEC.

3.1 Semantic Imprecision of Adjective Collocation in CLEC

The focus of collocational study is usually on the typical collocational behavior. The notion of typicality is different from possibility, since "there are virtually no impossible collocations, but some are more likely than others" (Sinclair, 1966: 411). So the analysis of the collocational behaviour of *big, large, great* will focus on their typical noun collocates.

3.1.1 The Collocational Pattern of 'big, large, great' in LOB

Firstly a brief analysis of the conventional use of these adjectives by native speakers will be made based on the evidence from LOB. The statistical evidence reveals the different collocational patterns of these three seemingly synonymous adjectives.

> In LOB Corpus, the overall frequency of *big* is 183, and most noun collocates are concrete ones, like *car, hand, school, country,* etc. Here the adjective *big* is used to describe the physical size of objects with the sense of 'of considerable size'.

The frequency of *large* is 423 in LOB. Its most significant noun collocates are *"scale, number, quantities, sums, majority, proportion, amount"*, and these seven collocates are all nouns relating to the amount or quantity of

something. Other collocates include the mass nouns like *audience, supply*, and countable ones like *school, country*. But on the whole, *large* is predominantly used in connection with the nouns indicating quantity.

The frequency of *great* is 685 in LOB. In contrast to *big*, the collocates of *great* are mostly abstract nouns, such as *danger, importance, influence, pleasure, value*, in which *great* is employed to emphasize the 'remarkable degree' of the nouns it modifies. The definition of *great* in *The American Heritage Dictionary*, however, goes as follows: the first sense is 'very large in size'; the second sense is 'large in quantity or number'; and not until the sixth comes the definition of 'remarkable or outstanding in magnitude, degree, or extent'. Nevertheless, the most frequent collocates indicate that this sixth sense is just the one most commonly used.

3.1.2 The Collocational Pattern of 'big, large, great' in CLEC

After a brief analysis of the collocational patterns of the above three adjectives in LOB, their collocational behavior in CLEC will be investigated. The preliminary research findings indicate that the learners show a feature of semantic imprecision in their use of adjective collocations. By **semantic imprecision** is meant the learners' utterance is unclearly expressed in terms of their selection of adjectives in adjective collocations. In other words, the learners have a strong tendency to use a general word to substitute a specific one. The significant collocates of the node 'big' are listed in Table 3.1.

Table 3.1: The Noun Collocates of 'big' in CLEC

Collocate	C_1	C_2	Z-score	Collocate	C_1	C_2	Z-score
cat	82	13	22.96	moon	583	9	4.63
building	268	18	17.01	river	269	4	2.99
smile	84	5	8.38	burden	172	3	2.95
noise	101	4	5.91	family	907	8	2.51
school	1789	20	5.17	hope	607	6	2.49

The total occurrence of *big* in CLEC is 497, and though most of its collocates do not violate the convention of native speakers, they show the learners' imprecision in the use of adjectives. The word combinations like *big burden/change/honor/hope/moon/noise/purpose/reason/smile* indicate that learners use this 'general' word (*big*) to substitute a more precise one. In contrast, native speakers will use a more precise adjective with different nouns: *"heavy burden; great/considerable change; great honor; best hope; full moon; loud noise; main purpose; good/major reason; broad smile"*.

The examples above suggest that one problem in the learners' use of adjective collocation is semantic imprecision. One reason is that the learners literally translate the Chinese words into English, which may be regarded as one form of mother tongue interference. In Chinese, we say '大雨', '大风', and learners translate them directly into *big hope, big noise*. The occurrences of the phrases like *big purpose/reason* in CLEC may be due to the fact that the learners are more familiar with the adjective *big* than with *main/major*. Consequently, they employ the word *big* without giving consideration to its typical collocational behaviour.

The overall frequency of *large* in CLEC is 362, and the unnatural collocations can be analyzed from two angles, based on the following concordances from CLEC:

1 e and other conditions result in the large changes of ?the life expectancy and
2 ith a bamboo basket --- all in vain. Large costs and ?manpower are worried by
3 ,-], the developing countries take a large improvement [cc4,?1-] [cc3,3-] in ec
4 ? 1-2] the job and when they have a large knowledge [cc4, 2-] in a field, he c
5 it by dreaming. But they have made a large mistake. A lot of scientists think
6 am of getting success , he must do a large practice [cc4,1-] ahead. On account
7 global shortage of fresh water is a large [cc4,-1] problem.? Because now we us
8 rious problem for all ?of us to pay large [cc4,-1] attention to. [sn8,s-] <ST
9 from it [pr3,d]. Fake commodities do large [cc4,-1] harm to our society. For ex
10 he top of the field quickly and make large progresses [cc4,1-]. But you also m

Firstly, the instances of 1-7, such as *large costs, large mistake,* point to the semantic imprecision in the learners' selection of adjectives, in that native speakers tend to use more precise adjectives, like *high (costs), serious (mistake),* to express that idea. One reason is that the learners have a limited range of word selection. Take the word *improvement* as an example: the learners can only use the frequent adjectives like *large, great,* while native speakers employ a wider range of adjective collocates like *great, marked, considerable, substantial*:

LEARNER	**NS**
large change	great/dramatic change
large costs	high costs
large improvement	great/marked/considerable/substantial improvement
large knowledge	wide/extensive knowledge
large mistake	serious/worst mistake
large practice	a lot of practice
large problem	serious/major/crucial problem

Secondly, the concordances from 8 to 10 indicate the confusion of the adjective *large* with *great*, as is suggested by the collocates like *attention, harm, progresses*. As mentioned above, native speakers commonly use *great* to collocate with these abstract nouns, but learners use the adjective *large* in these instances. In language classroom, most teachers explain *large* simply as an equivalent of *'very big'* without mentioning its collocational feature. The fact that collocation is a rather neglected field in English teaching, to some extent, accounts for the learners' collocational incompetence.

The frequency of the adjective *great* – 1274, is the highest one in CLEC among these three adjectives; however, there are few improper collocations with it. Most of its collocates are abstract nouns and they are perfectly acceptable in light of the collocational behaviour of the word *great* in LOB. The unnatural cases are *great burden/death.* As to the first one, native speakers prefer to use *heavy burden.* And *great death* can be more precisely rewritten as *glorious death.* These two instances also suggest that learners tend to use a general word (*great*) to substitute a precise one.

In summary, in the pattern of 'Adj. + Noun', Chinese learners show a great idiosyncrasy of semantic imprecision and create many imprecise word combinations. Teachers and learners should attach great importance to it and try to develop some methods to deal with it.

3.2 The Learners' Semantic Confusion of Polysemous Adjectives

Robins (1967) believes that word meaning does not exist in isolation, and they may differ according to the collocation in which they are used. The polysemous words present a complex collocational pattern because

each different sense attracts different sets of collocates. When an adjective has several senses, one sense usually occurs more frequently than the others. Native speakers generally can judge the most frequently used one intuitively, while learners might lack this ability. Especially when several polysemous adjectives have one seemingly shared sense, confusion of them will occur. Take the three adjectives *average, common, ordinary*, as examples, and an investigation into their respective collocational range and pattern will reveal the difference in the frequency of their several senses.

The frequency of *average* is 819 in JDEST, and its significant collocates can be classified into two sense groups. The first one is "of, relating to a number that typifies a set of numbers", which is the most frequently used sense in that most collocates fall into this category. The frequency of the second sense "usual or ordinary in kind or character" is much less than the first one.

The adjective *common* occurs 1209 times in JDEST, and its collocates can be divided into three categories according to its three different senses. The most frequently used sense is "belonging equally to or shared equally by two or more; joint". For example, in the collocation of *common sense/practice,* it means *joint, shared*. The second one is "ordinary, having no special designation, status, or rank". In the phrases like *common metals/materials*, it adopts this sense. The third one is the sense of "occurring frequently or habitually, usual" when the noun collocates usually refer to some phenomena.

The total occurrence of *ordinary* is 231 in JDEST. A large proportion of its collocates refer to materials or things which are used in everyday life, e.g. *typewriters, metals, paper, steel*, and in these collocations it has the sense of "usual, commonly encountered".

From the analysis above, we can see that these three adjectives have a shared sense of 'usual', but this sense is the most frequently used one with the adjective *ordinary*, while being a less frequently used sense with *common* and *average*.

In CLEC, the frequency of *average, common, ordinary* is 64 and 202 and 71 respectively. Because they are polysemous adjectives with one roughly same sense, learners are confused by their subtle distinction and use them with different nouns just at random, thus creating many anomalous ' Adj. + Noun' collocations.

Table 3.2: The Noun Collocates of 'average' in CLEC

Collocate	C_1	C_2	Z-score	Collocate	C_1	C_2	Z-score
age	114	12	51.07	population	502	8	15.81
amount	125	5	20.17	water	4465	15	8.80
years	1693	18	19.08	days	510	4	7.59
level	226	6	17.89	life	3265	8	5.15
infant	824	11	16.87	persons	715	3	4.54

Table 3.3: The Noun Collocates of 'common' in CLEC

Collocate	C_1	C_2	Z-score	Collocate	C_1	C_2	Z-score
phenomenon	54	7	24.19	women	810	5	3.41
sense	144	8	16.66	saying	401	3	3.07
bicycle	76	5	14.40	person	715	4	2.80
feeling	160	3	5.60	people	7539	19	2.25
ability	318	3	3.63	friends	614	3	2.15

Table 3.4: The Noun Collocates of 'ordinary' in CLEC

Collocate	C_1	C_2	Z-score	Collocate	C_1	C_2	Z-score
citizens	56	4	22.99	life	3265	8	4.76
radio	290	6	14.88	problem	691	3	4.34
person	715	5	7.49	students	1660	4	3.32
woman	280	3	7.38	school	1789	4	3.12
day	1973	6	4.83	people	7359	9	2.57

Firstly, the learners' problem lies in their free and careless selection of the three words when they intend to mean 'usual', whereas native speakers typically use 'ordinary'. For example, the learners exploit the expression of *average days/life/person*, while native speakers frequently use *ordinary days/life/person*. The frequency of the instances like *common man/people/teacher* is also very high in CLEC, which indicates that the learners arbitrarily equal *common* to *ordinary* in the use of these collocations.

The second problem lies in the learners' use of the collocation of 'ordinary + nouns referring to phenomena', as is shown in the following sentences from CLEC: 'So *traffic jam is ordinary*, especially in the morning when you go to work'; 'Now, *bicycle theft is ordinary* on campus'. As is discussed earlier, *common* can modify nouns denoting phenomena, but *ordinary* generally not. In these two sentences, *ordinary* should be replaced by *common*, because here *common* collocates with the noun phrases – 'traffic jam' and 'bicycle theft', to refer to the frequently occurring phenomena.

The third one is the confusion of *common* with *average* – the learners use *common ability*, while native speakers typically use *average ability*.

To sum up, these unnatural collocations are produced because of the learners' semantic confusion of the polysemous adjectives. The learners are confused by the respective senses of these polysemous adjectives and mix them up in their collocation with nouns. As to the possible solution, the present author believes that it is very helpful for learners to acquire the idiomatic collocations as a whole: 'node word, collocates and meaning' all at the same time, because it is difficult to distinguish clearly the several senses of one word in isolation.

3.3 Semantic Inharmony in 'Adv. + Adj.' Pattern in CLEC

One major contribution of corpus research to linguistic study is its discovery and validation of semantic prosody. A semantic prosody, according to Louw (1993: 157), is a "consistent aura of meaning with which a form is imbued by its collocates". Semantic prosody can be categorized into three types: a negative prosody, a positive prosody, and a mixed prosody (Stubbs, 1996: 176). In a negative prosody, almost all the collocates of a node have a negative semantic feature, while in a positive prosody, most of the collocates have positive semantic characteristics. The investigation of semantic prosody in CLEC indicates that Chinese learners display the feature of **semantic inharmony**, by which is meant learners use a node showing a negative semantic prosody with a collocate which has a positive semantic characteristic, or vice visa.

The semantic inharmony in CLEC can be revealed in the pattern of "Adv. + Adj.': learners may collocate an adverb that has a negative semantic prosody with an adjective which has a positive semantic feature, or vice visa. In this section, we will firstly establish the conventional semantic prosody of two adverbs, *rather* and *quite*. They are under investigation because they have a high occurrence in CLEC (*rather 209; quite 300*), and learners demonstrate an obvious semantic inharmony when using them.

According to the MI value of the adjective collocates of 'rather' in Cobuild, an overwhelming majority of the collocates have a strong pejorative sense, such as *unpromising, uninteresting, snobbish, tiresome, dull,* etc. Though there are a few exceptions, they can't deny the fact that *rather* displays a negative semantic prosody. In contrast, most adjective collocates of 'quite' show the commendatory attitude of the language user, e.g. *honest, happy, interesting, remarkable.* The adjectives with a laudatory sense account for a high proportion, indicating that *quite* generally has a positive semantic prosody.

Because Chinese learners lack such knowledge, the collocations they produce seriously violate the typical semantic prosody of *rather* and *quite.* There are sixteen significant adjective collocates of 'rather' in CLEC, and we can classify them into three categories according to their semantic characteristics: positive, negative and neutral. There are five adjective collocates with positive semantic characteristics: *proud, practical, good, young, fresh.* Those with negative ones include seven adjectives: *ashamed, noisy, poor, inhumane, sad, bad, evil.* Four adjectives show a neutral semantic feature: *flat, cold, low, high.* The overall collocational pattern of *rather* in CLEC presents a mixed semantic prosody, in that it attracts three categories of collocates. However, native speakers commonly use it in collocation with adjectives with a negative sense. Consequently, some word combinations produced by learners, such as *rather fresh/good,* create a serious semantic inharmony.

As to the semantic prosody of 'quite' in CLEC, it also presents a mixed prosody. Its significant collocates can be categorized according to their semantic characteristics. The following three categories indicate the learners' arbitrary use of collocations, and the category of 'negative collocates' are not in harmony with the typical collocational pattern of 'quite' used by native speakers.

quite {
- positive collocates (8): *optimistic, nice, convenient, helpful, suitable, beautiful, etc.*
- negative collocates (6): *annoyed, nervous, wrong, serious, limited, difficult*
- neutral collocates (8): *different, common, aware, familiar, long, simple, high, large*

On the whole, the learners lack an awareness of the typical collocational behavior of the node word. The semantic inharmony occurring in CLEC will create inharmonious and unnatural collocations, which is a serious problem in the learners' language production. Learners should acquire knowledge about semantic prosody to produce more natural, native-like word combinations. Therefore, it is important to introduce the notion of semantic prosody into language classroom to enlighten the learners on this subject.

4. Discussion and Implication

4.1 Possible Causes

There are many possible factors that may account for the learners' inadequate collocational competence. In this part, two main sources will be discussed, namely, the mother tongue interference, and the communication strategy-based errors.

One possible cause of the learners' problems in collocation is that the learner is carrying over the habits of the mother tongue into the second language. This is called *interference* and the implication of this term is that one's mother tongue habits prevent him in some way from acquiring the habits of the second language (Corder, 1981). Since collocation is language specific, the collocational behaviour of words varies with language. Nevertheless, the learners might frequently adopt the literal translation strategy to render a Chinese expression into English, thus producing many unnatural or unacceptable collocations. For example, Chinese learners literally translate the following expressions, '很大变化', '很大荣誉', '很大希望', '噪音大', into 'big change', 'big honor', 'big hope', 'big noise'.

Collocational errors can be intralingual or interlingual. Besides the above-mentioned mother tongue interference, which falls into the category of interlingual, many problems in the use of collocation are related to intralingual errors. James (2001) elaborates one form of intralingual errors – communication strategy-based errors, and classifies them into two types: holistic strategies and analytic strategies. For James (ibid.: 187), the holistic strategy can also be termed as approximation, that is to say, when learners lack the required form, they tend to use a near-equivalent L2 item they have learnt. They may use a near synonym, a super-ordinate term, or an antonym. For instance, in the collocation *ordinary life/person*, the learners substitute *average* for *ordinary* in that they regard these two adjectives as synonyms. By analytic strategy, which is also termed as circumlocution, the speaker expresses the concept indirectly. by allusion rather than by direct reference. The learners' collocational incompetence often forces them to create periphrastic utterances.

4.2 Possible Solutions

As to how to improve the learners' collocational competence, the present author holds that we should introduce concordance into classroom. In the past ten years an approach to language learning has emerged in which learners sometimes work with 'raw' information taken directly from corpora, which is called DDL, or data-driven learning. Exposed to a huge quantity of real data, learners can effectively acquire language when they are encouraged to follow an observe-hypothesize-experiment model, i.e. when they draw their own conclusions about word/phrase meanings and collocations by examination of authentic linguistic evidence. For teachers, they can use a concordancer to find examples of authentic usage to demonstrate features of vocabulary and typical collocations.

Secondly, learners should be provided with adequate high-quality input to improve their collocational competence. If learners haven't had enough input of collocations, they will produce labored, cumbersome speech, using unnatural collocations to express their ideas. When learners have more opportunities to be exposed to the conventionalized patterns and idiomatic expressions, they are more likely to internalize them and utilize them.

5. Conclusion

In the present paper the importance of teaching English collocations to second language learners has been explored. By comparing the collocational behavior of Chinese learners and native speakers, this research shows a quite noticeable discrepancy between Chinese learners' collocational knowledge and the convention suggested by the corpus evidence from the reference corpora. Three semantic characteristics of adjective collocation in CLEC have been investigated: semantic imprecision in the selection of adjectives, semantic confusion of polysemous adjectives, and semantic inharmony as regards semantic prosody. A brief exploration into the causes of the learners' collocational incompetence has been made, and possible solutions put forward as well. Through generalizing the features of the adjective collocations in CLEC, the present study has tried to depict a picture in this regard so that the learners as well as teachers will attach great importance to the enhancement of the learners' collocational competence.

References

Corder, S. P. (1981). *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

James, C. (2001). *Errors in Language Learning and Use: Exploring Error Analysis*. Beijing: Foreign Language Teaching and Research Press.

Leech, G. (1989). *An A-Z of English Grammar & Usage*. Longman: Nelson.

Louw, B. (1993). "Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies", in Baker, M., Francis, G. & Tognini-Bonelli, E. (Eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins: 157-176.

Robins, R. H. (1967). *A Short History of Linguistics*. London: Longman.

Sinclair, J. (1966). "Beginning the study of Lexis", in Bazell, C. E., Catford, J. C., Halliday, M. A. K. & Robins, R. H. (Eds.), *In memory of J. R. Firth*: 410-430. London: Longman.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell Publishers.

卫乃兴，2002，《词语搭配的界定与研究体系》，上海：上海交通大学出版社。

A Corpus-based Study of Chinese EFL Learners' Acquisition of Derivational Affixes

Cui Yanyan Huang Ruihong
Shanghai Jiaotong University

Abstract: After decades of neglect, lexis is now recognized as central to language acquisition. As an important component of lexical competence, knowledge of derivational affixes is indispensable to Chinese EFL learners to tackle the task of learning large numbers of English words. Although somewhat rule-governed, the system of derivational affixes is a complex one which often causes learners' lexical errors. This study is intended to investigate Chinese learners' acquisition of derivational affixes by analyzing the derivational errors in their essay writing. The corpus used for the study is Chinese Learner English Corpus (CLEC), and the concordance software is Micro-concord 1.0. The first step is to locate all the concordances of word building errors, then, the errors related to derivational affixes are scrutinized manually. Five categories of errors are identified and tagged: overgeneralization, confusion of affixes, incomplete knowledge of word class, incomplete semantic knowledge, and spelling errors. The results show that derivational errors are widespread in the five sub-corpora. Chinese EFL learners, even at an advanced proficiency level, have great difficulty in using derivational affixes. The sources of these errors are analyzed in the light of interlanguage, error analysis, vocabulary acquisition and language typology. The findings of the study may shed light on the acquisition of derivational affixes and vocabulary teaching.

Key words: corpus; derivational affixes; acquisition; derivational errors

1. Introduction

It is a universally accepted fact that vocabulary is the most essential element in language processing. Without knowledge of words, no language can be understood. As an important component of lexical competence, knowledge of derivational affixes is indispensable to learners to tackle the task of learning large numbers of English words. With a limited knowledge of derivational regularities, a learner can achieve a tremendous expansion of his/her vocabulary. Recent research into the acquisition and retention of foreign and second language vocabulary has shown that newly acquired words are better retained if they were initially inferred through linguistic cues rather than through context (Haastrup 1987). Nagy and Anderson (1984) hold that 84 percent of the prefixed words and 86 percent of the derivationally suffixed words are semantically transparent in "printed school English", i.e. their meaning can be inferred on the basis of their constituent morphemes. Obviously, derivational cues for the inference of words in a second or foreign language can be essential to vocabulary acquisition. The present study is intended to analyze Chinese EFL learners' acquisition of English derivational affixes.

2. Research Questions

English derivational affixes can be classified into two groups: one of them can change the word class of an underlying word; another group serve to alter the meaning of a word. Although somewhat rule-governed, the system of derivational affixes is a complex one which often causes learners' lexical

errors. Some derivational affixes can only be added to a specific subset of a class of words. Some affixes are very productive while some only have a very limited range. Consider the prefixes for negation. We say *unable, dishonest, intolerable, nonproductive, amoral, displeasing*. All of these prefixes negate the morphemes that follow. If more than one of them can be used, the two resulting forms have different meaning or different restrictions for word collocations, which makes their acquisition very difficult (Hatch & Brown 2001). There may be as many as four of five competing affixes that appear to do the same, but they are not interchangeable.

Since Chinese is an isolating language without morphological changes in a real sense, Chinese learners may find the acquisition of derivational affixes especially difficult. The error-proneness of derivational affixes of Chinese learners has been highlighted in some recent studies (Lu 1983; Zhou 2000; Huang 2001). However, their discussion is only from the perspective of linguistic study and at theoretical level. Empirical studies are needed to deepen the understanding of the learning and teaching of derivational affixes. In this paper we aim to throw some light on the features of the acquisition of the derivational affixes. The main questions we will answer are:

1. What are the characteristics of Chinese EFL learners' acquisition of derivational affixes?
2. What factors are responsible for the learners' derivational errors?
3. What is the implication of such study on vocabulary teaching and second language acquisition?

3. Method

The computerized corpus used for the present study is Chinese Learner English Corpus (CLEC), which is constructed by Shanghai Jiaotong University and Guangdong Foreign Studies University, with a total of 1.1 million running words. CLEC consists of five sub-corpora of essay writing from 5 groups of students. *ST2* in CLEC refers to senior high school students. Non-English majors in Band 4 and Band 6 are named as *ST3* and *ST*. First-year and second-year English majors are marked as *ST5*, and third-year and fourth-year English majors *ST6*. The running words of the five sub-corpora are 208088, 209043, 212855, 214510and 226102 respectively. The concordance software adopted in this study is Micro-concord 1.0. The first step of the research is to locate all the concordances of word building errors tagged *fm2* in the five sub-corpora, then, the errors related to derivational affixes are identified manually.

Through analyzing all the related concordances, we find that the derivational errors in CLEC can be classified into five categories. The lexical errors caused by the overuse of affixes is tagged as *fm2,O* (overuse). The errors due to learners' incomplete knowledge of the part of speech are marked as *fm2,PS* (part of speech). The confusion of affixes is tagged as *fm2,CA* (confusion of affixes), which is further divided into three sub-categories: *fm2,CA1* (confusion of prefixes), *fm2,CA2* (confusion of suffixes) and *fm2,CA3* (confusion of prefixes and suffixes). The fourth category, *fm2,Se* (semantic), is about the errors relevant to learners' incomplete semantic knowledge. The fifth category is marked as *fm2,Sp* (spelling mistakes). Then we use the software to group all the derivational errors into the above five categories.

4. Results and Discussion

4.1 Results

The results of the study are summarized in the following table:

Table 1: Raw frequencies of the derivational errors in the five sub-corpora

	fm2,O	fm2,PS	fm2,CA1	fm2,CA2	fm2,CA3	fm2,Se	fm2,Sp	Total
ST2	8	3	1	0	1	0	1	16
ST3	45	40	8	8	0	1	22	124
ST4	82	61	17	30	12	2	17	221
ST5	12	4	5	9	0	0	3	33
ST6	49	15	47	12	0	0	27	150
Total	196	123	78	59	13	3	72	544

The numbers in the above table are raw data of derivational errors occurring in CLEC. Yet the running words in every sub-corpus are not the same, so the data should be standardized for the purpose of making comparison among the five groups of students. The method to standardize the raw data is to suppose that we have a corpus with 1,000,000 running words, and each sub-corpus has 200,000 running words. For example, the actually occurred derivational errors of overuse of affixes in *ST2* are 8, and the running words of this sub-corpus are 208,080. Its standardized frequency can be calculated like this:

$$8/208088*200000 = 7.69$$

The following table shows the standardized frequencies of the derivational errors.

Table 2: Standardized frequencies of the derivational errors in the five sub-corpora

	fm2,O	fm2,PS	fm2,CA1	fm2,CA2	fm2,CA3	fm2,Se	fm2,Sp	Total
ST2	7.69	2.88	0.96	0	0.96	0	0.96	13.45
ST3	43.05	38.26	7.65	7.65	0	0.96	21.04	118.61
ST4	77.04	57.32	15.97	28.19	11.28	1.88	15.97	207.65
ST5	11.19	3.73	4.66	8.39	0	0	2.80	30.77
ST6	43.34	13.27	41.57	10.61	0	0	23.88	132.67
Total	182.31	115.46	70.81	54.84	12.24	2.84	64.65	503.15

From the above table, we can find that the acquisition of English derivational affixes is very complex and demanding. On the whole, the derivational errors are widespread among the five groups of students. The senior high school learners' (*ST2*) relatively few errors cannot suggest that they have mastered the system of derivational affixes, it only shows that they have an inadequate access to the system. Another possible reason might be that language learners often avoid using derivational affixes in the early stages of acquisition (Hatch & Brown 2001). In contrast, the highly proficient learners made more derivational errors. The possible reason is that they have realized that derivation is an important way of word formation in English and attempted to use it adequately in their writing like the native speaker. In spite of the idiosyncratic acquisition characteristics of the five groups of learners, there are still some features indicating that learner language is systematic. Among the five categories of derivational errors, the errors tagged as *fm2,O*, *fm2,CA* and *fm2,PS* are the frequently occurred ones in the five sub-corpora, which indicates that learners with different proficiency are faced with the similar learning difficulties. The learners' spelling errors in using derivational affixes are also apparent in the results, with the standardized frequency of 64.65. Only 2.84 errors are relevant to the learners' incomplete semantic knowledge. The sources of the five categories of derivational errors will be analyzed in the following section.

4.2 Discussion

Selinker (1972) used the term *interlanguage* to refer to learner language. According to him, interlanguage is systematic, which means that learners do select the interlanguage rules in predictable ways. Our findings

are consistent with this feature of interlanguage. From the above table we can find that the overuse of derivational affixes is the most frequently occurred lexical errors in the five groups of students. Richards (1974) categorized these intralingual errors under the term *overgeneralization.* Overgeneralization is a device used when the items do not carry any obvious contrasts for the learner. Ignorance of rule restrictions occurs when rules are extended to contexts where in target language usage they do not apply. Overgeneralization as a strategy of the learner's hypothesis-testing learning process can account for the overuse of some derivational affixes in the present study. Look at the following concordances:

1. ties [cc1,-1] producers and salers [fm2,0-]. For [pp2,-2] my view, our country
2. important not to do things hurrily [fm2,0-]. For example, when we studying, [s
3. e commodities and their productors [fm2,0-] for a long time. The "Ten-thousand
4. rt and the death rate is comparely [fm2,0-] high. People realized its seriousn
6. tch a success [cc3,2-]. Contrastly [fm2,0-], if you raise a purpose [cc3,2-] t
7. tly, [sn8,s-] I will also unendure [fm2,0-] it. My idea [wd3,-1] life should b

The above concordances show that the learners overuse the prefixes *-er/-ly/or/* and the suffixes *–un.*

The second frequently occurred derivational errors are marked as *fm2,CA* which are relevant to other features of interlanguage. Ellis (1985) points out that learner language is permeable and dynamic, in the sense that rules that constitute the learner's knowledge at any one stage are not fixed, but are open to amendment. Learner language system is constantly changing. This characteristic of interlanguage can explain why the confusion of affixes occurs so frequently among the five groups of learners in CLEC. Of course, the derivational complexity of the target language is also a cause of errors. For example, the prefixes for negation is very troublesome and demanding for the learner to master. As a result, the confusion of such prefixes take up a considerable amount of the category *fm2,CA.* The following concordances contain some examples:

1. fore doing, you become uncarefully [fm2,CA3-], and all this makes your [wd3,1
2. is that some have a bad acceptment [fm2,CA2-][cc4,1-] and they don't like to a
3. [fm2,-] proposition and uncorrect [fm2,CA1-] assay will often be projected in
4. his mathod [fm1,-] is unconvenient [fm2,CA1-], but it enables students to know
5. e spend much money buying unuseful [fm2,CA3-] commodities but also do great ha

In the above concordances, the learner's intended words are *careless, acceptance, incorrect, inconvenient, useless* rather than *uncarefully, acceptment, uncorrect, unconvenient, unuseful.*

The derivational errors tagged as *fm2,PS* and *fm2,Se* are caused by the learner's incomplete lexical knowledge. In Laufer's opinion (1997), knowledge of a word include its form, word structure, syntactic pattern, meaning, lexical relations of the word with other word, and common collocations. Jiang (2000) holds that second language learner's lexical representation are quite different from native speakers. In his view, an important feature of the lexical representation in first language is that the different types of knowledge mentioned by Laufer are highly integrated within each lexical entry, such that once the entry is opened, all the information automatically becomes available. The presence of these different kinds of information in the lexical entries and their automatic activation are critical for the appropriate and efficient use of these lexical entries in language production. Due to the lack of sufficient, highly contextualized input in the target language and the presence of an established conceptual/semantic system of L1 vocabulary, the learner's lexical representation is incomplete and cannot open automatically with all the lexical knowledge

available. In our study, some learners only know that a word can be used as a certain part of speech, and their poor command of word class results in derivational errors. The following concordances are examples of this kind of errors:

1. ugh the society is changing fastly [fm2, PS2] and we should learn different kn
2. they can be aware of the harmness [fm2, PS] and not to produce again. [se8, s]
3. [vp3, 1-] to change the job oftenly [fm2, PS]. Because they think differen job
4. a job is not only a challengement [fm2, PS] but also a chance to testing [vp9
5. dn' t bring our country advicements [fm2, PS], but also destroied [fm2,] [cc5,
6. reason that caused these changment [fm2, PS] exsits [fm1,] in pointment [fm2,

The typological differences between English and Chinese have influence on Chinese EFL learners' use of derivational affixes. We have mentioned in the first part that Chinese is an isolating language lacking morphological changes, while English is a morphological language for which derivation is a powerful way of word formation. In addition, the writing system of Chinese characters is quite different from English words. If Chomsky's theory has its sense and the universal grammar is still available, the Chinese learners have to reset their hypotheses about all the aspects of word formation of a foreign language like English. When learners want to retrieve a word, the basic form of a word in their mental lexicon is checked first. Then the system of derivational affixes is scrutinized to make the decision on which affix is suitable for the specific language context. The spelling consideration may be the last step of a word production. Such a process is painstaking and demanding which requires learners' continuous attention and takes up considerable space of the short-term memory. Furthermore, the word retrieval and production process is interfered with learners' first language. An affix may have more than one variants and its spelling is different depending on the morpheme it is added, which will increase the learners' memory load and lead to such spelling errors as the ones in the following concordances:

1. come up with series of solvations [fm2, Sp-] . Nowadays, not only governments
2. ce, we must improve our efficision [fm2, Sp-] of agriculture and industry so th
3. m1, -] and serious. The expandation [fm2, Sp-] of population and the development
4. s [fm1, -] , you will see happyness [fm2, Sp-] or sadness, pride or absesement [
5. we are [vp6, -] punished terriblely [fm2, Sp-] . <ST 3> <SEX 2> <Y 8> <AGE 19>
6. xpectacy [fm1, -] altered noticably [fm2, Sp-] . The second reason is the improv
7. the adventage [fm1, -] of immedicay [fm2, Sp-] while the newspaper is far less c

5. Implication

In the above sections, we have analyzed the features of Chinese learners' acquisition of English derivational affixes and explored the causes of their derivational errors by using the data provided by CLEC. The study of the acquisition of derivational affixes can provide insights into the relative importance of morphology teaching in SLA. Knowledge of processes underlying the learner's use of derivational affixes may support teaching, as it will make clear on which areas of derivational affixes teaching should concentrate and will help determine the best way of teaching them. As for Chinese EFL learners, the chief difficulties lie in overgeneralization and incomplete lexical knowledge. Therefore, the restriction of the derivational rules should be illustrated clearly to them. Secondly, this study can support the work that is being done in the area of vocabulary acquisition. As many words are related by form, studying the nature of these relations may shed new light on the processes and factors that are relevant to

the acquisition of vocabulary. Thirdly, the study of L2 derivational affixes may contribute to general theories of second language acquisition. The findings in the field of derivational affixes could be generalized to other fields.

6. Conclusion

The present study has investigated the acquisition of derivational affixes of Chinese EFL learners and analyzed the causes of derivational errors in CLEC in the light of interlanguage, error analysis, vocabulary acquisition and language typology. In addition, the implication is discussed with the hope of throwing light on vocabulary teaching and second language acquisition. We should admit that this study is a cross-sectional one that only takes into consideration the learners' productive ability of derivational affixes. Their receptive ability remains unknown. More studies can be made in a longitudinal way, and both the productive and receptive abilities of the derivational affixes should be investigated to have a complete picture of the acquisition of derivational affixes.

References

Ellis, R. 1985. *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

Haastrup, K. 1987. 'Using thinking aloud and retrospection to uncover learner's lexical inferencing procedures' in C. Farch and G. Kasper (eds.): *Introspection in Second Language Research.* Clevedon, Avon: Multilingual Matters.

Hatch, E. and C. Brown. 2001. *Vocabulary, Semantics and Language Education.* Beijing: Foreign Language Teaching and Research Press.

Jiang, N. 2000. 'Lexical representation and development in a second language'. *Applied Linguistics21/1*: 47-77.

Laufer, B. 2002. 'What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words' in N. Schmitt and M. McCarthy (eds.): *Vacabulary: Description, Acquisition and Pedagogy.* Shanghai: Shanghai Foreign Language Education Press.

Nagy, W. E., and R. Anderson. 1984. 'The number of words in printed school English'. *Reading Research Quarterly 19*: 304-330.

Rechards, J. (ed.). 1974. *Error Analysis.* London: Longman.

Selinker, L. 1972. 'Interlanguage.' *International Review of Applied Linguistics X*: 209-230.

黄远振，2001，词的形态理据与词汇习得的相关性，《外语教学与研究》第33卷 第6期。

陆国强，1983，《现代英语词汇学》。上海：上海外语教育出版社。

周维杰，2000，论英语词缀，《西安外国语学院学报》 第8卷 第4期

A Corpus-based Analysis of the Reportage of *SARS*[1]

Gao Chao
Henan Normal University

Abstract: The reportage of SARS has been studied based on the mini-corpus specifically designed. The interrelationships between the theme and key words, between the key key words and their associates and clusters have been examined; moreover, the linguistic characteristics of the word use have been demonstrated. The results suggest that: first, there is a close relationship between the use of lexical words and the representation of the topic; the key words which are often organized in clusters and used in association mainly represent the theme in terms of subject, time, place, the cause, the solution and the impact of the events described. Second, the linguistic feature analysis has shown that the reportage of SARS is highly creative and productive lexically. Third, other linguistic features are also examined, such as coinage, shortening words, words with semantic shift, non-common words or non common sense of common words, medical words, affective words, culture-loaded and nativised words and expressions. Finally, the "*noun...n + noun*" structure is preferred in the reportage. This study may be helpful in language learning and dictionary compilation.

Key words: SARS, key words, key key words, associates

1. Introduction

SARS (Severe Acute Respiratory Syndrome), a new communicable disease, spread in the world in the first half of 2003. According to the WHO, during the SARS outbreak from February to July, a total of 8,437 people worldwide got infected with SARS; of these, 813 died. In fact, SARS caused a general anxiety in the world and brought changes in people's discourse in terms of new-born lexis and expressions in linguistics. The research is much justified by the prevailing frustrations and lasting effect on people's mind. In addition, the rapid development in corpora provides great possibilities and potentials for better description and understanding of linguistic feature under investigation; and the corpus-based approach takes the advantage of "computers' capacity for fast, accurate and complex analysis and the extensive information about language use found in large collections of natural texts from multiple registers" (Biber et al., 1998, p. 233). Therefore, the corpus-based approach has made it possible to conduct new kinds of investigations into language use. So the paper tries to find out the interrelationships between the theme and key words, between key key words and their clusters and associates (Li, 2003, p. 283-293), and attempts to analyze the linguistic features of the word use in the reportage of SARS.

2. Research Methodology

To analyze the reportage, an observed mini-corpus of news articles from *China Daily* about SARS was constructed, of which the texts have similar style, subject and text length. The corpus contains 106 independent texts and totals 47,258 tokens sampled from May and July respectively. The reference corpus

[1] I am greatly indebted to my supervisor Dr. Li Wenzhong for his help in the revision for the paper.

used is British National Corpus, which includes over 4000 sample texts of modern British English and totaling more than 100 million tokens. The corpus concordance software used is wordsmith Tools, a user-friendly and powerful package developed by Mike Scott.

3. Analysis of Research Results

Based on the keyword list and keyword database, the observation is made for thematic study and linguistic feature analysis.

3.1. Thematic analysis

About 486 key key words were extracted, e.g. *SARS, China, year, fever, coronavirus, patients, measures, impact* (table1). The theme has been analyzed from eight aspects—the subject, place, time, the characteristics of the subject, the pathogen, patients, measures, and the impact of the event.

A list of Key key words in SARS reportage (part)

N	word	Of 104	As %	N	word	Of 104	As %
1	SARS	86	82.69	11	carriers	1	0.96
2	Beijing	23	22.12	12	patients	9	8.65
3	China	21	20.19	13	cases	5	4.81
4	first	1	0.96	14	deaths	2	1.92
5	half	1	0.96	15	panic	2	1.92
6	year	12	11.54	16	measures	2	1.92
7	infectious	1	0.96	17	prevention	5	4.81
8	fever	2	1.92	18	quarantine	3	2.88
9	coughing	2	1.92	19	impact	2	1.92
10	coronavirus	1	0.96	20	psychological	1	0.96

(Table1 shows the text number and the percentage of the reoccurrence of the key words in 104 different key wordlists.)

1) The Subject: The key key word ***SARS*** (K[1]=9,376.2; p=.000000) and its associates *health/ patients/ disease/ outbreak/ hospital/ epidemic/ cases/ medical,* (table 2), and its three-word clusters such as *the SARS crisis, SARS-hit countries, SARS-affected areas,* show that SARS was a new epidemic and hit over 30 countries and areas in the world.

The associates of *SARS* (part)

n	word	No.of Files	As %	n	word	No. of Files	As %
1	SARS	86	100.00	10	disease	8	9.30
2	Beijing	22	25.58	11	outbreak	6	6.98
3	China	18	20.93	12	hospital	6	6.98
4	health	9	10.47	13	tourism	6	6.98
5	patients	9	10.47	14	epidemic	5	5.81
6	economy	9	10.47	15	Guangdong	5	5.81
7	medical	5	5.81	16	measures	2	2.23
8	prevention	4	4.65	17	quarantine	2	2.23
9	exports	4	4.65	18	deaths	2	2.23

(Table 2 demonstrates the associates of SARS and text number and percentage of the occurrence of associates.)

2) Words indicating place: *China* (K=1,238.1; p=.000000), *Beijing* (K=1,632.6; p=.000000), *Guangdong* (K=479.4; p=.000000), *Xiaotangshan* (K=138.5; p=.000000), *Hong Kong, Taiwan* and *Singapore*. The

[1] Keyness is shortened as K in the paper.

countries most infected by SARS have been China, Hong Kong, Singapore, Taiwan and Toronto, and China's Beijing and Guangdong are the most SARS-hit areas. Xiaotangshan hospital is a special hospital devoted solely to SARS patients on the Chinese mainland. '

3) Time: The key key words *first* (K=65.3; p=.000000), *half* (K=96.1; p=.000000), *year* (K=390.2; p=.000000), *May* (K=68.2; p=.000000), and the key word *April* (K=137.3; p=.000000) indicate the time of SARS outbreak. It is said that the first SARS case was reported in November, 2002, but SARS mainly spread in the first half of 2003, especially in April and May.

4) The features of SARS: The key words *infectious* (K=96.9; p=.000000), *communicable* (K=32.5; p=.000000), *contagious* (K=32.0; p=.000000) show the nature of SARS. The key key words *fever/ coughing/ sneezing* indicate the symptom of the disease (In general SARS begins with a high fever, or headache; and after two or seven days, SARS patients may develop a dry cough). Other key words such as *touching/ mouth/ nose/ respiratory/ droplets/ air* demonstrate the ways that SARS seems to spread (The virus that is thought to be transmitted by respiratory droplets can spread through air when an infected person coughs or sneezes and the virus may also spread when a person touches an object contaminated with infectious droplets and then touches his mouth, nose and eyes).

5) Pathogen: The key words *coronavirus* (K=86.6; p=.000000) / *coronaviral, civet* (K=34.7; p=.000000), *carriers* (K=30.5; p=.000000) indicate the possible cause of SARS. It is reported that about 80% SARS cases have been estimated to be cases of coronaviral infection and civet has been clinically proven to be a carrier of the coronavirus, which shows the incompatible relationships between eating habits and hygiene rules, and between human beings and the natural environment.

6) Patients: Through the analysis of the key key word *patient* (K=368.2; p=.000000), *cases* (K=246.7; p=.000000) and key words *reported/ suspected/ probable/ unconfirmed/ confirmed,* we classify the SARS cases into four categories — reported cases, suspected/unconfirmed cases, probable cases and confirmed cases. The associates of *patients*: *SARS/ hospital/ medical/ deaths* and other words *suspect/ confirm/ panic/ death/ victim/ quarantine/ isolation/ die/ rescue,* show us the terror atmosphere and tense situation during SARS days.

7) Solution: From the concordance lines on the search word *measure* (K=144.8; p=.000000), we found that: the adjectives that modify *measures* mainly are *appropriate/ immediate/ decisive/ drastic/ effective/ harsh/ practical/ precautionary/ preventive/ sudden/ forceful* in turn and nouns are *prevention and control/ quarantine/ screening*; Other important words such as *facemask/ goggles, bleach/ disinfect/ isolation* have shown the specific measures adopted by the government to prevent the disease. In fact, the governments went all out to prevent and control SARS, even took harsh measures against those who denied to be quarantined.

1. xiaotangshan. They only said that '**appropriate measures**' would be taken shoul
2. nterview. Han said the **prevention and control measures** remain effective. But
3. sease. China has taken **immediate and decisive measures** to contain the spread
4. ovincial capital of Nanjing to take other **drastic measures** to control the spread o
5. l governments and departments to take **effective measures** to ensure the transpor
6. inese Government, who took surprisingly **harsh measures** against those who ne
7. "The punishment decisions, taken as **practical measures** by China's new gener
8. had put in place all the necessary **precautionary measures'** and identified back-
9. nday gave a positive comment on the **preventive measures** that the Macao Speci
10 ai Chee,said it is considering lifting **quarantine measures.** Bank spokesman Mi
11 n board have occurred since effective **screening measures** were introduced. Q:
12 t SARS, said they were surprised by the **sudden measures.** First the governmen

8) SARS *impact* (K=109.5;p=.000000): the key words *negative/ hit/ ravage/ dent/ slow/ loss, tourism/ retail/ sales/ transportation/ exports/ consumer/ service* show the negative effect of the disease: SARS mainly hit 6 sectors--tourism, exports, the consumer market, retailing sales, transportation and service industry. Other words such as *rapid/ economy/ insurance/ advertising/ growth/ external/ temporary* indicate the positive effect of the event: SARS effect is external and China maintains strong economic growth; in particular, the insurance and on-line advertising industry benefit from SARS. The words *psychological/ serious/ crisis/ trauma* demonstrate the psychological effect of SARS: due to SARS outbreak many people have developed psychological troubles.

Through the thematic analysis it is observed that: first, there is a close relationship between the word use and the representation of the topic; second, there is a relationship of association and co-occurrence between the key key words and other key words; third, some key key words are associates of each other; finally, the key words which are often organized in clusters and used in associations in the reportage of SARS mainly represent the theme presentation with regard to the subject, time, place, the cause, solution and the impact of the event.

3.2. Linguistic feature analysis:

On the basis of statistics and concordance analysis, the linguistic features that characterize the reportage of SARS are classified into ten categories:

1) Coinage: e.g. *SARS, netease.*

2) Semantic shift: e.g. *SARS* (sense1: a kind of communicable disease; sense2: smile and retain smile), *suspect/suspected* (if someone is suspected, maybe they are infected with some dangerous diseases such as SARS, but not confirmed; a "suspect" case of SARS is defined as a person who fits into one of the following two categories: a person who develops fever and one or more respiratory symptoms within ten days of returning from the areas where SARS cases are being reported, or a person who develops fever and one or more respiratory symptoms within ten days of having bad close contact with a "probable" case.), *probable* (a probable case means that the patient is likely to be infected with SARS virus; it is reported that probable cases of SARS often have a more severe illness, with progressive shortness of breath and difficulty breathing and in some cases, chest X-rays show signs of atypical pneumonia), *confirm/ confirmed* (if someone is confirmed, they are definitely infected with SARS), *screening*(*To screen* for SARS means to examine people to make sure whether they are have it or not.), *quarantine*(if a person is quarantined, he or she is being kept separate from other people for 10-14 days because they have or may have SARS disease; according to concordance analysis we find that there is a heavy use of *10-day, 14-day, home* and *SARS* on the left 1 position of *quarantine*, which shows that the period of SARS quarantine is 10 - 14 days or so and home quarantine is a kind of preventive measure).

3) Chinese borrowings: e.g. *xiaotangshan/ yuan/ renminbi/ huangqi/ yuyingcao.*

4) Shortening: e.g. *HKSAR, NIH* (National Institutes Health), *GDP, WHO.*

5) Nativised clusters and expressions: e.g. *daily necessity, prevention and control, quarantine and isolation, the united will of people.*

6) There is a preference for the non-common words: e.g. *coronavirus* (appearing once in BNC), *civet* (6 in BNC), *communicable* (54 in BNC), *precautionary* (136 in BNC), *quarantine* (147 in BNC).

7) There is another preference for using non-common senses of the words: e.g. *develop* (if you develop an

illness, you became affected by it--the 8th sense according to *Collins Cobuild English Dictionary*), *tract* (a system of organs and tubes—the last item of sense in Cobuild dictionary), *hit* (if something hit a thing or a place, it affects them very badly-- the 4th sense).

8) There is a heavy use of medical words: e.g. *virus/ carrier/ tract/ respiratory/ serum.*

9) There is frequent use of affective words: e.g. *hit/ fight/ forceful/ precautionary.*

10) There is a tendency of using "*noun...n + noun*" structure: e.g. China disease control center/ Singapore Computer Systems Spokesman/ Singapore Tourism Board.

On the basis of linguistic analysis, it is conclude that the reportage of SARS is highly creative and productive lexically. New words are added to English lexicon when the new situation arises; new concept is given to an old word form, thus the meaning of a form is multiplied. Other linguistic features are also examined in the reportage, for example, there is a preference for non common words, medical words, affective words, culture-loaded words and nativised expressions.

4. Conclusion

The corpus-based analysis on the reportage of SARS demonstrates the interrelationship between the word use and the representation of the subject-matter and shows us the linguistic features. The study is helpful for language learning and dictionary compilation. Vocabulary learning will be more effective if the students can focus on the clusters and associates in stead of a single word, can deeply understand the semantic extending or shrinking of a word and realize the significance of culture-loaded and nativised words.

References

Biber, D., Conrad, S., & Randi, R. (1998). *Corpus Linguistics*. London: Cambridge University Press. Beijing: Foreign Language Teaching and Research Press. 2000.

李文中, 2003, 基于英语学习者语料库的主题词研究[J]. 现代外语(3): 283–293.

Usage Contrast of *a * of NP* between the Writings of Chinese EFL Learners and Native Speakers[1]

Wang Fang
Henan Normal University

Abstract: From a corpus-based analysis, it has been found that the nominal structure *a * (NP1)of NP(NP2)* is frequently used in the writing of both Chinese students and native speakers, as it covers about 7% of the use of *of* in unprofessional written part of ICE-GB and 11.6% in the student writing sampled from COLEC, while the word *of* always stands among the top five in the word lists of all text genres. Here a usage contrast of a * of NP is made between Chinese students and native speakers, and the differences found are mainly as follows: 1). In an endocentric structure, with NP2 as a headword, and NP1 as a number or conventional measure to modify NP2 in various ways; It is found that there exists significant difference in the usage of this category between the Chinese students and native speakers. The main reason is probably Chinese students' overuse of the phrase *a lot of*, which covers 92% of the total occurrences of this category in COLEC, while native speakers tend to use more various modifications instead of *a lot of*. 2). In an exocentric structure, with neither noun seeming to be pivotal or dominant, and when NP1 is a noun derived from a verb, NP1 and NP2 are understood as being in a "verb-subject" or "verb-object" relationship: In this category, native speakers not only use the structure a lot more than Chinese students, but use more varieties of NP1, which makes the co-selections of NP1and NP2 become more diverse; while Chinese students tend to use an equivalent clause or a verb-form. 3). Generally, Chinese students tend to use NP1 and NP2 as concrete words while natives more abstract ones. Besides, in the NP2 part, Chinese students usually use a single noun without any modifications, while native speakers tend to use a noun phrase, with more adjectives coming in front of the noun. The findings will inform the students of the relevant information on the usage and provide opportunities to exercise on it.

Key words: corpus, standardized frequency, comparison, difference

I. Introduction

*a *(NP1) of NP(NP2)* is a frequently used structure in English, and there are hundreds of varieties of this structure in the writing of both Chinese ESL learners and native English speakers, for a corpus study based on the student writing sampled from COLEC(Yang, 2002, p. 62) and the non-professional written part of ICE-GB(Great Britain part of International Corpus of English) shows that the use of this structure covers almost 7% of the usage of *of* in ICE-GB and 11.6% in COLEC(College Learner English Corpus), while the word *of* always stands among the top five in all texts genres.

This paper firstly makes a firm classification of the usage of the structure *a * of NP*, and then, through coding, analysis and generalization of each concordance line in the two corpuses, finds out the significant differences on the usage of this structure between Chinese ESL learners and native speakers. The resulting reasons of these differences and implications in China's relevant English teaching are also considered.

[1] Thanks to Dr. Li Wenzhong for his instruction on the paper and valuable suggestions on its various drafts.

II. Classifications of the usage of the structure *a * of NP*

John Sinclair (1991) made a firm classification of the usage of the word *of*, a part of which can also be applied in the usage of the structure *a * of NP* here, despite of some overlaps though:

1. when NP2 appears to be a headword in the whole structure, while NP1, as determiners, numbers, etc come in front of the noun and modify its meaning in various ways, and the whole structure tend to be endocentric:

   eg: a lot of power, a range of alternatives, a kind of maturity

   However, further strands of categories can be made according the different types of NP1 and its relations to NP2:

   1). When NP1 is a number or a conventional measure:

   eg: a lot of time, a number of products, a quarter of such investment.

   2). When NP1s are some more lexically rich partives and qualifiers, which do not require special justification but indicate that this category, like most, has uncertain boundaries:

   eg: a group of young artists, a cup of coffee, a degree of concern, a layer of silk

   3). When NP1s are specifying some part of NP2 or on a component, aspect, or attribute of NP2:

   eg: a study of American football players, a rest of the evening, a list of kings

   4). When NP1s are seen as offering some kind of support to NP2, rather than just specifying some relevant aspect of NP2. Always NP1 are more general and reduced in meaning, while NP2 have more restrict selections:

   eg: a kind of maturity, a question of origins, a matter of his own choice, a sense of unity, a behavior of cheat.

2. In most of the above cases, NP2 will be accepted as the headword, but there remain many cases where neither noun seems to be pivotal or dominant, and where the structure is exocentric and simply requires both of them:

   eg: a disorientation of time, a unity of different characters.

   Still further categories can be made according to the co-selection of NP1 and NP2:

   5). When NP1 are nouns derived from verbs, and the NP1 an NP2 are understood as being in a "verb-subject" or "verb-object" relationship:

   eg: a change of the lithotomical rock, a choice of wrong intervals, a loss of money, a confession of yourselves.

   or NP1 can be a noun derived from an adj:

   eg: a possibility of another economic boom

   6) When *a * of NP* is an alternative way of stating that NP2 possesses NP1:

   eg: a baby of her own, a son of their neighbors

   7) When NP1 and NP2 are all concrete words and NP1 is a cell of NP2:

   eg: a member of a group, a citizen of the country

   8) When NP1 and NP2 are all abstract words and NP1 are some kind of restrictive descriptions of NP2:

   eg: a consequence of these movements, a feature of the sensual celebration, a population of organisms, a standard of goods quality

   9) When the structure appears to be a NP1 of doing, or NP2 are nouns derived from verbs:

   eg: a method of escape, a process of learning

III. Result analysis

Table 1: percentage of each category in ICE-GB and COLEC:

classifications		ICE-GB	COLEC	p value
Focus noun	1 *	12.5	58.56	.000
	2	18.75	15.74	7.910
	3	7.14	3.86	2.200
	4	11.6	8.83	4.857
	5 *	23.2	4.41	0.038
	6	1.78	1.65	0.750
Double-headed	7	0.89	3.31	0.883
	8 *	15.1	1.93	.000
	9	8.92	1.65	0.280

Table 1 shows that there are significant differences between Chinese ESL learners and native speakers in the above three categories. Further analyses on these three categories are considered here:

1.In category 1: when NP1 are conventional measures.

Table 1 shows that in this category, Chinese students use a lot more than native speakers, while a further calculation shows that the usage of the top phrase *a lot of* in COLEC covers 92% of this category, and the top phrase *a number of* in ICE-GB 64%. Then the usage of these two phrases (mainly from the perspective of collocations) worth comparing between Chinese students and native speakers:

Table 2: Collocations and standardized frequencies of the top phrases in ICE-GB and COLEC:

top phrases		ICE-GB		COLEC	
		collocations (lemma)	frequency	collocations (lemma)	frequency
a lot of	1	time	2.29	money	14.75
	2	power	2.29	time	8.05
	3			problem	7.38
	4			trouble	6.71
a number of	1	reason	4.59	book	0.67
	2	angle	2.29	man	0.67
	3	definition	2.29	praise	0.67
	4	factor	2.29	products	0.67

From this table we find it seems that native speakers don't have too much interest in the usage of *a lot of*, which is followed only by two words: time and power, while Chinese students seem to be crazy for it: they put almost everything after *a lot of*, which probably is because Chinese students are too familiar with this phrase, and it can be used conveniently without the danger of making mistakes. But, in fact, in most cases, it can't make a firm and delicate description of the objects concerned. So here when further concordances are made on the high-frequency collocation words in ICE-GB, it is found that the native speakers have more varieties of modifications of these words, for example, they will use *numerous problems, various problems; considerable amount of time, extended periods of time and long(er) time* instead. These words are obviously more exact and delicate than *a lot of*. As for the usage of *a number of*, the difference mainly lies as: Chinese students tend to use more concrete words after *a number of*, while native speakers usually put words that carry abstract concepts instead. This is possibly because low-level ESL grammar books tend not to cover abstract words, but introduce words that have more concrete meanings, which, actually, are much less common for natives.

2. Significant difference also exists when NP1 are nouns derived from verbs, as shown in category 5:

Table 3: Standardized frequencies of category 5 in ICE-GB and COLEC:

	ICE-GB	COLEC	P value
Varieties of NP1	52.73	6.03	.000
Occurrences in this category	59.61	10.73	.000

Table 3 shows that in this category, native speakers not only use this form of structure a lot more than Chinese students, but use more varieties of NP1, and thus the co-selection of NP1and NP2 are more diverse. This is probably because Chinese students tend to use a clause or a verb-form instead of the noun-form verb, the latter, actually, can make our descriptions simple. For example, a Chinese student writes as follows:

So we should choose proper activities to take part in, because it can do good to our study.

It will be better if we say:

A choice of proper activities will do good to our study.

Another example:

If one often changes his jobs, he will spend a lot of time on learning how to do a new job.

It will be better if we say:

A frequent change of jobs will waste a lot of time on transitions.

3. As for the finding in category 8, it echoes the analysis of the collocation of *a number of*: Chinese students tend to use more concrete words, while native speakers more abstract.

Besides the above cases, during the process of encoding the whole copra, it is found that in the NP2 part, Chinese students, in most cases, use a single noun without any modifications, while natives tend to use a noun phrase, with more adjectives coming in front of the noun, which makes the descriptions more adequate and complete.

IV. Implications

The previous section demonstrates mainly the differences between native speakers and Chinese ESL learners on the usage of a * of NP, which can possibly form some new emphasizes for the teaching of this structure: Firstly, teachers should make the students raise the awareness of avoiding awkward and too general description, instead, they should try to increase the varieties of NP1 to make a more restrictive co-selection of NP1 and NP2. Secondly, students should be informed that NP1 and NP2 can not only be concrete words, but can be words that carry fairly abstract concepts. Thirdly, students should learn to use noun-form verbs as NP1, instead of composing equivalent clauses. In such a way, the sentence structure can be lightened with a more simple form. Finally, students should be encouraged to use some modifications of nouns to make the descriptions more adequate and vivid.

References:

Sinclair, J. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press, 1991: 81-98.

李文中，濮建忠，2001，语料库索引在外语教学中的应用[J]。解放军外国语学院学报，(2)：20-25。

杨惠中，2002，语料库语言学导论。上海：上海外语教育出版社。

Problems in Chinese Learners' Use of the English Existential Sentences

Lei Xiuyun
Shanghai Jiaotong University

Abstract: The existential structure is an important item in English study. This paper examines the existential sentences in the Chinese Learner English Corpus, looks at mistakes Chinese learners of English make and attempts to interpret what these mistakes indicate and why Chinese learners tend to make such mistakes.

Key words: existential sentences; Chinese Learner English Corpus

I. Introduction

The existential sentence is a very active sentence pattern. Equivalents of the English existential sentence which represent the existence or non-existence, occurrence or non-occurrence of something in some place can be found in nearly all languages. Basically, the existential sentence in English has the following structure:

There + existential verb + indefinite NP

Where the existential verb is, more often than not, the verb *be*. It can be also other verbs, such as verbs denoting existence or position, e.g. exist; verbs denoting motion or direction, e.g. come, go etc., verbs denoting the occurrence, development or actualization of events, e.g. arise. The noun phrase following *be* is usually indefinite and is referred to as the notional subject and existential object. To present new information is believed to be the main discourse function of existential sentences.

The English and Chinese existential sentences have many similarities. Typically the Chinese existential sentence has the following structure: NP1 + VP + NP2 (e.g.: 树上有只喜鹊.), where NP1 shows the direction or position and is usually a definite NP; VP is usually a punctual verb denoting motion; while NP2 is the agent and most likely an indefinite NP.

The English existential sentence involves many theoretical problems, one example of which is the interpretation of the function and meaning of *there*. The theoretical exploration of the form and meaning of the existential sentence structure is still going on and well attended. In daily English teaching, however, it is generally believed that the existential sentence doesn't present much difficulty to the students, and therefore, the structure tends to be left to itself after being introduced to the students.

An examination of the existential sentences in Chinese Learner English Corpus (CLEC) shows that usage of the existential *there* is not so easy to the learners as many people may have expected. A variety of errors show up in the corpus concerning the use of the English existential *there*. And the fact that learners have problems in their writing with such a basic structure of English demonstrates the loopholes in our English teaching and learning at the basic level and shows some defects in the writing abilities of Chinese learners of English represented by the corpus.

II. Methodology

The present paper focuses its attention on the existential sentences in the CLEC. Chinese Learner English Corpus contains about 1 million words of running text with the following construction:

Table 1: Composition of the CLEC

	Number of Words	Source
St2	208,088	middle school students
St3	209,043	first and second year non-English majors
St4	212,855	junior and senior non-English majors
St5	214,510	first and second year English majors
St6	226,106	junior and senior English majors

In the current research mconcord and wordsmith are the software used to count the number of occurrence of the existential *there* in CLEC or a particular section of the CLEC. Concordance lines of the existential sentences found were classified into different groups according to sentence structure or error type. The outcome was then compared with statistics about the existential *there* in "Longman Grammar of Spoken and Written English" (LGSWE). Statistics in LGSWE were obtained from Longman Spoken and Written English Corpus which has more than 40 million words of text covering such registers as conversation, fiction, newspaper language and academic prose. Whenever needed, the author also referred to the use of the existential sentences in the JDEST (a corpus of academic English) corpus. The outcome of comparison turned out to be very interesting and instructive, showing us, from one particular perspective, what problems there are in our English teaching and learning and toward what we should make our efforts.

III. Analysis and Discussions

1. Overuse of the existential *there*

According to LGSWE, the existential *there* occurs 3000 times per million words on average in conversation and 2500 times per million words in academic prose (LGSWE : 948).The author randomly selected over 1 million words of text from the JDEST corpus and found the number of occurrence of *there* (both existential and locative) is 2130 times.

In the over 1 million words of CLEC, however, the existential *there* appeared 3710 times, higher in frequency than even in the native speakers' conversation. The author also checked the St2, St3 and St6 sections respectively and obtained the following result:

Table 2: Use of Existential *there* in Three Sections of CLEC

	No. of words	No. of ex. *there*	No./million
St2	208,088	897	4311
St3	209,043	753	3603
St6	226,106	679	3003

As can be seen from the table, with the improvement of the learners' English proficiency, the number of existential *there* tends to decrease in the corpus. But even for senior English majors, the use of existential *there* in one million words is still as high as 3003, similar to that in native speakers' conversation and well above the frequency of existential *there* in native speakers' writing. In a word, the results show that Chinese learners of English, even at an advanced proficiency level, tend to overuse the existential *there*.

Authors of the LGSWE believe that, "The reason for the high frequency of existential *there* is no doubt that

it agrees with the looser syntactic organization of conversation: ..... The use of the existential clause makes it possible to present one unit of information at a time. ..... In contrast, the lower frequency of existential *there* in the written registers is the result of planned language production, so that the writer can pack more information into a single clause." (LGSWE : 953)

The fact that the frequency of the existential *there* in CLEC is higher than in native speakers' conversation shows that Chinese learners tend to overuse the existential *there*, which, from one aspect, indicates that Chinese learners of English at this level have great difficulty with the style of written English. Their written English is oral like, and sentences are loosely connected, even broken. This can be illustrated by a sample section from CLEC, in which we see the above mentioned problems as well as other mistakes.

> *Some people have been in the same job for their life. Why?* ***There are all kinds of reasons.*** *For example, some are good at the job, they wouldn't change it. Some think they always work the same surrounding, relationships and companions are familar to them. It is easy to deal with problems in work or life. Furthermore, the others want to change the job, owning to the need of work etc., they have to continue his job. Some people would like to change their jobs constantly. Maybe they like the challeging change. In the addition, maybe they didn't like the old job. Maybe they long for high salary and equal opportunity. They think new jobs are fit for them.* ***I think there are advantages and disavantays in two ways.*** *If you feel present job isn't satisfied or you want to learn different field knowledge, you should change it as soon as possible. On the contrary, if you like present job, you are satisfied with it.* ***You shouldn't change it like*** *others. At last, I hope everyone find a job that you like as soon as possible. You will become a proficient employee.*

From the essays in CLEC we can see that besides those appropriately used existential *there*, there are others that are used too casually and should well be eliminated from formal writing, for instance, the sentence: "Why? There are all kinds of reasons." in the sample above. On the other hand, learners sometimes use the existential *there* in places where it's not appropriate. For instance:

> *There isn't our country's today without this policy.*
> *There are some reasons as follows.*
> *But there's only 3 percent of all the water on the earth we can make use of because 97 percent is sea water or salt water.*

2. Tendency to use only the simple forms of the existential *there*

The existential sentence structure varies mainly in the following two ways: "adverbials which are essential for the meaning of the existential clause (chiefly time and place adverbials) and postmodifying elements in the notional subject" (LGSWE: 949). According to LGSWE, the basic pattern of the existential sentence (there + existential verb + NP) mostly occurs in conversation. In conversation, information is provided to the listener in smaller chunks, and some information is sometimes left for the listener to infer rather than clearly stated. The basic form is, therefore, frequently used. In academic prose, the basic form is found least frequent because sentence structures in written English tend to be more complicated. Now that *there* and *be* in the existential sentence are nearly meaning-empty, it takes other sentence elements to carry the information. These elements include adverbials and postmofidiers of the NP. LGSWE also found that the structure "there + be + NP + adverbial" is more frequently used than the structure "there + be + NP +

postmodifier" in conversation. In written English, things are just the opposite. The latter structure is more frequently used than the former. Based on the statistics given by LGSWE, existential sentences without any expansion, those with postmodifiers and those with adverbials as expansions account for the following percentages of all existential sentences respectively in each register.

Table 3: Use of structural expansions in existential sentences

	CONV	FICT	NEWS	ACAD
No expansion	25%	20%	15%	10%
Postmodifier	23%	40%	50%	50%
adverbial	40%	25%	30%	20%

The author checked sections St3 and St4, altogether 421898 words in CLEC, and found the existential *there* occur about 1200 times, among which the basic form without expansion accounts for about 23%, those with adverbial expansions account for 23% and those with postmodifiers 51%. It can be seen that the non-English major college learners use the simple existential structure much more frequent than native speakers typically do in academic writing. Sentences like the following are easily found in CLEC:

Someone like doing the same job in his life. They don't want to change. However, there are many reasons. First ...

Xinxiang is a industrial city. There are plenty of factorys. They can produce a lot of production which sell to throughout our country.

And the baby's infant mortality was 10percent in 1990 while it was 20 percent in 1960. why did this happen? There are two causes. Firstly, ...

The three types of existential sentences account for 19%, 21% and 55% respectively in St6. The percentage of the simple structure decreased a few points, but is still much higher than in a typical academic prose. Close examination shows that although existential sentences with adverbials or postmodifiers in CLEC are in accordance with native speakers' writing in frequency, their structures are typically simple and thus the sentences short. This is true of the essays at all levels in the corpus. This shows that the learners are still inadequate in their writing skills. Many of their sentences are simple and lack variation and coherence.

3. A high-frequency item like the existential *there* can be error-prone rather than totally safe

The investigation into the existential *there* in CLEC also shows that learners make all kinds of mistakes with the existential structure.

1) Verbs may appear immediately after the NP, for instance:

But there are still many person don't know it.(St3)
There is well known words says..... (St3)
There are a lot of new things wait for us. (St4)
... there are some unfair phenomina existed. (St6)
And it is no more than throwing the money into the sea for there would be no wonder occurred. (St6)

The author believes that transfer plays an important role in this kind of misuse. As we mentioned before, the Chinese existential sentences have the following structure: NP1 + VP + NP2, but it can have a verb phrase as an expansion, for example: 院子里有很多人跳舞. Some Chinese learners must have been influenced by the Chinese version of existential sentence in their use of the English equivalent.

2) Problem of subject and verb agreement

Some students still have problem deciding whether to use a singular or plural verb after *there*.

e. g. We can't deny that until today there is still those people. (St6)
However, there has been no laws in China on this issue, ... (St6)
Although there are still a lot of work to do before the .... (St6)

As a rule, in the existential sentence, the verb usually agrees in number with the NP on its right. There are, however, exceptions. The rule is generally followed in formal English, while in spoken English, even well educated native speakers take *there* as the singular subject of the existential sentence and thus use a singular verb accordingly. In this case, the often used form is "there's". Furthermore, when the NP is composed of two or more coordinate nouns, and the first of them is singular, the verb of the sentence is generally singular no matter it is in written or spoken English. That learners have difficulties using the verb in the right number in this simple structure may manifest the difficulties in their transplanting the concept of number into their Chinese ideological system, since in Chinese, the existential verbs like other verbs do not involve number.

Besides, in CLEC, the following sentences can be found:

If there no fresh water, we all would die.
I fact, there many way for us to get to know.
People always think that there have a lot of fresh water on the earth...

The author feels such mistakes are at least partly due to that the most frequently used Chinese existential verb is 'you' , while its English equivalent consists of two parts and neither part has to do with the Chinese 'you'. So some learners tend to forget one part or use 'have' instead of a form of 'be' for the second part.

3) There + be + definite NP

The NP in the existential sentence is generally an indefinite noun phrase. For instance, the following sentences are regarded as ungrammatical:

* Are there the ten students in the class?
* There is the phone on the bed.
*There's every student in the garden.
*There are most students here.

In some cases, however, a definite NP can appear in an existential sentence. For example:

... there is every possibility that the country would now be wracked by civil war.
Then there was the war and the evacuees, such beautiful children, and my letters to you and yours to me and in the end there was this, Willie, what we have now.

According to LGSWE, definite NP is used in the existential sentence in the following cases:

i. as the opening line of a conversational story
There was this prince.....
ii. to bring something known back to mind instead of asserting its existence
A: You won't get so much for twenty-five pound in Marks and Spencers.

B: Well, it's not that. What do they sell that you want?

A: There's the food place, isn't there? (LGSWE : 953)

In some cases, with the postcopular NP being not indefinite, the interpretation changes. Sentences like "There's the man you wanted to see." are presentational rather than existential.

In CLEC, there aren't many instances of usage of definite NP's in the existential sentence But some definite NP's can indeed be found in some existential sentences in CLEC. For instance:

There are the several reasons. (St3)

I believe where there is the will there is the way. (St3)

There are the friends I've had. (St6)

Such inappropriate usage of definite NP's in existential sentences indicate that learners, on the one hand, use the definite NP without knowing why, and on the other, they lack the ability to use the existential sentence with a definite NP to express their feelings in the right way.

IV. Conclusion

This study conducts a comparative analysis of use of the existential sentences in CLEC and the native speakers' use of them. Results show that Chinese learners, even the senior English majors, are not totally free from problems with this simple structure. This has interesting pedagogical implications because although high-frequency patterns are encountered very early in instructional programs, once they have been taught, they tend to be neglected. This is particularly unfortunate because these patterns are extremely complex and learners are at a risk of having only a very crude knowledge of them. Teaching at the advanced level may still need to lay some emphasis on fleshing out the learners' incomplete knowledge about the very basic things, and raising learners consciousness as to areas in which L1 and L2 do not correspond may be a never-ending task in the learners' language learning career.

References

Biber, D. et al., 2000, 《朗文英语口语和笔语语法》[M]。北京：外语教学与研究出版社。

桂诗春、杨惠中，2003，《中国学习者英语语料库》[M]。上海：上海外语教育出版社。

Register Misuses of *Because* in Chinese EFL Learners' English Writing — A Comparative Study of CLEC[1] and Brown

Dawang HUANG
City University of Hong Kong
Faculty of Foreign Languages, Ningbo University

Abstract: The learning of register knowledge has always been an obstacle in the development of students' writing ability, owing to the multi-faceted display of register variance. This paper, adopting the theoretical framework of Schleppegrell (1996), mainly analyzes the register misuses of *because* in Chinese EFL learner's writing based on the "Chinese Learner English Corpus" (CLEC) with a reference corpus—Brown. It is found that, (1) a surprisingly high frequency of *because* with a ratio of 2.14 per 1,000 tokens is evinced in CLECT-ST3, (in contrast with 0.87 per 1,000 tokens in Brown); (2) within such instances of *because* in CLEC-ST3 colloquialism stands out. More than one quarter of these 'because's are used to initiate independent clauses, the typical structure of whose prior utterances is a single word—'why' or a short 'why' interrogative; (3) generic features of academic writing see frequent violation, similar to those mistakes of American ESL learners on the whole. Chinese EFL learners tend to frequently employ *because* to link clauses whose logic relationships are loosely meaning-based; and (4) the premodifiers of *because* in CLEC-ST3, confined to few degree adverbs, demonstrate an intensifying semantic prosody which might result in readers' misbelief that Chinese learners are over-certain with their claims. All the above four categories of mistakes can possibly be attributed to the negative transfer of Chinese as well as such developmental factors as learners' unfamiliarity with academic norms.

Key Words: because; register; writing; EAP; Learner English

1. Introduction: Grammatical Correctness and Register Features

That a grammatically correct sentence is not necessarily a well-written one is known to concern register-based factors. In other words, people are not expected to adopt formal style in oral communication whilst people would avoid casual style in written communication. Oral English of Chinese EFL learners was once criticized for its written flavor before a nationwide reform on ELT across China (approximately the mid-and late-1980s), but what is the condition thereafter?[2] When correcting (tertiary-level) students' writings in English, we teachers usually feel that, those writings, loosely structured and abundant with simple words, turned to be at a low level in style on the whole (to be exact, having a touch of colloquialism) (Cai, 2000, p.302).

Hyland (1998, p.243) points out that 'students learn most of their discourse knowledge when expressing themselves in writing ...'. College English[3] teaching in China, however, currently allocates little time in specifically training students' writing abilities, and the limited time spent on composition training usually goes to the commenting upon students' essays in mock tests without recurring to the instruction of discoursal features like register.

2.The Study

2.1 Research Framework

In traditional grammar, conjunctions mainly connect words and other linguistic units, and subordinators are strongly expected not to begin a single sentence in written English. However, those conjunctions as discourse markers (including some subordinators) primarily fulfill the functions like pragmatic relations instead of semantic ones, and commonly introduce a complete sentence. This type of usage is typical of oral English, as a matter of fact. By means of their semantic meaning, conjunctions presuppose the existence of certain propositional units: for instance, *because* tends to indicate the existence of 'assertion' and *because* clause provides reason. In addition, the intensity of semantic foregrounding of conjunctions vary along with different discourses; once conjunctions act as cohesive discourse markers, these conjunctions will render pragmatic support to the ongoing communication and the under-construction discourse whilst their semantic function undergoes weakening (Schiffrin, 1987, pp.182-227). Also, the utterances initiated with and preceded by *so* or *because* in the above cases find no discrepancy as regards the degree of importance. Conjunctions in academic writing, however, are generally assumed to be a meaning-based marker, seldom fulfilling their pragmatic functions.

Schleppegrell (1996), following the analytic model of discourse marker (Schiffrin, 1987), explores the misuse of 'because' in the writings of American ESL students. It is found that, in her self-collected database (142 essays altogether) ESL students use two times as many 'because's as do students of native speakers of English, and the writings of ESL students usually show an 'oral' tone. Schleppgrell categorizes the misuses of 'because' in ESL student writings into three groups, all belonging to the typical performance of 'because' in oral style (i.e., violating expectations for academic registers) whilst contributing to an 'oral' tone (1996, pp.275-280)[4]

1) Knowledge-based linking. As an internal conjunction, *because* reflects the rhetorical organization of text, and links a proposition and the speaker's attitude toward that proposition, or provides information about the knowledge base on which the speaker makes an assertion. Such use is common in speech, but typically not in writing. An example of this is given below: *Schedules [in American schools] are flexible because students who don't like history can take geography instead.* In the above sentence, we expect the *because* clause to provide a justification for flexibility in student scheduling. Instead, the writer gives an example of how schedules are flexible.

2) Adding information in independent segments. ESL students sometimes initiate independent clause(s) with *because* to add background or motivate circumstances as they proceed. Although the prior and following clauses of *because* herein are meaning-based, punctuation does provide evidence of writers' perceptions of sentence boundaries and lack of knowledge in registers. For example: *An example of technology taking the place of man exists in the computers. Kids today get recognition for papers or essays they write but they don't get the satisfaction.*

   Because the computer does most of the work in establishing the essay together.

3) Linking larger segments of discourse. With a broad scope in linking the prior and following clauses, *because* often introduces sequential clauses which are not associated with their prior clauses, and simply helps structure the text. For instance: *Finally, I don't think we are robbed out of our 'satisfaction of the technology because let say you are a doctor and because of the technology you are able to help more patients.*

2.2 Research Method

This paper plans to follow the research framework of Schleppegrell (1996) and adopt her three-point analytic model of misuse of *because* in ESL student writings. First of all, I will present a categorization of the *because* data according to their formal features (e.g., the distribution of because within sentences as well as co-occurring adverbials of because). Detailed procedures are listed as follows:

1) Identify the research subject;[5]
2) Retrieve all the data of *because* from CLEC-ST3, a subcorpus of "Chinese Leaner English Corpus" (CLEC) (and from Brown, a reference corpus hereby);[6]
3) Categorize and analyze the data via *Wconcord* and *WordSmith*;[7]
4) Explain the misuse of *because* in CLEC-ST3 under the analytic model of Schleppegrell (1996).

In the course of the categorization of the *because* data according to their formal features, I left the primary work to such concordancing software as *Wconcord* and *WordSmith* but affirm the results at the end through the hand-count of two English teachers. The method for the categorization according to the formal features of *because* is listed below in detail:

Group A: preceded by such punctuation marks as '.', '?', and '!' whilst initiating (an) independent clause(s) (e.g., *As a student in university, we should learn much knowledge. Because we will use them in the future.*);

Group B: preceded by such punctuation marks as '.', '?', and '!" whilst initiating (a) dependent clause(s) (e.g., *Also it is necessary to know the world outside the campus. Because we will enter the society in the future, we must adapt to ....*);

Group C: preceded by the punctuation mark of ',' (e.g., *I started my college life, I looked a newspaper at first, because the newspaper is a way by which I could know the ....*);

Group D: locating within sentences whilst immediately preceded by no punctu- ation marks (e.g., *we graduate from the college. I think we should know about society because it is good to know the difficulties earlier. We can get to ....*);[8] and

Group E: set-phrase of 'because of' (e.g., become family teachers so we can know the society more deeply because of all types of persons. At the same time, we can earn ....).

(Note: The number of *because of* calculated in Groups A, B, C and D will be subtracted.)

2.3 Results and Discussion

Table 1 shows that, the data of *because* incompatible with the norm of academic writing (i.e., Groups A, B, and C) add up to 246 in the column of CLEC-ST3, amounting to 54.89% of all the instances of *because* in CLEC-ST3. In the column of Brown, however, the number of *because* falling into Groups A, B, and C reaches 270, accounting for 30.59% of all the instances of *because* in Brown Corpus. Besides, a significant difference has been found in the above two corpora as regards Group D; that is, as to *because* in the mid of a sentence preceded by no punctuation marks, it appears 111 times in CLEC-ST3 (24.78% of all the instances of *because* in CLEC-ST3) in contrast with 402 times in Brown (45.53% of all the instances of *because* in Brown). These contrasts suggest the overwhelming existence of an oral tone in Chinese EFL learners' writings, violating the expectation of the target discourse community.

Table 1: Distribution of *because* in Two Corpora*

CORPUS / DATA / GROUP	Brown			CLEC-ST3		
	raw freq.	% in *'because's*	‰ in Brown	raw freq.	% in *'because's*	‰ in CLEC-ST 3
A	11	1.25	0.01	124	27.68	0.59
B	86	9.74	0.08	49	10.94	0.23
C	173	19.59	0.17	73	16.29	0.35
D	402	45.53	0.40	111	24.78	0.53
E	211	23.90	0.21	91	20.31	0.44
Total	883	100.00	0.81	448	100.00	2.14
it/this be because+clause	20 (17+3)	2.27(1.93+0.34)	0.02	24 (11+13)	5.36 (2.46+2.90)	0.11
premodifiers+because	72	8.15	0.07	9	2.01	0.04

(*Brown Corpus is of 1,014,312 tokens as reported in *ICAME Corpus Collection* (2nd ed.), and CLECT-ST3 is of 209,043 tokens as reported in Gui and Yang (2003).)

The issue of standardized frequency of *because* has to be noted yet. Comparatively speaking, *because* appears 2.14 times per 1,000 tokens in CLEC-ST3, much higher than the standardized frequency of *because* in Brown – 0.87 times per 1,000 tokens. It indicates the over-reliance of Chinese EFL learners on the use of such an overt maker of cause.

2.3.1 The locus of English reason clauses is fairly flexible though the choice of the initial or mid section of a sentence reflects writers' style/register preference; whereas Chinese reason clauses are typically sentence-initial. Most Chinese grammarians claim that, the coding mechanism of 'yuanwu (object-observing) and quxiang (concept-selecting)' in Chinese shows the iconicity of our physical world: reasons and conditions first while results and conclusions second (Lu, 2001, p.257). Therefore, the overuse of *because* within Groups A, B, and C in CLEC-ST3 can be partly attributed to the negative transfer of the Chinese language. A number of corresponding misuses in CLEC-ST3 are listed as follows:

(1) ... walk into the world outside the campus, to serve the people. Because they have been studying in campus for years, they know very little ...

(2) ... practice. But we do grasp so much words we need. This because we remember these words not so firmly. The word like "yes" .... (Influenced by the Chinese structure of 'zhe4 shi4 yin1 wei2'—'this is because'.)

Another important difference is concerned with the subsequent unit initiated with *because*: English *because* can only bring clause(s) while Chinese *because* (yin1 wei2) is empowered to introduce clauses as well as nominal phrases. The negative transfer of the Chinese language can also be seen in the sentences below:

(3) I must be getting to know the society necessary. Because of I will be going to the society I think that ....

(4) ... world's knowledge. Now I very difference to find a job. Because of it need very high quantity. So we ...

(5) ... the child who want to go to school but hasn't chance because of economic. We should help them to appeal the society.

2.3.2 Table 1 shows that the inter-sentential use of *because* (i.e., Group D) is quite common in CLEC-ST3. However, the logical relationship between prior and following clauses (i.e., X and Y) connected by *because* seems vague in Chinese EFL learners' writings, consequently failing to reflect the notion of subordination. It is related to students' lack of awareness of register difference, but we assume the underlying causes are developmental factors of learners. A detailed analysis of the misuse of *because* among Chinese EFL student writers is to be presented below under the model of Schleppegrell (1996).

Class One: Knowledge-based linking

(6) I am not able to watch TV, because I live in school. *(The comma immediately before 'because' indicates the informal style.)*

Example (6) lays its focus on explaining that the writer cannot watch TV, then providing a foundation for his assertion that he knows little about the outside world. But the reason the writer presents is indirect for the notion of "school dormitories are not equipped with TV sets" is not necessarily a common sense for others. In addition, 'be able to' refers to ability, but the writer hereby wants to express possibility—'cannot'.

(7) The world outside the campus is important to us [A] because we will serve the country and the people after graduation [B].

Similarly, the clauses [A] and [B] within Example 7) are not directly related for [B] provides readers with additional information instead of illuminating the importance of the outside world for students. Example (7) would be acceptable if *because* is replaced by *since*.

Class Two: Adding information in independent segments

(8) But in fact, fresh water is limitted around us. Because more and more people are born, much fresh water industry needs, and ...

The clause introduced by *because* in Example (8) furnishes the prior utterance—'fresh water is limited around us' with a direct and concrete reason. However, punctuation provides evidence of the writer's perceptions of sentence boundaries, linking an inter-sentential relation. From Table 1, the errors of such class in CLEC-ST3 occupy a considerable proportion of 22.7% (in contrast with 1.2% in Brown). We assume that the errors probably result from the negative transfer of oral style in English since it is indeed common to use *because* to initiate independent clause(s) in oral English whilst to maintain a close connectivity with its prior utterance(s).

Class Three: Linking larger segments of discourse

(9) Our universties must get to know the world outside the campus, because we are going to work, we must know the world enough.

In Example (9) *because* extends a large scope of two clauses, though loosely linked with its prior utterance. Furthermore, there exist run-on sentences. The revised version of Example (9) can even remove the overt causal marker and then utilize the implicit causal structure of infinitive: Our universities must also impart students with the knowledge outside campus to better equip the students for their future job.

2.3.3 As mentioned in Section 2.3.2, a sharp contrast between CLEC-ST3 and Brown is found concerning the errors of Class Two, which is worthy of special attention. Amongst Chinese student writers their use of *because* in Class Two usually has a relatively small scope
and the typical pattern is *because* initiating an independent clause whilst closely linked with its short prior utterance. Furthermore, the short utterances are usually composed of *why* or a short interrogative introduced with *why*.

(10) But once you do it skillfully, you will think it is too easy. Why? Because you have found the spirit through

practicing.

(11) Why has this change happened? Because the health of people is getting better.

The above two examples show the inter-sentential use of *because* and place the *because* clauses at the point of information focus. In order to avoid over-stressing reason clauses once and again, we may as well utilize embedded clauses or nominalized structures to integrate the prior and following utterances. Although similar sentences can be observed in Brown, the register in which they occur is evidently fitting—oral English:

(12) Asked why, he replied primly: " Because that's no activity for a gentleman". [Brown1_f.txt]

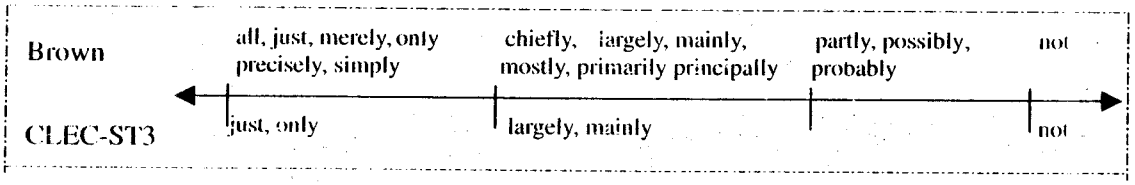
(13) Why? Because your soul was made to be filled with God Himself, not religious functions "about" Him. [Brown1_d.txt]

Example (14) demonstrates that in a formal context, even with a preceding utterance of 'why' interrogative, the following utterance is still a complete sentence.

(14) Why did the Belgians grant independence to a colony so manifestly unprepared to accept it? In one large oversimplification, it might be said that the Belgians felt, far too late, the gale of nationalism sweeping Africa. [Brown1_a.txt]

2.3.4 Lastly, a discussion is to be held on the co-occurring expressions (especially adverbials) with *because* in Chinese EFL learners' writings. Typical premodifiers of *because* in academic writings comprise both enhancing markers (e.g., *only* and *just*) and mitigating markers (e.g., *partly* and *largely*). It is also claimed that deliberate imprecision is a characteristic of academic discourse (cf. Hyland, 1998, p.9).

Table 2: Typical Premodifiers of *because* in Two Corpora



Concurrent Premodifiers	Brown	CLEC-ST3
enhancing/mitigating markers	just (8), simply (7), largely (6), only (6), all (3), primarily (3), precisely (2), chiefly (1), mainly (1), merely (1), mostly (1) , principally (1)	only (3), just (2), mainly (1), largely (1)
	partly (11), perhaps (5), probably (3), maybe (1), possibly (1)	/
	not (11)	not (2)
合计	72	9

Compared with the data of *because* in Brown from Table 2, CLEC-ST3 only has 9 concurrent premodifiers of *because*, most of which are confined to few degree adverbs. What's more, these premodifiers demonstrate an intensifying semantic prosody which might result in readers' misbelief that Chinese learners are over-certain with their claims. It does not mean that the sentences reflecting such an intensifying prosody are grammatically wrong; however, we mean to postulate that Chinese EFL learners are imperfectly aware of the norm of specific lexicons in academic discourses of English and its corresponding variation in registers.

3. Conclusion

Hong Kong tertiary-level student writers were found to overuse certain causal markers, but without discerning *because* with other causal markers like connectives (*as*, *since*, and *for*), compound prepositions (*because of/as a result of*), nouns (*cause/reason*), prepositions (*with/without/by*) and verbs (*cause/result in/contribute to*) (Flowerdew, 1998; Hyland and Milton, 1997). Still, a distinct underuse of such logical connectives as *therefore*, *thus*, *however*, and *yet* can be observed in formal EAP (English for Academic Purposes) of Swedish EFL learners (Altenberg and Tapper, 1998). It shows a choice of registers though these four connectives basically function at the inter-sentential/discoursal level.

Schleppegrell (1996) and this paper unfold that *because*, typically having an intra-sentential use in academic writing, is commonly misused by ESL/EFL students to either function at the inter-sentential level or convey pragmatic senses. This fact nevertheless reflects the option of registers, proving the existence of register misuses of *because* among different groups of EFL/ESL learners.

Register misuses identified in this paper can be attributed to the negative transfer of mother tongue – hereby referring to the Chinese language (locus and chunks of *because*) as well as the developmental factors in the acquisition of English (lexical variety and register knowledge). While this paper primarily aims to identify the use of *because* in Chinese EFL learners' writing, future analysis should consider the pedagogical solution to such register misuses and the longitudinal development of Chinese EFL learners in the acquisition of register knowledge.

ACKNOWLEDGEMENTS

I would like to extend my heartfelt appreciation to Dr. Ken Hyland (City University of Hong Kong), Dr. Mary J. Schleppegrell (University of California, Davis) and Dr. Albert T.-Y. Wong (University of Hong Kong) for their assistance during the writing of this paper. Thanks are also due to my colleagues at the Research Institute of Theoretical Linguistics, Ningbo University, for very helpful comments on an earlier draft of this paper.

NOTES

1. CLEC refers to Chinese Learner English Corpus (Gui and Yang, p.2003). This corpus with a size of approximately 1,000,000 tokens collects the writings of English learners in China at different levels (i.e., middle school—ST2, band four of College English—ST3, band six of College English – ST4, elementary phase of English majors—ST5, and advanced phase of English majors—ST6).
2. Communicative language teaching underwent strong advocation since then in the Chinese mainland.
3. College English in this paper refers to the courses of English for non-English majors in Chinese universities.
4. Three examples listed in Section 2.1 are quoted from Schleppegrell (1996), whereas all the other examples in this paper, if not specified, are taken from CLEC-ST3. Original mistakes are retained but the tagging of errors has been omitted.
5. The reason for choosing CLEC-ST3 as my subject mainly lies in the fact that CLEC-ST3 contains CET4 (College English Test, Band Four) writings of tertiary-level students in the Chinese mainland. CET4 examinees comprise the largest group of English learners at tertiary-level since it is compulsory

for college students of non-English majors to pass CET4 before graduation while CET6 (College English Test, Band Six) is only elective. Also, the data in CLEC-ST3 are topic writings of exposition and argumentation finished within 30 minutes, similar to the characteristics of the database of Schleppegrell (1996) (i.e., with the same style and being time-pressured).

6. All the data of CLEC-ST3 quoted in this paper are taken from the CD-ROM attached with Gui and Yang (2003), while all the data of Brown from the CD-ROM of ICAME Corpus Collection (2nd Version) issued by the HIT Centre of Bergen University, Norway.
7. The software of Wconcord was co-developed by Mr. Zdenek Martinek (the University of West Bohemia, Czech) and Prof. Les Siegrist (Technische Hochschule Darmstad, German), downloadable from_http://www.linglit.tu-darmstadt.de/ wconcord.htm; and the software package of WordSmith, developed by Dr. M. Scott (University of Liverpool, UK), is recognized as one piece of the most powerful corpus software across the world.
8. For example, *because* in '... can play it better and better, just because he can become ....' is classified into Group C while *because* in 'Secondly, because people in the developing country..., ...' into Group A.

REFERENCES

Altenberg, B. & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learners' English on computer* (pp.80-93). London and New York: Longman.

Cai, Jigang. (2001). *A contrastive study of English and Chinese writing*. Shanghai: Fudan University Press.

Cheng, X.G. & Steffensen, M.S. (1996). Metadiscourse: A technique for improving student writing. *Research in the Teaching of English*, 30(2), 149-181.

Ferris, D.R. (1994). Rhetorical strategies in student persuasive writing: Differences between native and non-native English speakers. *Research in the Teaching of English*, 28(1), 45-65.

Flowerdew, L. (1998). Integrating 'expert' and 'interlanguage' computer corpora findings on causality: Discoveries for teachers and students. *English for Specific Purposes*, 17(4), 329-45.

Gui, Shichun & Yang, Huizhong. (Eds.). (2003). *Chinese learner English corpus*. Shanghai: Shanghai Foreign Language Education Press.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam/Philadelphia: John Benjamins B. V.

Hyland, K. & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, 6(2), 183-205.

Lu, Chuan. (2001). *Paratactic networking of Chinese grammar (HAN YU DE YI HE WANG LUO)*. Beijing: Commercial Publishing House.

Schiffrin, D. (1987). *Discourse markers*. New York: Cambridge University Press.

Schlepegrell, M.J. (1996). Conjunctions in spoken English and ESL writing. *Applied Linguistics*, 17(3), 271-285.

Zhao, Weibin. (2003). A quantitative analysis of logical connectors in Chinese ESL learner writings. *Foreign Language Education*, 24(2), 72-76.

An Empirical Study of the Social Constraints on the Lexical Realization of Thanking in English Conversation

Liang Hongmei
South China Agricultural University

Abstract: This article represents a multiple-source approach to an empirical study of the social constraints on the lexical realization of thanking in English conversation. A questionnaire survey among 28 native English-speaking academics and the corpora LLC:c and COLT were employed as the sources of data to explore the social constraints on the lexical realization of thanking. The focus was on the following research questions: 1) What lexical items can be used to express thanking? What are the native speakers' comments on these forms? 2) How can the lexical realization of thanking be constrained by such social factors as speech domain, degree of formality of communicative setting, degree of familiarity between interlocutors, change of time, and speakers' social class, age and gender? The major findings of the study show: 1) The lexical realization of thanking can be diversified and is closely related to the degree of formality as well as the degree of gratefulness. 2) It is also loosely related to such social factors as speech domain, change of time, and speakers' social class, age and gender.

Key words: English conversation, lexical realization of thanking, social constraints

1. Introduction

Aijmer (1996), in her study of conversational routines based primarily on the original version of London-Lund Corpus of Spoken English (LLC:o), has analysed thanking both quantitatively and qualitatively from varying perspectives. Nevertheless, some aspects - such as the relationships of speech domain, the degree of formality and the degree of familiarity between interlocutors to the verbal realization of thanking, the other social constraints such as speaker's individual differences on the lexical realization of thanking still remain understudied. Moreover, since the data of LLC:o were collected mainly in the 1960s and the 1970s, some aspects of the use and realization of thanking may have by now changed. Given these facts, the complete version of the London-Lund Corpus of Spoken English (LLC:c) and the complete version of the Bergen Corpus of London Teenage Language (COLT), both of which consist of approximately 500,000 words; and a questionnaire survey among 28 native English-speaking academics in the UK and the USA(this number does not include the three informants of the pilot survey) have been employed as the sources of data to explore the social constraints on the lexical realization of thanking in the following aspects:

1) What lexical items can be used to express thanking? What are the native speakers' own comments on these forms?
2) How can the lexical realization of thanking be constrained by such social factors as speech domain, degree of formality of communicative setting, degree of familiarity between interlocutors, change of time, and speakers' social class, age and gender?

The social factors to be analyzed are further specified as follows:

a) Five speech domains in LLC:c, viz. face-to-face conversation, telephone conversation, public discussion, answerphone and Dictaphone recordings, and public speeches.

b) Two degrees of formality of setting in LLC:c, viz. formal (public) and informal (casual).

c) Two degrees of familiarity between interlocutors in LLC:c, viz. equal and disparate.

d) The change of time, viz. LLC:c, representing the past, and the questionnaire survey representing the present. The questionnaire survey was conducted two years ago while the materials of LLC:c were recorded from the 1950's to the 1980's. Therefore comparison might be made between them in terms of lexical forms.

e) The social characteristics of the interlocutor, viz. social class, age, and gender. The discussion of the social class as a factor will be based mainly on the data from the questionnaire survey; that of age will be based on the comparison between LLC:c and COLT (only the tokens of teenagers) as well as the information from the questionnaire survey; and that of gender will be based on the data from LLC:c, as well as from the questionnaire survey.

By so doing, we aimed at finding out either one or both of the following:

a) how different social variables might be related to the interlocutor's choice of forms;

b) how different social variables might be related to the frequency distribution of different types of forms.

2. Sources of data

The present study has chosen LLC:c as the main source of data for the reason that two important types of information have been provided in the users' manual accompanying the CD-ROM version: firstly, the information about the texts in terms of speech domains, recording setting (surreptitious or non-surreptitious), years of recording, topics of some of the public conversations and monologues; secondly, the information about the speakers' background such as their sexes, their social roles or professional backgrounds, their ages, and the social distance and power relations between the participants.

COLT, the first large English corpus focusing on the speech of teenagers, was developed by a group of scholars headed by Stenström in Bergen University of Norway. It was collected in 1993 and consists of the spoken language of 13 to 17-year-old teenagers from different boroughs of London. Since the complete version of COLT consists of roughly the same number of words as that of LLC:c, comparison can be made in terms of the forms of thanking between these two corpora to see the possible relation of age to the realization of thanking.

In order to obtain a systematic inventory of the types which are available in English language to express gratitude verbally in conversation and to make up for the limitations of LLC:c, we have conducted a questionnaire investigation among some native English speakers by e-mail. In the questionnaire, we have invited the informants to add more forms to the given list of thanking (complied by the present researcher according to the forms summarized by Stenström (1994:107) and Aijmer (1996:39) in their studies based primarily on LLC:c or LLC:o, and to make the points clearer, I presented the forms in more specified ways in the questionnaire) and to give their comments on the usage of the various forms as well. Up till now, I have got 28 valid replies. The informants are from several universities from the UK and the USA such as Aberdeen University, Manchester Metropolitan University, University of North London, Bristol University, York University. From the responses to the questionnaire survey, we found that the age, and the occupation and education background of these informants are similar to those of the speakers in LLC:c (which represents conversation among educated British speakers, most of them are in the age range of 20 to 70 and have an academic background). Since the questionnaire data was collected two years ago, while the data in LLC: c was mainly collected in the 1960s and 70s, comparison might be made between them.

3. Data retrieval

The Oxford English Software MicroConcord (MCO), designed by Mike Scott and Tim Johns, has been used to retrieve all the instances of thanking in LLC:c and COLT and to sort out and calculate statistics of the questionnaire information.

The retrieval of the gratitude expressions containing *th*ank** has been conducted in the 12 different files of LLC:c respectively, for the compliers have divided the corpus into 12 different files according to the speech domains, recording setting of the data, and social distance or power relations between the participants. In order to retrieve all the instances spoken by either sex respectively, <S1> was used to mark all the instances of thanking spoken by the female speakers found in the dialogue texts, and <S2> was used to code those spoken by the male speakers.

205 tokens containing thank* were retrieved from COLT, however according to the discourse context in which they occurred, we eliminated 6 irrelevant tokens and 22 instances of narrative gratitude expressions. In order to explore the possible differences in the realization of thanking between teenagers and adults, we picked out only those instances uttered by teenagers.

The 28 informants' responses to the questionnaire survey were edited and saved as two ASCII-formatted documents in one file. Both the information of the informants and their views on the different types of questions were numbered and coded with consistent markers with the use of < >. By so doing, MCO can present the informants' various comments on the same question in concordance lines and that provide us with easier ways to find out the obvious tendency displayed in those comments and to make classifications of them.

Given the fact that 'thank you' and 'thanks' expressions are the major forms of thanking in LLC:c and COLT, the discussion of the social constraints on the lexical realization will focus mainly on these types of expressions.

4. Results

The Lexical Realization of Thanking

The informants of our questionnaire survey have added various verbal expressions of thanking to the lists summarized by Stenström and Aijmer. The number in the brackets indicates the number of informants recommending this expression:

> *Ta (5); Cheers (5); Cheers, thanks (1); Cheers, mate (1); Thanks, mate (1); Thanks a bunch (1); Much obliged (2); Ta very much /muchly (2); Nice one (1); Cool (2); Lovely (2); Great (3); Ace (1); (That's) fantastic (2); Good job/show (1); How kind of you ... (1); Ciao (1); I am indebted to you for ... (1); Much appreciated (1); Good enough (1); Super (1); Superb (1); Brilliant (1); Wonderful (2); It's lovely/beautiful/just what I wanted (1); Thanks a million (2); I really appreciate that (1); That really means a lot to me (1); Thanks a whole lot (1); I would like to express my gratitude to ... for ... (1); You are (have been) a great help to ... (2); That's very kind/good of you (1); Gracias (1); (Address term), thank you very much(1).*

Nevertheless, 'thank you' and 'thanks' are found to be the main gratitude expressions in both LLC:c and COLT in spite of the fact that some other illocutionary force indicating devices such as 'grateful', 'gratitude', 'nice', 'kind', 'good', 'ta', 'cheers', 'lovely', 'great', 'super', 'wonderful', and 'glad', are used occasionally.

Native Speakers' Comments on Gratitude Expressions

The informants' comments on the forms are various and some are even contradictory.

Table1 Informants' comments on the gratitude expressions in LLC:c

Type of Comments	Forms	Number of informants
1. UNCOMMON	thank you very much indeed	2
	thank you + address term very much indeed	10
	thanks very much indeed	6
	thanks awfully	10
	thank you so much indeed	16
	ta	3
	cheers	3
2. LESS COMMON	thank you very much indeed	3
	thank you +address term very much indeed	3
	thanks very much indeed	5
	thanks awfully	2
	thank you so much	3
3. MOST COMMON	thanks	8
	thanks for + V-ing	3
4. SARCASTIC	thank you very much indeed	4
	thanks awfully	3
	thank you so much	2
	thanks a lot	7
5. OLD-FASHIONED	thank you very much indeed	2
	thanks awfully	11
	thank you so much	3
	many thanks	2
6. CLASS-SPECIFIC	thanks awfully	10
	thank you so much	4
	thank you so much indeed	2
7. FORMAL	thank you	4
	thank you for + NP	5
	thank you very much	6
	thank you very much indeed	10
	thank you +address term very much indeed	8
	thank you so much	3
	I am very grateful to ...	10
	It is/ was very nice of you ...	4
	It is/was very kind of you ...	4
8. WRITING	thank you for +NP	3
	many thanks	4
	I am very grateful to ...	6
	It is / was very nice of you ...	3
	It is / was very kind of you ...	2
9. INFORMAL	thanks	13
	thanks for +NP	5
	thanks for +V-ing	5
	thanks very much	8
	thanks very much indeed	4
	thanks awfully	2
	many thanks	4
	thanks a lot	8
	ta	4
	cheers	5

The analysis of the comments indicates that varying gratitude expressions can have different degrees of popularity, different degrees of formality, some of the forms are used more in writing than in daily conversation, some can be used or are mainly used satirically, some are associated with certain classes, and some are considered to be old-fashioned.

Speech Domain and Thanking

Table 2 summarizes the distribution of thanking instances containing "thank you" and "thanks" separately in the five speech domains of LLC:c, and in further categories of text types within them, the degrees of familiarity between interlocutors, respectively. Comparing the frequency of THANK YOU and THANKS per 10,000 words, Table 2 shows that answer phone and Dictaphone obviously has the highest frequency of thanking containing 'thank you and thanks' (53), telephone conversation ranks second (30), face-to-face conversation ranks third (4.89), public discussion ranks fourth (2.94), and public speeches ranks last (0.56). From this comparison we can see that thanking can also be constrained by speech domain.

The forms used in the answer phone calls tend to be fairly formal. The possible explanation for it is that if thanking is uttered for fulfilling of one's request (either before or after the fulfillment of the requests), it involves higher degree of gratefulness and thus more formal forms are likely to be used. On the other hand, the thanking expressions in Dictaphone tend to be lengthier than those used in casual conversation, with three of the 5 examples consisting of 5 to 6 tone units. Public discussion consists mainly of radio discussions or interviews, and committee or academic meetings, and the use of thanking in it is relatively role-specific. In radio discussions and interviews, the majority examples indicate that at the end of the discussion, generally speaking, it is the broadcaster or interviewer (the one who presides) that should say "thank you" first to the participant for giving comments (because by so doing they might be able to change to a new topic or close the conversation smoothly). Vice versa tends not to be the case at all.

Table 2 Distributions of THANK YOU and THANKS in various speech domains in LLC:c

Speech type	number (number per 10,000 words)		
I DIALOGUE	THANK YOU	THANKS	TOTAL
(A) face-to-face			
equals	70 (3.59)	24 (1.23)	94 (4.82)
disparates	19(4.75)	2(0.5)	21(5.25)
Subtotal	89 (4.13)	26(1.21)	115(4.89)
(B) telephone			
personal friends	5(3.33)	12(8)	17(11.33)
business associates	56(28)	28(14)	84(42)
disparates	40 (26.67)	9 (6.67)	49 (32.67)
Subtotal	101(20.2)	49(9.8)	150(30)
(C) public discussion			
equals	20(4)	0(0)	20(4)
disparates	5(1.43)	0(0)	5(1.43)
Subtotal	25(2.94)	0(0)	25(2.94)
II MONOLOGUE			
(D) Casual (Answer phone & Dictaphone recording)			
	50 (50)	3 (3)	53 (53)
(E) Public			
spontateous	7(0.82)	0(0)	7(0.82)
prepared	0(0)	0(0)	0(0)
Subtotal	7(0.56)	0(0)	7(0.56)

Degree of formality of setting and thanking

The communicative setting in LLC:c is usually at a person's house, at work, at radio stations, at meetings, or at other public places. The term "public" indicates that the setting are formal; and the term 'casual' on the other hand indicates that the setting are less formal and informal. According to the summary in Table 2, we may find that the bald form "thanks' and its various expanded forms do not occur in the conversations and speeches under the heading of "public". This fact might indicate that "thanks" and its various

intensified forms are less formal than the bald form "thank you". Generally speaking, at meetings is regarded as more formal than at offices and personal houses.

Degree of familiarity between interlocutors and thanking

The notion of formality is not only associated with the communicative setting, but also with the degree of familiarity between interlocutors. Table 2 indicates that the frequencies of thanking between disparates are higher than those between equals and personal friends in both the face-to-face conversations and telephone conversations. This superficial comparison might indicate that thanking is used more with people you do not know well than with people you know well.

In the aspect of the choice of various forms, the comments made by the questionnaire informants indicate that formal forms are used with strangers and people you don't know well, while informal forms are used with intimates, friends, and people you know well. In LLC:c the frequency of "THANK YOU" between equals is slightly lower than that between disparates in face-to-face conversation (3.59 < 4.75), while in telephone conversation the frequency of "THANK YOU" between personal friends is much lower than that between disparates (3.33 < 26.67). On the other hand, the frequency of "THANKS" between equals is a little bit higher than that between disparates in both the face-to-face conversation (1.23 > 0.5) and telephone conversation (8 > 6.67). The differences in this respect seem to indicate that the less formal form "THANKS' is used more between equals and personal friends than between disparates; while "THANK YOU" is used less between equals and personal friends than between disparates.

Change of time and thanking

From the summary of the informants' comments as presented in Table 1, we might easily see the changes in the use and realization of thanking due to the change of time. On the one hand, some of the forms might disappear or just come into being in some certain regions of the English speaking world and might thus become uncommon or even unknown to people in other regions. On the other hand, some of the forms occur in LLC:c such as thank you very much indeed, thank you so much indeed, thanks awfully are considered to be old-fashioned by quite a number of informants.

Social class and thanking

11 of the 24 informants (about 46%) describe that the phrase 'thanks awfully' is mainly used by upper class or regard it as an uncommon and class-specific term; 5 of the 24 informants (about 20.8%) consider 'thank you so much (+ address term)' to be an upper-class phrase; and 2 of the 24 informants indicate that 'thank you so much indeed (+address term)' is class-specific or an upper-class phrase.

Age and thanking

Comparison of the forms between LLC:c and COLT indicates that teenagers seem inclined to use *thanks* and *thank you* without intensifiers more often than adults, and among them, the bald '*thanks*' and *thanks + address term* seem to be more common with teenagers than with adults. Questionnaire data indicates that some of the forms such as *ta, cheers, great, cool, ace* might be used mainly among teenagers and young people; while some other phrases such as *thanks very much indeed, thanks awfully, thank you so much, thank you so much indeed, much appreciated* could be more common with older people.

Gender and thanking

In Aijmer (1996), it is suggested that the sex of the participants might be needed to explain the use of thanking and it is quoted that the study of Greif and Gleason (1980) of children's acquisition of politeness found that mothers used more polite 'thank yous' than the fathers did. In LLC:c, the female and male ratio is 275:356 (about 0.77 times) and the females' and males' thanking token ratio is about 156:139 (about 1.12 times). From this superficial comparison, women used more '*thank yous*' than men did in LLC:c. This finding tends to be in consistent with that of Greif and Gleason. In addition, it is found that the forms: *thanks very much indeed* (5 examples) and *thanks awfully* (3 examples) were all used by women; that the forms: *thanks so much* (1 example), *thanks so much indeed* (1 example), *many thanks* (1), *many thanks for +V-ing* (1 example) were used by men only. In the questionnaire survey, some male informants reflect that *thanks awfully* (1 informant), *thank you so much* (2 informants), and *lovely* (1 informant) seem more common with women. However, these examples and comments are too few for us to come to any conclusions.

4. Conclusion

The major findings of the study show:

(1) The lexical realization of thanking can be diversified and is closely related to the degree of formality (co-determined by the social relation of the interlocutors and the inherent properties of the communicative setting) as well as the degree of gratefulness (co-determined by the social relation of the interlocutors and the inherent properties of the object of gratitude).

(2) Its lexical realization is also loosely related to a number of social factors such as speech domain, change of time, and speakers' social class, age and gender. Some particular forms seem to be more commonly used by interlocutors in a particular type of speech, in a specific time, of a specific class, at a specific age, or having a specific gender.

References

Aijmer, K. 1996. *Conversational Routines in English: Convention and Creativity*. London & New York: Longman

Stenström, A-B.1994. *An Introduction to Spoken Interaction*. London: Longman.

Greif, E. B. and Gleason, J. B. 1980. 'Hi, thanks, and goodbye: More routine information'. *Language in Society* 9: 159-66.

Acknowledgements

Thanks to Professor He Anping, my supervisor, for her invaluable advice and support and to all the informants for their assistance in providing one of the main sources of data used in the present study. Special thanks are also due to Professor Anna Brita Stenström and Professor Knut Hofland, for providing me with the initial access to COLT on-line and for taking their precious time to reply to my inquiry about the LLC corp

Collocation Patterns of Delexical Verbs in Chinese EFL Learners' Writing

Deng Yaochen
Dalian Maritime University

Abstract: The present study focuses on the collocation patterns of delexical verbs used by Chinese EFL learners of English. A corpus-based Contrastive Interlanguage Analysis (CIA) approach is adopted in the study. By comparing the collocations in NS and NNS free production, striking patterns of collocations by Chinese learners are revealed, in terms of quantity, the degree of appropriateness and the degree of accuracy. The study shows that, compared to Canadian native speakers, Chinese learners not only show a strong tendency to overuse the collocations of the delexical verbs, but also allow these verbs more freedom to collocate with a wider range of nouns. Furthermore, a large proportion of the collocations by Chinese learners, especially those unshared by native speakers, are often of a kind typically found in speech rather than in writing. It was found that the less advanced learners, to an even greater extent than the more advanced learners, appeared to use collocations from the wrong register. Although Chinese learners are generally less competent in collocation, yet the present study indicates a clear development pattern with time and increased proficiency. It was further found that, for the collocations of delexical verbs, general verb effect is the vital factor for misuses.

Key words: collocation, delexical verb, corpora, error analysis

1. Introduction

For some time now, it has been widely acknowledged, in the field of EFL teaching, that collocations are an important part of native speakers' competence, and that they therefore should be included in foreign and second language teaching (i.e. Kennedy 1990; Cowie 1992; Bahans 1997; Granger 1998a). However, collocation is indeed one of the obstacles for successful language acquisition, and it has been shown that collocational errors make up a high percentage of all errors committed by L2 learners (Grucza & Jaruzelska 1978 cited in Biscup 1992; Marton 1977; Gui & Yang 2003). Although a few cases of empirical study of learner collocations have been conducted in Europe, on the basis of a reasonable amount of natural production data (Chi *et al.* 1994; Howarth 1996; Granger 1998;Lorenz 1999), yet the patterns of collocations produced by Chinese EFL learners have rarely been investigated. Therefore, it is still largely unclear what features differentiate Chinese learners from native speakers in terms of collocation and what are the problems that Chinese learners of English have in dealing with collocations.

The present study, with a corpus-based approach, attempts to investigate the collocation patterns in Chinese EFL learners' interlanguage. It reports on a study of the verb + noun collocations of six high frequency delexical verbs, i.e. *do, get, give, have, make, take*, aiming to reveal some distinctive patterns of collocations and to identify possible causes of misuses.

2.Theoretical background

The term *collocation* was first introduced by Firth, who considered that collocation is a phenomenon of 'word accompaniment' and a 'mode of meaning at the syntagmatic level', it is 'association' of co-occurring linguistic items between which there is a certain 'mutual expectancy' (Firth 1957). Firth's definition and

explanation of collocation lays a theoretical foundation for further research. For years, subsequent linguists have attempted to define collocation from various perspectives, which resulted in a large number of different but related definitions. All these definitions, though applicable to the collocation study for different purposes, are problematic, to some extent, in theory and practice. On the basis of systematic comparison and evaluation of previous divergent definitions, Wei (1999) put forward a comprehensive working definition of collocation as follows:

> A collocation is a conventional syntagmatic association of a string of lexical items which co-occur in a grammatical construct with mutual expectancy greater than chance as realization of non-idiomatic meaning in texts. (Wei 1999)

The most distinguishing feature of the working definition is the integration of quantitative with qualitative criteria. What's more, the quantitative measure of 'mutual expectancy greater than chance' is considered to be an important criteria for a corpus-based collocation study.

The 'conventional' feature is regarded as the first criteria for a collocation in Wei's definition. Collocations, in nature, are some conventional way of saying things. The conventionality is most clearly embodied in the collocations of delexical verbs, such as *do, get, give, have, make, take*, etc (ibid).

Delexical verbs are termed in this way 'because of their low lexical content and the fact that statements of their meaning are normally derived from the words they co-occur with (MaCarthy 1998). For example, in *take a photograph, take a decision,* it is *photograph* and *decision* that retain their normal meanings while *take* loses its meaning. Wei (1999) takes '*make a decision*' and *'*do a decision*' as an example to illustrate the conventionality of the collocations of delexical verbs. Wei (*ibid*) argues that it is completely a convention in English to talk about '*make a decision*', not '*do a decision*', although any speaker of English would understand the latter unconventional expression. Arguably, there are no interesting structural properties of English that can be gleaned from this contrast. The choice of these verbs is mostly arbitrary and semantically unmotivated (Allerton 1984). Therefore, the collocations of these verbs are usually problematic for learners.

Inspired by Sinclair's underuse hypothesis (1991), Altenberg and Granger (2001) conducted an empirical study on the use of delexical verbs by foreign learners. The result of the study conforms Sinclair's hypothesis: both the Swedish and French-speaking learners underuse delexical structures.

In the present study, guided by the working definition of collocation by Wei (1999), we aim to throw some light on the collocation patterns of six high frequency delexical verbs in Chinese English learners' free written production. We focus on the verb + noun collocation patterns of delexical structure exclusively, excluding free combinations and idioms of these verbs. We expect to answer the following 4 questions:

1) Do Chinese EFL learners tend to over- or under-use the collocations of these delexical verbs, as compared to native speakers?
2) What patterns of collocations differentiate learners from native speakers?
3) How do the pattern change with time and increased proficiency?
4) What factors are attributable to the collocation errors?

3. Research Design

3.1 Corpora

To answer the research questions, a corpus-based Contrastive Interlanguage Analysis (CIA) approach

(Granger 1998b) is adopted in the study. We will compare not only the NNS learner corpora with the NS corpus, aiming to identify the collocation patterns of delexical verbs in the interlanguage, but also the NNS corpora between learners at different proficiency levels forevidence of the development pattern.

The NNS data are drawn from the Chinese Learner English Corpus (CLEC), which consists of over a million words of written samples collected from almost all proficiency levels of school learners of English in China. Two sub-corpora of CLEC, ST4 (for advanced non-major learners) and ST6 (for advanced English major learners), standing for two different proficiency levels, are used as source of evidence for collocation patterns.

The NS control corpus, created by the author, is made up of 350 essays, all of which are free writings of Canadian native speakers of English at more than 10 universities. The topics are comparable to those in CLEC. The types and tokens in the NS and NNS corpora used for comparison in the study are listed in Table 1.

Table 1 Learner and native speaker corpora

Corpora	Types	Tokens
NS	10231	154039
St4	8499	252247
St6	10756	220672

What is unique in the present study is not only the comparison between L2 learner corpora and native speaker corpus for collocation patterns but also the comparison between L2 learner corpora and L2 learners' mother tongue corpus, in my case, Chinese, for evidence of L1 transfer. The Chinese mother corpus is sampled from *People's Daily* in 1998.

In judging the appropriateness of the verb-noun collocations produced by Chinese learners, several other auxiliary corpora were referred to: LOBa and BROWNa for written English; the spoken sample of BNC, for spoken variety; JDEST, for science variety. If there turned out to be one or more instances of the suspicious verb-noun combination in one register, the pair was considered to be acceptable. Otherwise it was marked as a deviant collocation.

3.2 Procedures for data collection

To obtain the data necessary for the present study, the Concordance software and three Foxpro programs were applied. The first step in our analysis of the data was to lemmatize all the delexical verbs under investigation in the three corpora, i.e. NS, St4 and St6, with the substitute function of WinWord. The advantage of lemmatization is that it is then possible to create a concordance of each lemma rather than having to create concordances for each verbal form. In the second step, the concordances for each delexical verb in question were created, and all the co-occurrence word types in a span of 4 words on the right were extracted. The co-occurrence frequency of each word type was computed before they were saved in the file CO-OCCUR.txt. After that, the first Foxpro program RIGHT.prg was run to extract all the nouns in the file CO-OCCUR.txt, which are possible collocates of the delexical structure. From the small number of nouns, we can manually select the collocates in compliance with the norms of the study. In a third step, the second Foxpro program COLLOCATION.prg was run in each corpus to calculate the collocation strength, i.e. the degree of typicality, of each selected collocate, measuring by T-score. Then the collocates with the frequency of co-occurrence greater than 3 and the T-score greater than 2 are chosen as typical ones and sorted by T-sore for further comparison. In the fourth step, in order to discover the difference in the collocations used by Canadian native speakers and Chinese learners, the third Foxpro program

COMPARE.prg was run to separate all the collocations into three groups: overlapping, only used by learners and only used by native speakers. The application to the data of the last three steps outlined above resulted in 5 files for each verb in each corpora: (1) delexical collocates (2) typical collocates (3) overlapping collocates (4) collocates only by learners and (5) collocates only by native speakers.

In exploring the collocation patterns in terms of over-use and under-use compared with native speakers, all frequency differences across the samples have been tested by means of the chi-square test, with 5 percent ($p<0.05$) as the critical level of statistical significance. An asterisk in the table below marks statistically significant differences between each learner group and the native speaker control corpus.

4. Results and Analysis

4.1 Frequency of the delexical verbs

Table 2 shows the frequency of each delexical verb in the NS and NNS corpora.

Table 2 Frequency of delexical verbs in NS and NNS corpora

Verb	NS	St4	St6
Do	130	1068	369
Get	78	389	306
give	64	136	196
have	610	1169	767
make	137	682	157
take	103	437	207

Table 3 Chi-square value for de-lexical verbs in NS and NNS corpora

Verb	NS vs St4	NS vs	St4 vs
do	184.5*	106.94*	5.58*
get	97.37*	182.53*	30.38*
give	40.93*	66.23*	1.98
have	200.77*	1.51	273.21*
make	0.56	0.5	3.09
take	126.8*	2.3	110.42*

The table brings out an overall tendency of over-use in the verb + noun collocations of delexical structure by Chinese EFL learners. But there are specific features with different groups and different verbs. The values for chi-square in Table 3 confirm that at the significance level of 0.05 ($p<0.05$), the learners in St4 group use more collocations of 5 verbs out of 6 than Canadian native speakers. Only one verb, i.e. *make*, is not overused significantly. In contrast, the more advanced learners, in St6 group, use the collocations of only 3 verbs significantly more than native speakers, i.e. *do, get, give*. The comparison between learners at two different proficiency levels shows that there is no significant difference in the use of two verbs, that is, *give* and *make*. From the result, we can see that Chinese learners, whether in St4 group or in St6 group, are comparatively successful in the acquisition of the verb *make*, but only in terms of the quantity of use. The result is due to the fact that this verb is usually treated as a key verb and has received more attention, whether in teaching or in learning. So learners have got more exposure to this verb than to the others.

The finding of overuse pattern in the collocations of delexical structure by Chinese learners, especially by learners in St4 group, is quite different from Sinclair's underuse hypothesis and the result of Altenerg and Granger's study (2001) based on Swedish and French-speaking learners, which indicates that this phenomenon is L1-related.

There are two possible explanations for the overuse pattern by Chinese learners. Firstly, according to Hasselgren (1994), learners, even sufficiently advanced, hugely overuse these verbs just because these verbs are 'learnt early, widely usable, and above all safe'; therefore learners tend to cling on to them like 'lexical teddy bears'. Secondly, the overuse is related to the limited vocabulary of learners. When having difficulty in nuancing the common notions, learners have to resort to these items. And they employ

repetition of these verbs as a strategy for communication.

From the result of the above comparison, we can see that although Chinese learners have an overall tendency of overuse of these collocations, yet, with the development of interlanguage, the degree of overuse is decreasing. When learners reach a higher proficiency level, in our study, the level of St6 group, the number of overused verbs decreases to only three. Furthermore, even if overused, the occurrences of these collocations are much fewer than those in St4 group (except the verb *give*). This finding gives us some hints that in terms of the collocations of delexical verbs, Chinese learners, whether at St4 group or at St6 group, are at different development stages. The subsequent analysis in the paper will guide us to search for more evidence for this pattern.

To have a deeper understanding of the distinctive patterns of the collocations by native and non-native students, it is necessary to examine the difference of typical collocations from the perspective of category and type.

4.2 Comparison of typical collocations of each verb in NS and NNS corpora

After the third and fourth steps of data collection, all the typical collocates of each verb are extracted for further comparison. The top 15 collocates of HAVE by T-score are listed in Table 4 for example. The collocates in italics refer to those unshared by native speakers, while those in bold , unshared by learners.

Table 4 Top 15 typical collocates of HAVE in NS and NNS corpora

NS	T-score	St4	T-score	St6	T-score
trouble	6.17	chance	11.50	right	18.42
fun	4.82	*rest*	6.61	chance	12.21
chance	4.74	Idea	6.58	effect	6.56
idea	4.13	influence	6.26	opportunities	5.49
contact	3.75	Right	6.15	advantages	5.27
tendency	3.69	trouble	6.13	trouble	5.11
energy	3.32	doubts	5.85	ability	3.87
doubt	3.10	choice	4.98	*adaptation*	3.87
validity	2.82	*conditions*	4.63	connection	3.86
connection	2.66	*belief*	4.31	idea	3.44
memories	2.47	*lunch*	4.31	*welcome*	3.29
choice	2.33	opportunities	4.16	choice	2.90
influence	2.23	experience	4.13	knowledge	2.68
interests	2.05	ability	3.95	*hope*	2.54
meeting	2.05	*disadvantage*	3.76	influence	2.37

By comparing the typical collocations in NS and NNS free production, four striking patterns of collocations by Chinese learners are revealed.

Firstly, the collocations produced by Chinese learners, whether at St4 group or at St6 group, are all characteristic of a wider range of collocates than those by native speakers, although the disperse degree of St6 group is a little lower than that of St4 group. This can be accounted for again with the reason mentioned in 4.1, the lack of sufficient vocabulary. When learners are not sure which verb should be used to collocate with nouns listed above, they will recourse to these delexical verbs. It also indicates that Chinese learners have a low level of collocation competence. The typical collocations, which are frequently used by native speakers, are not psycholinguistically salient to the learners and they cannot spring to the learners readily.

Secondly, from table 4, we can find that although some of the typical collocates in NNS corpora are also shared by native speakers; however, the T-scores are different from those in NS corpus. Besides, these collocates are in different positions in the list of top 15 collocates, which means they are different in typicality for native speakers and Chinese learners. Take the collocate *trouble* for example. In NS corpus, it is on the top of the list, while in the NNS corpora, whether in St4 group or in St6 group, it is in the middle.

Thirdly, the most important, a large proportion of the collocations produced by Chinese learners, especially the unshared ones, are different from those by native speakers in register. The collocations by learners are more often of a kind typically found in speech rather than in writing. For example,

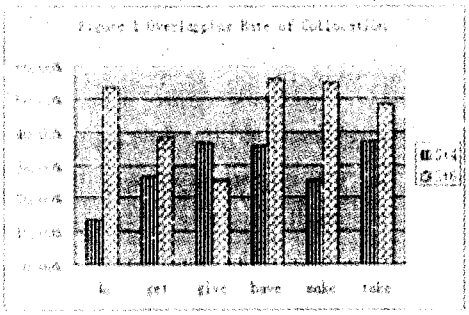
> *do shopping, do reading; have a look, have no idea, have lunch, have the right, have a rest, have a mind; make a bargain, make a telephone call, make contact, make a decision, make effort, make no difference, make fun, make a bad impression, make a killing, make a living, make a mistake, make a statement, make a fortune, make a profit, take advantage, take care, take a chance, take control, take effect, take an interest, take a look, take a notice, take pains, take part, take place, take a long time, take a break, take drugs, take photos, take a risk, take the trouble.*

According to the results of corpus-based studies by Biber *et al* (2000), all these collocations are dominate in conversation rather than in written exposition. The statistics of this kind of collocations in NNS corpora shows that the less advanced learners, to an even greater extent than the more advanced learners (40%:25%), appeared to use collocations from the wrong register. It can be concluded that the perception of these essays as less idiomatic could partly be attributed to the fact that the collocations were perceived as belonging to a more informal register, because 'these learners simply employ the restricted lexicon of speech, writing down talk' (Cobb 2003). On the contrary, the collocates which are not shared by learners are more frequently used in writing. For instance, the noun *tendency* is a typical collocate of the verb *have*, but doesn't appear in the collocate list of learners. The corpus-based study by Leech *et al* (2001) shows that *tendency* is 3.2 times more often used in writing than in speech (the distinctive value is 183).

The fourth pattern is devoted to the development evidence for Chinese learners' collocation competence. We argue, in this paper, that the more typical collocations learners share with native speakers, the more competent the learners are in collocation. Therefore, the rate for collocation overlapping by learners at St4 group and St6 group are calculated respectively and listed in Table 4. The result is visualized in Figure 1 for easy interpretation.

Table 5 Overlapping Rate of Collocates

Verb	St4	St6
Do	13.64%	53.33%
Get	26.32%	38.24%
Give	36.84%	25.00%
Have	35.85%	55.88%
make	25.00%	54.54%
Take	37.03%	48.28%



It is obvious from Table 5 and Figure 1 that in 5 verbs out of 6, the learners at St6 group share much more collocations with native speakers than those at St4 group. The overlapping rates for 4 out of 6 verbs in St6 group are up to or over 50%, but in contrast, in St4 group, the highest rate is not higher than 40%. Except the verb *give*, the rest all display the signs of progression. Figure 1 shows that the development pattern can be clearly represented by the cases of the verbs *do*, *make* and *have*. With the development of the

interlanguage, the use of collocations by Chinese learners are changing not only in quantity (from overuse than native speakers to approximate to native speakers) but also in quality (approximate to native speakers gradually). But the verb *give* is an exception, the overlapping rates of which in both St4 group and St6 group are relatively low (36.84%, 25%). And the learners at St6 group allow the verb even more freedom to collocate with nouns. The verb *give* is problematic for Chinese learners. It provides us some pedagogical implications.

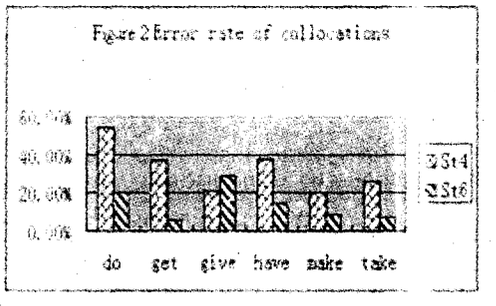
4.3 Misuses in collocations by Chinese EFL learners

Careful study of the collocations in NNS corpora show that Chinese learners not only overuse the collocations of delexical structure, they also misuse them. In this part, we will first investigate the development patterns of collocations in relation to the rate of misuses, and then the factors of misuse will be explored.

To understand the accuracy of collocation use by Chinese learners and explore the causes of misuse, all the deviant collocations are extracted manually from the learner corpora. Table 6 indicates the rate of collocation errors committed by learners at two different levels. The pattern of change is visualized in Figure 2.

Table 6 Rate of Collocation Errors

Verb	St4	St6
do	54.54%	20.00%
get	36.84%	5.90%
give	21.05%	28.57%
have	37.74%	14.70%
make	20.00%	9.00%
take	25.92%	6.90%



Compared with Figure 1, Figure 2 shows opposite trend, rising in the former, decreasing in the latter. But they symbolize the same pattern of development. The learners at St6 group make fewer collocation errors than those at St4 group. 5 verbs of 6 under investigation (i.e. *do, get, have, make, take*) display significant differences between the two groups. Once again, the development pattern is testified. But the verb *give* is again an exception. No evidence of development in the acquisition of this verb can be found in our study. So, we can conclude tentatively that, firstly, with the advancement of language acquisition, the collocation competence of Chinese learners is developing gradually. Secondly, the verb *give* is a great problem for Chinese learners in terms of collocation.

Having described the development patterns from the perspective of collocation errors, the next step, which is the most important for a complete SLA research, is to explore what are the final causes for the learners to commit collocation errors. Careful examination shows that the errors are partly interlingual and partly intralingual. All the deviant collocations are classified into three categories for analysis according to different causes of misuse: *general verb effect; Li influence and overgeneralization*. They will be illustrated separately.

1) General verb effect

Statistics shows that, among the deviant collocations of delexical verbs in the present study, 64% are due to what might be called 'general verb effect'. The following samples are cited for illustration:

his example, do [fm3,3-2] you *make* [cc3,-2] *a conclusion* that some
And some *penalty* should be *done* [cc3,4-] to those cheat in [wd3,-] fake
nothing and have to *have medicine* [**cc3**,1-] every day,
I will do my best to *get* [cc3,-1] *sucess* [fm1,-]
no reason for you to *do* [**cc3**,-1] *crime*, even if you we
industry have [vp3,-] been *take* [**cc3**,-3] very big *progress*.

These errors happen when the learners have not internalized the specific verb to describe an action and resort to using these delexical verbs for general purpose. This can be surmised from the fact that this kind of verbs are often seen in learner corpora in place of more specific verbs for a certain state or action.

In the samples cited above, the student using *make a conclusion* does not know *conclusion* is habitually collocated with the verb *draw* in English, so he uses *make a conclusion* instead, assuming that in English *make* is an all-around verb suitable in many kinds of context.

2) L1 interference

According to Nesselhauf's (2003) result of a recent empirical study based on German-speaking advanced learners of English, L1 influence is a vital factor in collocation errors, which is stronger than what some researchers suspected (cf.e.g. Biskup 1992).

Statistics in our study confirms Nesselhauf's conclusion. About 20% of the collocation errors in our study result from L1 influence. The following examples are representative.

Government [np6,-s] must *give* [**cc3**,-3] more *attention* to
same jobs all days long can't *give* them *energe*. [fm1, -] [cc3,2-]
, the fake commodities *have* [**cc3**,-2] great *harms*, so,
satisfied me. So I wanted to *do* more beautiful *dreams* [**cc3**,3-]

If we compare the above problematic collocations with their Chinese counterparts, we can see clearly all these deviant collocations could be the results of mother tongue influence, since they are all direct or literal translations of the assumed Chinese equivalents, which would be unacceptable or ridiculous with a native speaker of English. In order to verify our statement, we take **do dreams* as an example. This deviant collocation is clearly an unacceptable deviant of *dream a dream*, corresponding to '做梦' (*zuo meng*)in Chinese. Part of the concordance lines of '做梦' in Chinese mother tongue corpus are listed below:

wei shen mo huizuo meng de yi zhong jie shi
yan jiu zhong fa xian,zuo meng shi da nao di ji
lao renzuo meng ye mei you xiang
di fang de hai zizuo meng ye bu zhi dao
hun ran da dun,zuo zhe fa cai meng
na hui er zheng tianzuo zhe zuo jia meng
hua wei pao ying bizuo hang liang meng

It is clear that when learners attempt to express the meaning of *dream a dream*, but they don't know which verb is the most appropriate to collocate with the noun *dream*, they treat L1 transfer as a strategy to fill the gap. The verb *do* is the most frequent verb corresponding to '做' in Chinese in learners' lexicon, as a

result, the deviant collocation *do a dream* is produced.

3) Overgeneralization

Our study shows that deviant collocations of delexical verbs can also result from overgeneralization. When learners apply the patterns of some collocations or phrases that they are familiar with to other words, collocational overgeneralization may occur. The deviant collocation *do good to*, for example, is obviously the reduplication of the mode *do harm to, get favor*, similarly, is generated from the collocation *do sb. a favor*; and *take some means* can be traced back to *take some measures*.

5.Conclusion

By comparing the collocations in NS and NNS corpora, the striking collocation patterns of delexical verbs have been revealed, in terms of quantity, the degree of appropriateness and the degree of accuracy.

The study shows that, compared to Canadian native speakers, Chinese learners not only show a strong tendency to overuse the collocations of the delexical verbs, but also allow these verbs more freedom to collocate with a wider range of nouns, which indicates that the typical collocations for native speakers are not salient in learners' lexicon. Furthermore, a large proportion of the collocations by Chinese learners, especially those unshared by native speakers, are often of a kind typically found in speech rather in writing. The less advanced learners are more likely to use the collocations from the wrong register.

An obvious development pattern is discovered by comparing collocations produced by learners at different proficiency levels. In all three aspects involved in the study, i.e. the frequency, the overlapping rate and the degree of accuracy, the learners at St6 group are at a higher developmental stage than those at St4 group.

After analyzing the deviant collocations used by Chinese learners, it is found that although L1 influence plays an important role in collocation errors, yet, for the verbs under investigation in the present study, many more deviant collocations are attributable to the general verb effect. The low rate for overlapping and the high rate for errors indicate that Chinese learners are generally less competent in collocation than native speakers. We can conclude that even advanced learners have difficulties in the production of collocations. Collocations indeed deserve more attention in foreign language teaching and learning.

References

Allerton, D. J. (1984). *Three (or four) levels of cooccurrence restriction*. Lingua 63: 17-40.

Altenberg, B. & Granger, S. (2001). *The grammatical and lexical patterning of MAKE in native and non-native student writing*. Applied Linguistics 22: 173-194.

Bahans, J. (1997). Kollokationen und Wortschatzarbeit im Englischunterricht. Tübingen: Narr.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (2000). Longman Grammar of Spoken and Written English. Beijing: Foreign Language Teaching and Research Press.

Biscup, D. (1992). *L1 influence on learners' renderings of English collocations: A Polish/German empirical study*. In P. J. L. Aunaud and H. Bejoint (eds). Vocabulary and applied linguistics. Houndmills: Macmillan.

Chi, M. L. A., P. K. Wong, and C. M. Wong. (1994). *Collocational problems amongst ESL learners: a corpus-based study*. In L. Flowerdew and A. K. Tong (eds). Entering Text. Hong Kong: University of Science and Technology, pp. 157-65.

Cobb, T. (2003). *Analyzing late Interlanguage with learner corpora: Quebec replications of three European studies*, Canadian Modern Language Review 59: 393-423.

Cowie, A. P. (1992). *Multiword lexical units and communicative language teaching*. In P. J. L. Arnaud and H. Bejoint (eds). Vocabulary and Applied Linguistics. Houndmills: Macmillan, pp.1-12.

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

Granger, S. (1998a). *Prefabricated patterns in advanced EFL writing: collocations and formulae*. In A. P. Cowie (ed.). Phraseology. Theory, analysis, and applications. Oxford: Clarendon Press, pp. 145-60.

Granger, S. (1998b). *The computerized learner Corpus: a versatile source of data for SLA research*. In S. Granger (ed.) Learner English on Computer. London and New York: Longman.

Hasselgren, A. (1994). *Lexical teddy bears and advanced learners; a study into the ways Norwegian students cope with English vocabulary*. International Journal of Applied Linguistics 4: 237-60.

Howarth, P. (1996). Phraseology in English Academic Writing. Some implications for language learning and dictionary making. Tübingen: Niemeyer.

Kennedy, G. D. (1990). *Collocations: Where grammar and vocabulary teaching meet*. In S. Anivan (ed). Language Teaching Methodology for the Nineties. Singapore: SEAMEO Regional Language Centre, pp. 215-29.

Leech, G., Payson, P. and Wilson, A. (2001). Word Frequencies in Written and Spoken English. Harlow: Pearson Education.

Lorenz, G. R. (1999). Adjective Intensification—Learners Versus Native Speakers. A corpus study of argumentative writing. Amsterdam: Rodopi.

MaCarthy, M. J. (1998). Spoken Language and applied Linguistics. Cambridge: Cambridge University Press.

Marton, W. (1977). *Foreign vocabulary learning as problem no. 1 of language teaching at the advanced level*. Interlanguage Studies Bulletin 2(1): 33-57.

Nesselhauf, N. (2003). *The use of collocations by advanced learners of English and some implications for teaching*. Applied Linguistics. 24: 223-242.

Wei, N. X. (1999). Towards Defining Collocations: A Practical Scheme for Study of Collocations in EAP Texts. Unpublished Ph. D. thesis. Shanghai Jiaotong University.

Gui, S. C & Yang, H. Z. (2003). Chinese Learner English Corpus. Shanghai: Shanghai Foreign Language Education Press.

Acknowledgements

I would give my heartfelt thanks to my supervisor Professor Fan Fengxiang, who led me to the interesting new research approach——corpus approach, which is proved to be very useful for EFL teaching and learning. Without his patient instruction, insightful criticism and constant encouragement, this paper would not have come into being. Thanks also go to Professor Wang Haihua for her useful comments on a previous version of this paper.

Statistical Study on TAM of English *If*-conditionals

Cao Jingxiang
Dalian University of Technology

Abstract: This paper is a report of a computer-aided investigation into the Tense-Aspect-Modal (TAM) system of the finite verb phrases in the *if*-conditionals in six corpora: Brown-A, LOB-A, T4, MEE, MTE and NEC. Of all the grammatical sub-systems, TAM is probably the most complex and frustrating to the linguist. This paper is focused on the *if*-conditionals, for TAM in this particular structure has aroused much interest of English teachers and test designers. The research is based on a simple TAM framework in which the marked and unmarked TAM forms are in contrast, Past v.s. Non-past, Modal v.s. Non-modal, Perfect v.s. Non-perfect, Progressive v.s Non-progressive, and there are 48 theoretically possible TAM forms. But the corpus investigation shows that, in the case of *if*-conditionals, only a small part of them are commonly used, whereas others are extremely rare cases or even do not occur at all.

Key words: Tense Aspect Modal If-conditionals Corpora

1. Introduction

Of all the grammatical sub-systems, Tense-Aspect-Modal (henceforth TAM) is probably the most complex and frustrating to the linguist. In this paper the author is going to have a closer look at the operation of the TAM in English. The focus will be put on the *if*-conditionals from six corpora available. For TAM in this particular structure has aroused much interest of English teachers and test designers. There is, almost in every national or international English test, one or two items on the choice of the verb forms in *if*-conditionals. It seems that there are too many 'unusual' uses of the verb forms in the structure.

1.1. TAM Framework

Every English finite verb phrase contains four grammatical categories: Tense, Aspect, Modality, and Voice. Voice is not investigated here since it is not directly related to temporality. Past and Non-past is of the category of Tense. Past Tense is marked by the past form (V-*ed*) of the finite verb; Non-past is the unmarked one (with *-s/-es* marking singular number). Aspect covers two oppositions: Progressive / Non-progressive (Simple), and Perfect / Non-perfect (Simple). Progressive is marked by *be* V-*ing*, and Perfect is marked by *have* V-*en*. Modal Auxiliaries are investigated here as Tense is represented by the Modal forms in the clauses with Modal Auxiliaries. Those sentences without Modal Auxiliaries have unmarked Modality, which termed as Actuality, and which in the semantics of possible worlds is valued only in the actual world (Steedman 1997:912).

In this frame, there are theoretically 8 possible combined Tense Aspect forms, corresponding to the traditional 8 primary tenses (Palmer 1988). Combined with 10 Modal Auxiliaries (Coates 1983), theoretically there are another 40 TAM forms. Therefore, there are 48 possible TAM forms. In practice, as can be seen from later corpus investigation, only a small part of them are commonly used, whereas others are extremely rare cases or even do not occur at all.

1.2. Research Goals and Data

This paper is a report of an investigation into the distribution of the TAM forms in *if*-conditionals. The investigation took the form of a large-scale corpus-based project, looking at finite verb phrases in the six corpora. The research goals are specified as: i) distribution and frequencies of TAM forms of the finite verb phrases in *if*-conditionals in six corpora, and ii) features of TAM forms of *if*-conditionals.

The data is provided by six corpora: Brown-A, LOB-A, T4, MEE, MTE and NEC. Brown-A and LOB-A are respectively the first part of the Brown corpus of American written English and the Lancaster-Oslo/Bergen (LOB) corpus of British written English. Both consist of 44 texts of newspaper reportage. T4 is the fourth part of the JDEST corpus, which is established by Shanghai Jiaotong University and is of English of Science and Technology. MTE, MEE, and NEC stand for Maritime Treaty English, Maritime Engineering English, and Nautical English Corpus respectively. The three corpora are set up by Dalian Maritime University and may be grouped together as Maritime English. Their respective sizes are given in Table 1

Table 1: The Sizes of the Six Corpora

Corpus	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
Size(w)	90, 735	91, 022	121, 693	503, 192	214, 027	672, 417	1, 693, 086

1.3. Procedures and Techniques

A corpus-based approach is adopted in this study, but the original corpora are not ready for TAM analysis since the elementary unit of the study is not words but clauses and sentences. So the running texts have to be processed by computer programs, and manual work is necessary where computer programs are not available or inadequate.

The study was carried out through the following steps:

1. Break the original corpora into sentences to provide a mini-context.
2. List all the sentences with the word *if*.
3. Sample the *if*-sentences if the output file in step 2 is over 500 sentences.
4. Encode the sentences with a set of TAM codes (See Appendix 1).
5. Sort out the coded sentences.
6. Fill all the frequencies into a table and do some calculations accordingly.
7. Interpret the results and draw conclusions.

In Step 1, a sentence is roughly defined as a string which begins with a capitalized letter and ends with any of the three punctuation marks (.?!) followed by a blank; therefore, the results are not quite accurate in that some outputs are incomplete sentences, but they are enough to provide a mini-context for the *if*-clauses, and the larger context can be traced out if necessary in the original corpus with the help of the referential code in front of the sentences. The number of the sentences in the original corpora is of little relevance to the study.

In Step 3, the sampling is not truly random, but is a variation of simple random sampling known as systematic or quasi-random sampling, where units are taken at equal intervals throughout the numbered population. It is still valid in that there is no periodicity in the population, i.e., the units with the properties under investigation do not tend to recur at regular intervals (Butler: 1985). The interval is 5 for the MTE corpus, and 8 for the MEE and NEC corpora. No sampling is needed for the corpora Brown-A, LOB-A and T4, since the number of sentences with *if* extracted from them is under 500.

The sorting part is a complex task. First, not all the theoretically possible forms occur in the data, so new codes have to be added to the list to tag the new forms. Second, not all the mathematically possible combinations of the Tense forms of the two clauses occur in the samples, so all the actually occurring combinations and the sentences with the same codes have to be sorted out.

The coding is based on the form of the finite verb phrases of the *if*-clauses and their superordinate clauses after excluding the sentences with *even if* or *as if*. Some cases need clarification. First, not all *if*-clauses function as a condition in the context; some may function as an object of a verb or preposition, for which a special code "++++" is added in front of the sentence. Second, there exist some sentences in which two or more *if*-clauses share one matrix clauses. The two clauses may be connected with the matrix clause with the same *if* or with two separate *if*s, and the two have the same or different TAM forms. Even when there is one *if*-clause, it may have two or more coordinate predicates, of which the TAM forms may be the same or different. Since *if* is the focus of the study, in the case with two *if*s the sentence will be coded and counted twice accordingly, and in the cases with two clauses sharing one *if* and with two predicates, they will be coded once in accordance with the first clause or predicate. Third, as with the *if*-clauses, the matrix part may also be complicated. One *if*-clause may modify two matrix clauses or one matrix with two predicates. In both cases, the coding is based on the first TAM form of the matrix part.

One more point that needs clarifying is that the coding is made to facilitate later analysis, so it is not exclusively based on the forms. Some subjective judgment is unavoidable. For example, **v** is the code for *be to* v. structure, but not all sentences with the structure are coded with **v** . And the code for Present Subjunctive Mood (**#** |v.|) is only added to the clauses whose subject and predicate are not in normal agreement numerically, thus neglecting the clauses with plural subjects except those with *be* as the predicate.

Computer programs are used to scan the huge files in the corpus, match words or phrases, count them, pinpoint which texts and lines they occur in, and store the context in which they occur. SNOBOL4 Programs are written and run in this study to break the running text into sentences, to search for sentences with *if* in the corpora, and to sort out the TAM forms in MEE and the other five corpora. SNOBOL is an acronym for StriNg Oriented symBOlic Language. The word *string* is a computer term, which means a sequence of characters such as a word or line of text. SNONOL4 was written to handle text rather than numbers and is therefore particularly suitable for studies in the humanities, and thus is employed in this study.

2. Research Findings

With the help of the codes designed as in Appendix 1, the actual combinations of the TAM forms of the finite verb phrases of the *if*-clauses and their respective superordinate clauses are found out and their gross frequencies are counted. The final results are given in Appendix 2.

The results show that in the *if*-clauses, there are actually only 16 types of codes found in the six corpora. Form A (Simple Present) is apparently the most popular one while W (*must* v.) and I (*shall* v.) are the rarest cases except those with zero frequency, e.g., K (*may* v.), P (Present Perfect Progressive). The matrix clauses have a lot more possible TAM forms than the *if*-clauses. Of the 30 coded forms, only six forms --- # (V.), u (*were* |singular|), p and q (Present/Past Perfect Progressive), m (*could have* V-*en*), and f (Past Progressive) --- have no actual occurrence.

X represents partial ellipsis of the *if*-clauses, z is the code for non-finite clauses, and eq represents mathematical equations. Their TAM is not literally given and they have to be excluded from the main discussion, hence resulting in the Table 2.

Table 2: The Total Number of TAM Forms of If-clauses and their Matrixes

Clause	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
If-clauses	84	57	210	157	77	132	717
Matrixes	84	65	224	175	107	160	815

2.1. Tense

For computational convenience, classification of the TAM forms is based exclusively on forms. Non-past Tense form includes A, C, E, G, I, K, P, R, $, V(present), #, and *, and Past form includes B, D, F, H, J, L, M, Q, S, T, U, and V(past). Accordingly, the distribution of Tense forms are given in Table 3.

Table 3: Percentage of Non-past Forms

Clause	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
If-clauses	59.52	57.89	80.95	86.62	98.70	87.12	80.89
Matrixes	50.00	52.31	74.11	76.57	98.13	73.13	73.37

As can be seen from the table, Non-past forms strikingly dominate both clauses of *if*-conditionals in the six corpora. In the *if*-clauses, all the four corpora (T4, MEE, MTE,NEC) of scientific writing have over 80 percent of Non-past forms, with the highest up to 98.70 percent in MTE. Its percentage in the two corpora (BROWN-A, LOB-A) of news reportage English is comparatively lower, both under 60 percent. In the matrix clauses, the discrepancies are not so conspicuous. The apparent difference proves the common intuition that Non-past form is preferred in EST, which is mainly on generic description and general arguments where the propositions are valued in the past, at the present, and in future, i.e., the inclusive present.

2.2. Aspect

Aspect is the grammatical category which in English includes two pairs of oppositions: Progressive/Non-Progressive and Perfect/Non-Perfect. As coded in this study, only four of the codes contain Progressive forms: E, F, P and Q. Theoretically, there are a lot more possible Progressive forms, e.g., Modal Progressives, but there is even no actual occurrence of F, P, and Q. Its rare occurrence is in the form of E (Present Progressive), whose distribution is shown in Table 4.

Table 4: Percentage of Progressive Forms

Clause	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
If-clauses	1.19	3.51	0.00	2.55	0.00	0.00	0.98
Matrixes	1.19	0.00	0.00	0.00	0.00	0.63	0.25

Zero occurrence of Progressive forms prevails in the corpora (e.g., T4, MTE, NEC). Even in those with Progressives, the percentage is clearly low, with the highest of 3.51 in LOB-A. It can be claimed safely that Non-progressive forms are preferred in general in the if-conditionals.

Perfect forms include C and D (Present/Past Perfect), M, N, O, and T (*could, should, might/would have* v-*en*), and P and Q (Present/Past Perfect Progressive). Only C and D are found in the conditional *if*-clauses in the six corpora, and the rest forms have zero occurrence. In the matrix clauses, The Perfect appear in the form of M, N, O, and T (*could, should, might/would have* v-*en*). The distribution is given in Table 5.

Table 5: Percentage of Perfect Forms

Clause	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
If-clauses	4.76	3.51	1.90	1.91	15.58	4.55	4.32
Matrixes	3.57	7.69	1.34	0.64	0.00	2.50	1.96

Compared with the Progressive forms, the Perfect forms seem to have more frequencies in the six corpora. In all the six corpora except LOB-A, the Perfect form is more likely to occur in the subordinate clauses than in the matrixes. In contrast to its zero occurrence of Progressive forms, the percentage of Perfect forms in the *if*-clauses in MTE is the highest, which seems to imply that the contrast of Perfect / Non-perfect in the condition is of much significance in treaty English.

2.3. Modal Auxiliaries

Besides Tense and Aspect, Modality denoted by the Modal Auxiliaries is also an essential part of the finite verb phrases. The Modal Auxiliaries actually occurred in the six corpora include *can, could, shall, should, may, might, will, would, must,* and *ought to*. But no *may, might*) or *ought to* is found in the conditional *if*-clauses. The percentage of Modal clauses (those with a Modal Auxiliary) is given in Table 6.

Table 6: Percentage of Modal Auxiliaries

Clause	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
If-clauses	11.90	7.02	5.71	7.64	9.09	6.82	7.53
Matrixes	65.48	67.69	55.80	62.29	89.72	59.38	64.29

Modal Auxiliaries are not very frequently used in conditional *if*-clauses in the six corpora, all except Brown-A have a percentage of less than 10. There is also an apparent difference between Brown-A (written American English) and LOB-A (written British English) in the use of Modal Auxiliaries in conditional *if*-clauses. Strikingly in contrast with the *if*-clauses, the matrix clauses in all the six corpora have a preference to marked modal forms. Actually, their percentages of all the corpora are over 50, with MTE at the top. The high percentage in a way justifies the common intuition that conditionality is part of the implication of modality (Palmer 1988). With the similar frequencies of Modal Auxiliaries, the six corpora may vary in their choice of particular modals. Table 7 lists the most frequently used modals in the conditional *if*-clauses of the corpora, and Table 8 the first three most frequently used Modal Auxiliaries in the matrix clauses in descending ranking. Figures in the bracket give the actual frequencies.

Table 7: List of the Most Frequently Used Modal Auxiliaries in Conditional *if*-clauses

Corpus	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
Modals	*can*(4)	/	*can* (4)	*should*(5) *can*(4)	*can*(5)	*can*(22)	

Table 8: Ranking on Frequency of Modal Auxiliaries Used in the Matrixes

Rank	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
1	would(26)	would(18)	would(35)	will(35)	shall(68)	should(26) will(108)	
2	will(9)	will(9)	will(32)	should(23)	may(25)	will(22)	would(103)
3	could(7)	could(5)	can (17)	may (17)	should(2)	may(11) shall(78)	

Modal *can* seems to enjoy the highest frequency in the *if*-clause in all the corpora except MEE where *should* is the most frequently used. Whereas *will* and *would* are the most frequently used, and *shall* is the most common one in MTE, substantiating the authority of the legal documents.

3. General features of TAM in *if*-conditionals in the six corpora

After the long computations and analyses, some general features of TAM in *if*-conditionals have been found. The most striking feature is that of simplicity, simple in Tense, Aspect, and even in Modality in the case of the *if*-clauses.

In terms of distribution of Tense forms, the unmarked Tense form, the Non-past, dominate over the Past form. The Non-past form occupies a percentage of 80.89 in the *if*-clauses and 73.37 in their matrix clauses in all the six corpora as a whole. It may be a rash conclusion that the Non-past Tense overwhelms the Past Tense in all the *if*-conditionals , but it is at least safe to claim that the Non-past Tense prevails over Past Tense in *if*-conditionals of scientific writing, which echoes the common claim that the employment of the Non-past Tense is a register feature of EST.

In terms of Aspects, Progressive and Perfect, the unmarked forms have an absolute advantage over the marked forms in *if*-conditionals, both in the subordinate clauses and in the matrix clauses. In the *if*-clauses, the Non-progressive and the Non-perfect account for 99.12 and 95.68 percent respectively. In three corpora, MTE, NEC and T4, no occurrence of the Progressives is found in the *if*-clauses. In the matrix clauses, the Progressives account for only 0.25 percent in the six corpora as a whole, and in four of them, LOB-A, T4, MEE and MTE, not a single instance is found.

In terms of Modality, or Modal Auxiliaries, to be exact, the situation of the *if*-clauses is quite different from that the matrix clauses. The percentage of Modals in the six corpora as a whole is 7.53 in the former and 64.29 in the latter. The modals particularly used in the clauses are not the same: the first three most frequently used in the *if*-clauses are *can*, *could*, and *should*, while those in the matrix are *will*, *would*, and *shall*.

References:

Butler, C. S. 1985. *Statistics in Linguistics*. Oxford: Basil Blackwell Ltd.

Coates, J. 1983. *The Semantics of the Modal Auxiliaries*, London: Croom Helm.

Palmer, F. R. 1988. *The English Verb*. Beijing: Longman, World Publishing Corp.

Steedman, M. 1997, 'Temporality' in *Handbook of Logic and Language* (ed. By J. Van Benthem and A. ter Meulen). Elsevier Science B.V.

Appendix 1 Codes

a	simple present	b	simple past
c	present perfect	d	past perfect
e	present progressive	f	past progressove
g	can v.	h	could v.
I	shall v.	j	should v.
k	may v.	l	might v.
m	could have v-en	n	should have v-en
o	might have v-en	p	present perfect progressive
q	past perfect progressive	r	will v.
s	would v.	t	would have v-en
u	were(singular)	v	be to
w	must v.	x	ellipsis
y	must have v-en	z	non-finites
$	directive	*	ought to v.
#	v.		

Appendix 2 Gross Frequencies of TAM Forms in If-Conditionals

The *if*-clauses is represented by the upper case and the figures on the left of the slash while the matrix clauses by the lower case and the figures on the right of the slash.

Code	Brown-A	LOB-A	T4	MEE	MTE	NEC	Total
Total	90	67	226	177	108	167	835
A/a	39/22	25/16	157/91	114/56	58/11	100/37	493/233
B/b	24/4	18/3	18/4	6/0	0/0	8/2	74/13
C/c	1/0	0/1	1/0	3/0	12/0	5/3	22/4
D/d	3/0	2/1	3/0	1/0	0/0	1/0	10/1
E/e	1/1	2/0	0/0	4/0	0/0	0/1	7/2
G/g	4/4	1/2	4/17	4/11	4/0	5/6	22/40
H/h	2/7	1/5	4/5	3/1	1/0	2/6	13/24
I/I	1/0	0/1	0/1	0/0	1/68	0/8	2/78
J/j	2/4	0/2	1/13	5/23	0/2	1/26	9/70
K	3	3	9	17	25	11	68
l	1	0	1	1	0	0	3
n	0	1	0	0	0	0	1
o	0	1	0	0	0	1	2
R/r	0/9	1/9	1/32	0/35	1/1	0/22	3/108
S/s	1/23	0/17	2/32	0/16	0/0	0/8	3/96
t	3	1	3	0	0	0	7
U	0	1	8	6	0	3	18
V/v	6/0	5/0	11/1	6/0	0/0	5/1	33/2
W/w	0/1	1/1	0/12	0/4	0/0	1/7	2/25
#	0	0	0	5	0	1	6
X/x	6/1	10/1	16/1	20/0	31/0	35/0	118/3
y	0	1	0	0	0	0	1
z	5	1	0	2	1	7	16
$	2	0	3	10	0	21	36
*	0	0	0	1	0	0	1
eq			1				1

An Empirical Study of the Use of Past Tense in the TEM Band-4 Oral Examination of Chinese Students

Chen Xuan
Nanjing University

Abstract: The present study aims at finding out the pattern in English majors' use of simple past tense in the Band-4 oral English test in China. Specifically, the influence of factors like task types, oral English proficiency, linguistic and contextual factors on the use of simple past tense was explored.

The analysis was based on a mini-corpus extracted from the Chinese English learners' oral English corpora. Results have shown that a number of factors had an effect on the subjects' use of simple past tense. Firstly, the past tense marking rate of verbs in the story-retelling task is higher than that in the task of monologue. Secondly, irregular verbs have a higher past tense marking rate than regular verbs, and dynamic verbs higher than stative verbs. And the subjects tended to mark the verbs in the past tense more often when they were describing a specific past event than when they were describing a person's habitual behavior in the past. Thirdly, temporal adverbials indicating the past time were found to have a positive influence on the marking of verbs in the past tense, but temporal adverbials of frequency did not seem to have this effect. Besides, the position of verbs in a clause unit did not seem to have an effect on past tense marking.

Discussions of these results are provided in terms of the allocation of the limited attentional resources.

Key words: past, tense, verbs, factors, influence

Introduction

Simple past tense is one of the most basic grammatical phenomena English learners have encountered. It is usually introduced at a rather early stage of English learning in China. So the underlying assumption is that this grammar point is easy to grasp. But in fact, it is found that the use of past tense is one of the major problematic areas in Chinese English learners' oral performance (Wen & Wu 1999; Zhu 2000).

Researchers have identified some factors that influence students' past tense marking. Bayley (1994) found that variation in interlanguage tense marking was systematically conditioned by a range of linguistic, social and developmental factors. Among them, verb saliency, grammatical aspect and learners' proficiency level were found to influence the process of marking past tense (Price 1998). And factors like verb salience, lexical aspect, temporal adverbials and narrative structure were found to exert a certain influence on college students' use of simple past tense in writing (Cai 2002).

It would be interesting to find out what factors have an effect on the use of simple past tense in learners' oral performance, so that we can search for countermeasures to improve learner's performance.

The present study explores the influence of various linguistic and contextual linguistic factors on the use of simple past tense by sophomore English majors in their Band 4 oral English test.

Research methods

1. Research questions:

1) Do task types and language proficiency level influence the use of simple past tense? If yes, how?
2) Do linguistic factors such as formal features and semantic features of verbs, grammatical functions of simple past tense exert an influence on the use of simple past tense? If yes, how?
3) Do contextual factors such as the presence or absence of temporal adverbial, the verb position in a clause unit[1] have any influence on the use of simple past tense? If yes, how?

2. Data collection

The data for analysis were extracted from the corpus of Band 4 oral English test for college English majors. This corpus includes both audio data and their transcripts. The test participants were randomly assigned to groups. The data under analysis are from one such group of the year 2001.

3. Data analysis

The oral test of year 2001consisted of three tasks: retelling a story about a forgetful person, a monologue on the topic "an Unusual Teacher", and a dialogue with another testee. Each task allowed three minutes' preparation. The time limit was three, three and four minutes respectively. Since the focus of the present study is on the use of the simple past tense, the third task was excluded from analysis since it did not require the use of simple past tense.

The transcripts of the audio data were first checked against the original tapes for accuracy of transcription. Because of the inferior recording quality of some tapes and the intrinsic nature of some verbs' pronunciation, the verb ending of some words could not be heard clearly. Then these tapes and words were excluded from analysis. Altogether there were data from 31 subjects.

After the data were checked, verbs used in contexts requiring the simple past tense were tagged according to the factors under study. Then the frequency of each type of tagged verbs was computed with the help of Wordsmith Tools. The past tense marking rate of each type of verbs was then calculated[2].

After that, the difference between higher level (top ten of the group) and lower level (bottom ten of the group) subjects was compared with the help of SPSS 10.0 tools.

Results and discussion

1. The influence of task types and proficiency level on the use of simple past tense

Task types were found to make a difference in the use of simple past tense. In general, verbs in Task 1 have a much higher past tense marking rate than verbs in Task 2 (74.5% vs. 59.5%).The past tense marking rate of higher level subjects is significantly higher than that of lower level subjects (t=-2.976, p= .003) in Task 1, but the difference between the two proficiency level subjects does not reach significant level in Task 2.

---

[1] A clause unit is a structure consisting of an independent clause with any dependent clauses embedded within it (Biber et al, 1999:1069).

[2] Marking Rate of a type of verb = (total of correctly marked tokens + total of wrongly marked tokens)/ (total of marked cases + total of unmarked tokens)

One reason for the effect of task type is that the retelling task is much easier than the monologue task. Before retelling, students could listen to the story twice and take some notes. They could refer back to their notes while retelling. This seemed to greatly reduce the difficulty level of the task. But in the monologue task, they had only three minutes for planning. While performing the task, they had to allocate their limited attentional resources to both the content and the form of their speech. Since meaning is always the priority in communication, we can see how this goal is achieved at the expense of form.

What is interesting is that oral proficiency level made a difference in Task 1 but not in Task 2. This seems to suggest that when the task is easier, more proficient speakers have more spare attention to pay to the form of their language than less proficient speakers. But when the task is difficult, this advantage disappears. And this suggests that difficulty level of the task has a direct impact on the accuracy of language.

2. The influence of linguistic factors on the use of simple past tense

2.1 Irregular verbs were found to have a higher past tense marking rate than regular verbs in both tasks (75.3% vs. 71.8% in Task 1; 61.3% vs. 56.6% in Task 2).

2.2 Dynamic verbs were found to have a higher marking rate than stative verbs in both tasks (77.2% vs. 67.1% in Task 1, 65.5% vs. 50.8% in Task 2).

2.3 Verbs can be classified into three types according to their semantic function: material processes, mental processes and relational processes (Halliday 1994). Because of the particular topic of the two tasks, quite a number of verbs describing the verbal process were involved in the task. They should belong to the larger group of verbs describing mental processes, but they were singled out for comparison with other verbs describing mental processes.

Table 1: Past tense marking rate of verbs describing different processes

	Task 1				Task 2			
Verb types	CM	WM	WU	MR (%)	CM	WM	WU	MR (%)
Mate	160	0	46	77.6	161	1	105	62.8
Ment	160	3	71	69.7	52	1	39	57.6
Verbi	145	1	36	80.2	77	1	36	68.4
Relat	178	3	67	73.0	125	2	112	53.1

(mete= material processes; ment= mental processes; verbl= verbal processes; relat= relational processes)

As can be seen from Table 1, verbs describing verbal processes and material processes enjoy higher past tense marking rate than verbs of mental processes and relational processes.

2.4 Among verbs with three different grammatical functions, the past tense marking rate of verbs describing a specific past event enjoy the highest past tense marking rate, verbs describing the habitual behavior in the past have the lowest marking rate, and verbs describing a past state are in the middle.

These results are not difficult to explain if we take into consideration the limited attentional resources in language processing. Since the learning and use of the irregular past form depends more on memory than on rules, irregular past forms are ready-made and the retrieval of them is easier and costs less attentional resources than on-line application of the past tense rule. As for dynamic verbs and verbs denoting verbal and material processes, they are more concrete and have a stronger link with time than stative verbs and verbs of mental or relational processes. Thus they enjoy a higher past tense marking rate. The same is true

for verbs describing a specific past event, they also have a stronger link with the past time than verbs of past state or that describing the past habitual behavior. The very "habitualness" tends to cut the link between the verb and a specific past time, hence the low marking rate.

3. The influence of contextual factors on the use of simple past tense

It was found that no clear pattern had emerged from the past tense marking rate of verbs in different positions of a clause unit. This result indicates that the position of a verb does not have a clear effect on the marking of simple past tense. That is to say, speakers can be reminded to attend to the tense of the verb no matter where the verb occurs in a clause unit.

As can be seen from Table 2, in both tasks, verbs modified by temporal adverbial of frequency have the lowest past tense marking rate, even lower than verbs not modified by any temporal adverbial. The modification of verbs by temporal adverbial only made a difference in the use of simple past tense in Task 2, but not in Task 1.

Table 2: Past tense marking rate of verbs in different linguistic contexts

	Task 1				Task 2			
	CM	WM	WU	MR (%)	CM	WM	WU	MR (%)
Vta0	28	0	9	75.7	37	0	10	78.7
Vtac	26	0	8	76.5	34	0	12	73.9
Vta	86	1	28	75.7	43	0	23	65.2
Vtaf	40	1	28	59.4	35	1	30	54.5
Vnm	468	5	152	75.7	273	5	224	55.4

(Vta0= verbs inside the temporal adverbial clause; Vtac= verbs in clauses modified by a temporal adverbial clause; Vta= verbs modified by a temporal adverbial; Vtaf= verbs modified by a temporal adverbial of frequency; Vnm= verbs not modified by any temporal adverbial)

The fact that the presence or absence of temporal adverbial of past only made a difference in Task 2 but not in Task 1 is not difficult to understand. Since Task 1 is story retelling, the story had a clear setting in the past. Besides, students could be constantly reminded to use the correct tense by the notes they had taken. Therefore, the overall past tense marking of verbs was high, and the effect of temporal devices became less evident. But in Task 2 (monologue), there were less reminders of the use of tense, then the effect of temporal adverbial indicating the past tense became salient.

As for temporal adverbial of frequency, its semantic function is opaque since it can denote both happenings in the past and at present. And this may have led to confusion in the use of past tense.

Conclusion

The result from this study indicates that students' use of the simple past tense in their oral production is conditioned by a certain linguistic and contextual linguistic factors. The accuracy of ready-made verb forms is higher than forms requiring on-line computation. And all the factors that can remind the speaker of the form of verbs they are producing are somehow related to the semantic aspect of the verb. This again corroborates the assumption that when meaning and form are competing for the limited attentional resources, meaning always takes priority.

One implication of this study is that to increase the accuracy of the oral performance, learners need to

achieve a certain automaticity by performing some easier tasks first. Otherwise, other goals are always achieved at the expense of accuracy.

References:

Bayley, R. (1994). Interlanguage variation and the quantitative paradigm: past tense marking in Chinese-English. In Tarone, E. E., S. M. Gass & A. D. Cohen. (Eds.), *Research Methodology in Second-Language Acquisition.* Lawrence Erlbaum Associates Publishers.

Biber, E., Johansson, S. & Associates. (Eds.). (2000). *Longman grammar of spoken and written English.* Beijing: Foreign Language Teaching and Research Press.

Cai, Jinting. (2002). *The Effect of Multiple Linguistic Factors on the Simple Past Use in English Interlanguage.* Unpublished doctoral dissertation. PLA Foreign Languages University.

Halliday, M. H. K. (1994). *An Introduction to Functional Grammar.* London: E. Arnold.

Price, Hsuehmei Liu. (1998). *Past tense marking in English by Chinese learners.* From MAI 37/06, p. 1974. Dec. 1999. MA. University of Richmond.

Zhu, Minghui. (2000). *Factors that affect the marking of the Band-4 spoken English Test for English Majors.* Unpublished Master thesis. Nanjing University.

文秋芳，吴彩霞. 对全国英语专业口语水平的评估：兼评《大纲》对口语的要求. 外语教学与研究. 1999. (1): 29-34.

Modal Verbs in Contrast: a Corpus-Based Study

Liu Hua
Ningbo University

Abstract: This paper starts with a brief survey of major modal categories and modal verbs in English and Chinese. Then a comparison is drawn between modal verb use by NSs and that by Chinese learners of English at the advanced level through a careful look at two corpora---FLOB and the ST6 component of CLEC. Statistics reveal clear discrepancies in the use of modal verbs by NSs and ST6 learners, who use modal verbs twice more than NSs. A detailed breakdown shows ST6 learners tend to express modality with high and median modals, while NSs prefer those of median and low value. In addition, NSs choose to view their opinions from a wider range of alternatives with plenty of freedom, whereas ST6 learners restrict their choice to particular auxiliaries. Several factors contribute to these differences, including the sizes of vocabulary of NSs and ST6 learners, the nature of the texts in the corpora, L1 negative transfer, and L2 learners' inability to comprehend and manipulate modal nuances. Such differences partly explain the gap between the interlanguage and the target language and some communication failures between Chinese learners and NSs.

Key words: modality, modal verbs, differences, overuse, underuse

1. Modality and Modal Verbs

1.1 The concept of modality

Regardless of the many discussions and impressive amount of literature in the area, a consensus is hardly reached on the concept of modality, its classification and functions. Various linguistic schools have given their understanding of this subject from different perspectives (Liang, 2002). Quirk, for example, says modality is used to express the speaker's judgment of the 'likelihood of the proposition being true' (1985: 219). Bussmann, however, provides a much broader definition for the term, noting that modality not only communicates the speaker's attitude but encompasses mood as well (1996: 308-9). In this respect Halliday seems to share Bussmann's idea, as he sets out the discussion of modality in relation to the mood of the sentence (1985: 68-90). To Halliday, modality conveys the speaker's views and is located in the middle between the positive and negative (1985: 334). Some scholars insist on treating modality as a fully independent semantic category though (Li Ji'an, 1999). Although there is no unanimity over its meaning, many linguists agree that modality expresses the speaker's opinion or attitude towards the proposition concerned or the situation described by the proposition, including the speaker's intention and volition (Siewierska, 1991: 123).

Compared to English, the term 'modality' in Chinese is always mentioned in conjunction with 'mood'. Early documented studies of 'words of mood' (such as *le*, *ba*, *a*, *de*, etc.) showed they communicate 'certainty' and 'uncertainty', two interactional functions that are closely related to modality (Qi, 2002). Generally, mood and modality in Chinese convey the speaker's thoughts, feelings and attitude, and therefore every Chinese utterance is said to carry mood and modality (Ma Cuiling, 2002).

1.2 Semantics of modality

In English, linguists offer different divisions to the semantic distinctions made by modality. Epithets such as intrinsic / extrinsic, modulation / modalization, deontic / epistemic, inherent / objective / epistemological, deontic / dynamic / epistemic, etc., are used to group together the sub-categories of modal meanings, areas that have to do with obligation, permission, volition, ability, possibility, intention, willingness, and necessity (Quirk, 1985; Halliday, 1985; Siewierska, 1991; Papafragou, 1998). In the last group, deontic and dynamic uses of modality are often classified together as agent-oriented modalities, or root modalities. Thus the distinction between epistemic and root modalities, a line this paper will follow whenever general types of modal meanings arise for discussion. In this dichotomy, the former is related to the degree of speaker commitment to the truth of the proposition, while the latter involves some kind of human control over events.

By contrast, distinctions as such cannot be easily identified in Chinese, where modal meanings, as the subsequent section will show, are often interpreted and categorized in terms of the meanings of modal operators, or 'helping verbs', as some Chinese scholars call them (Ding, 1999: 89-93).

1.3 The expression of modality

According to Halliday (1985: 334), a native speaker of English has an infinite number of ways to express his opinions, or mask his opinions. This statement entails that modality can also be communicated through numerous ways, but chiefly through modal verbs, mood adjuncts and extended predicators. Among these modal expressions, modal verbs have so far received the greatest attention in grammatical research, thanks to the limited size of the group and its being overt indicators of modality, so much so that some linguists are complaining that the study of modality has been displaced by the study of modal verbs (Li Ji'an, 1999).

In Chinese, modality is mainly expressed by modal verbs and adverbs like *neng*, *yinggai*, *keyi*, *gan*, *ken*, *yuanyi*, *yao*, *bixu*. Linguists disagree on the exact affiliation of such words, which consequently take on myriad names including 'helping verbs', 'modal particles', 'adverbs', 'restrictive words', 'basic verbs', and 'words of evaluation' (*Modern Chinese*, 2000; Chen Guanglei, 2001; Chen Wangdao, 1978; Wang, 1954; Lu, 1956; Ding, 1999; Qi, 2002).

1.4 Modal verbs

In English, the countable group of modal verbs, also called 'modal auxiliaries', belong to the 'closed classes' and are remarkably distinguished from full and primary verbs (Quirk, 1985: 67, 96, 120). Among them, *dare* and *need* are borderline cases or 'straddlers' (Jacobsson, 1979), or marginal modals, while *have to* is a semi-auxiliary. Although each verb is polysemous, Quirk manages to put them into three categories of distinct meanings (1985: 221):

permission ←→ *possibility, ability*: can/could, may/might
obligation ←→ *necessity*: must, have (got) to, need, should/ought to
volition ←→ *prediction*: will/would, shall

Palmer (1983: 29) divides modals into primary and secondary, noting the former group consists of 'present' modals and the latter 'past' as well as more 'modalized' ones which carry a stronger note of politeness. Halliday (1985: 338), on the other hand, attaches degrees of value to the modal verbs:
high: must, ought to, need, have to (also 'dare')
median: will, would, shall, should, be to

low: may, might, can, could

Ding (1999, 89-93) divides the Chinese modal verbs into three groups by their meanings:

possibility, ability, permission: *neng, nenggou, hui, keyi, kenneng, de*

volition, willingness, prohibition: *gan, ken, yuan, yuanyi, yao, dei*

necessity: *ying, yinggai, yingdang, gai*

2. Modal Verbs Used by Native Speakers of English and ST6

2.1 Purpose, methodology and corpora

This paper compares modal verb use by native speakers of English (NSs) and Chinese learners of English at the advanced level, specifically 3rd- and 4th-year students at university whose major is English. It identifies overused and underused modal verbs by Chinese students and attempts to explain such findings. Two corpora are used to obtain evidence of modal verb use---the 1-million word FLOB and the ST6 component of CLEC (the Chinese Learner English Corpus), the latter a 120,000-word collection of free compositions by Chinese advanced learners of English. The concordancing tool WordSmith is used to find out the frequency of each modal verb and generate necessary keyword lists.

2.2 Overall findings

The following table presents frequency statistics about modal verbs found in FLOB and ST6, together with other information about the two corpora:

	ST6_freq.	FLOB_freq.	ST6_norming freq. count (per million)	FLOB_norming freq. count (per million)	ST6/FLOB norming freq. count ratio
total modal verb	6,102	15,279	25,003	12,347	2.02
ST6_tokens	244,055	FLOB_tokens	1,237,426		
ST6_types	11,733	FLOB_types	45,089		
ST6_type/token	4.81	FLOB_type/token	3.64		
ST6_standard	40.25	FLOB_standard	45.52		

Table 1: Frequency statistics of modal verbs in FLOB and ST6

Overall figures in the above table show that ST6 students use modal verbs twice more than NSs (ST6/FLOB ratio=2.02), or in other words, Chinese English learners at this level are twice more likely to express their opinions or attitude by modals. Empirical evidence supplied by other researchers is supportive to the results here, which points to the fact that Chinese learners of English do heavily rely on such auxiliaries in communicating modality to the extent that other expressions or grammatical constructions which also carry modal meanings rarely emerge or are even completely absent in their corpus (Yu, 2002). The above finding may also remind us of what Holmes said about the use of modals in English, that the total frequency of modal verbs used by native speakers is lower than that of any other word classes (Holmes, 1988).

As modals are often considered to convey subjective modality (Li Jie, 2002), their high frequency in ST6 gives a note of subjectivity to these Chinese learners' interlanguage. Messages presented in their writing thus sound rather subjective or personal.

2.3 A detailed picture

As overall statistics are often misleading, the following table provides a more detailed, and therefore a more meaningful picture of how modals are used by NSs and ST6 students:

modal verb	ST6_freq.	FLOB_freq.	ST6_norming freq. count (per million)	FLOB_norming freq. count (per million)	ST6/FLOB norming freq. count ratio
dare	21	18	86	15	5.92
must(mustn't)	198	816	811	659	1.23
ought(not)	9	58	37	47	0.79
need(needn't)	99	290	406	234	1.73
have(got) (has/had)	310	760	1,270	614	2.07
will('ll, won't)	1,170	2,741	4,794	2,215	2.16
would ('d, wouldn't)	402	2,690	1,647	2,174	0.76
shall(shan't)	17	200	70	162	0.43
should(shouldn't)	1,165	1,148	4,774	928	5.15
be(am/is/was/were)	178	497	729	402	1.82
may	281	1,190	1,151	962	1.20
might(mightn't)	66	642	270	519	0.52
can(can't, cannot)	1,939	2,458	7,945	1,986	4.00
could(couldn't)	247	1,771	1,012	1,431	0.71

Note: Instances where some words (e.g. 'may', 'can', 'will', 'might', and 'need') are used as nouns are excluded.

Table 2: Frequency statistics about each modal verb in FLOB and ST

The above numbers show ST6 students do not overuse all the modals, contrary to what we can infer from the overall ratio. They use some modals with a noticeably high frequency, but others less frequently than native speakers. To make things clearer, overused and underused modals are placed in separate tables:

modal verb	ST6_norming freq. count (per mil)	FLOB_norming freq. count (per mi)	ST6/FLOB ratio
dare	86	15	5.92
should(shouldn't)	4,774	928	5.15
can(can't/cannot)	7,945	1,986	4.00
will(won't/'ll)	4,794	2,215	2.16
have(got) to	1,270	614	2.07
be to	729	402	1.82
need(needn't)	406	234	1.73
must(mustn't)	811	659	1.23
may	1,151	962	1.20

Table 3: modal verbs _a contrast between ST6 and FLOB (ratio>1)

modal verb	FLOB_norming freq. count (per mil)	ST6 _norming freq. count (per mil)	FLOB/ST6 ratio
shall(shan't)	162	70	2.31
might(mightn't)	519	270	1.92
could(couldn't)	1,431	1,012	1.41
would(wouldn't/'d	2,174	1,674	1.30
ought	47	37	1.27

Table 4: modal verbs -- a contrast between FLOB and ST6 (ratio>1) ('d=would)

3. Analysis

3.1 Overused modals

Table 3 gives detailed information on overused modals in ST6. The biggest difference between FLOB and ST6 lies with *dare*, a marginal modal. Next comes *should*, which is over five times more frequent in ST6. *Can* takes up the third place, its frequency in ST6 four times that in FLOB. *Will* and *have to* follow then, their frequency differences notably smaller than the top three. Other modals that are also more often used in by these Chinese learners are *be to*, *need*, *must* and *may*.

3.1.1 Extent of overuse

To see the extent of this overuse, we may refer to a keyword list based on a comparison between FLOB and ST6, with a number of irrelevant keywords omitted from it:

N	KEY WORD	FREQ.ST6	ST6.TXT %	FREQ.FLOB	FLOB.LST %	KEYNESS
1	s	2,741	1.12	246	0.02	8,300.2
4	China	796	0.33	65		2,435.7
9	**can**	**1,529**	**0.63**	**1,772**	**0.14**	**1,599.3**
10	money	763	0.31	306	0.02	1,583.7
12	**should**	**1,143**	**0.47**	**1,115**	**0.09**	**1,397.0**
21	women	595	0.24	424	0.03	915.8
35	**will**	**1,167**	**0.48**	**2,284**	**0.18**	**617.2**
40	suffering	213	0.09	44		548.9

Note: The frequency counts of 'can' and 'will' include instances where the two words are used as nouns.

Table 5: keyword listing _source text=ST6, reference corpus=FLOB, total keywords=40, listed in the order of keyness of each keyword in the source text

Among the first 40 keywords in ST6, i.e. the top 40 words that are unusually frequent in this corpus compared with FLOB, *can*, *should* and *will* are found, with keyness value of 1,599, 1,397, and 617 respectively, a further proof that these modals are overused by ST6 learners.

Can, *should* and *will* are not only overused by ST6 learners, they are also exceptionally often in corpora of other levels of Chinese learners, as is shown in the following keyword list:

N	KEYWORD	FREQ.CLEC	CLEC.TXT %	FREQ.FLOB	FLOB.LST %	KEYNESS
1	s	17,843	1.48	246	0.02	23,033.9
5	**can**	**8,357**	**0.69**	**1,772**	**0.14**	**4,829.1**
10	do	5,655	0.47	1,388	0.11	2,884.8
20	**will**	**5,993**	**0.50**	**2,284**	**0.18**	**1,819.8**
21	commodities	1,329	0.11	5		1,816.5
28	**should**	**3,815**	**0.32**	**1,115**	**0.09**	**1,632.6**
40	get	2,486	0.21	655	0.05	1,184.2

Table 6: keyword listing _ source text=CLEC, reference corpus=FLOB, total keywords=40, listed in the order of keyness of each keyword in the source text

3.1.2 Value of the modals

If we assign value to the overused modals by ST6, displayed in Table 3, we may have the grouping on the left and the one on the right below when we give value to the underused modals shown in Table 4, those which natives speakers use more often than ST6 learners:

high: dare, have to, need, must	*high*: ought to
median: should, will, be to	*median*: shall, would
low: can, may	*low*: might, could

We can see among the 9 overused auxiliaries by ST6 learners, 4 are modals of high value and 3 of median. In contrast, 4 out of 5 modals often used by native speakers have low or median value. It can therefore be concluded that when expressing their opinions by modals, Chinese learners at the advanced level tend to choose modals of high and median value, while native speakers prefer those of median and low value. As some linguists observe, modal value is closely associated with politeness of the utterance (Li Jie, 2002). The higher the value, the more certain the speaker is about the proposition. The lower the value, the more modulated or tempered the utterance is and the speaker sounds more polite. Naturally the large number of high and median modals in ST6 gives such learners' English a touch of directness, self-assuredness and even brusqueness.

3.1.3 Implications of overuse

Among the 'full' modals, *should* stands out with its strikingly high frequency, over five times more often in ST6 than FLOB. Most texts in ST6 are persuasive by nature and as a result, most cases of *should* in these compositions also carry a persuasive tone and are often employed by these Chinese authors to express obligation or duty. However, *should* as a persuasive marker usually makes one's utterance over-direct and offensive (Zhang, 1981). The overuse of *should* in ST6 thus causes the same effect as too many high and median modals.

Readers of ST6 texts may end up with another feeling: these Chinese students seem obsessed with obligations and duties! Indeed, the majority of the overused modals by ST6 students are agent-oriented root modals, particularly *should*, *have to*, and *must*, which express obligation and necessity. The persuasive nature of most ST6 texts can partly explain the overuse of such modals (some topics assigned to the students actually contain *should*, like '*Should* euthanasia be practiced'), but more importantly, Chinese culture typically allows or even encourages the choice of such expressions when help and suggestions are to be offered, because under its norms, giving 'help' and 'suggestion' is a friendly and altruistic action and therefore the person who offers them rarely feels the need to be indirect (Yu, 2002). In Chinese a speaker would not hesitate to insert *yinggai*, or *yao* (both express obligation and necessity) between the subject and main verb, and consequently *should*, *must* and *have to*, their English equivalents, appear very often in an English corpus by Chinese learners. Cultural difference and negative transfer of L1 are another factor that justifies the overuse.

Both *can* and *may* are overused in ST6, but *can* is used with a remarkably higher frequency (7,945 vs. 1151 times). A bold conjecture is that *can* is often used in place of *may* in ST6 because both modals can communicate possibility and permission and the principal difference is stylistic, with *can* being more colloquial. Thus the preference of *can* over *may* gives a coiloquial feature to the texts by Chinese learners, a finding similar to Ma Guanghui's conclusion based on an analysis of 66 linguistic features in English compositions by American and Chinese university students that Chinese learners' writing is 'participatory' or 'colloquial' (2002).

3.2 Underused modals

3.2.1 Modals of politeness

As is shown above, underused modals in ST6 are mostly of low or median value (with the exception of *ought to*). It is generally agreed that 'past forms' of modals like *might*, *could*, and *would* are often used to express hypothetical meaning in both main and subordinate clauses. They tend to add to the utterance implications of tentativeness or politeness and what distinguishes them from the present forms is a greater degree of modalization rather than time (Quirk, 1985; Searle, 1975; Palmer, 1983; Qin, 1994; Lei, 2001). Compared to texts produced by native speakers, ST6 learners thus sound more assertive and yet less polite.

The statistics in Table 4 indicate that among all the underused modals *shall* is the rarest one in ST6. This might be accounted for by the fact that *will* instead of *shall* is often used by Chinese learners to express intention, which in turn explains the high frequency of *will* in the Chinese corpus. In the case of *would*, its low frequency in ST6 in comparison with English Corpora of Native Speakers (ECNS) is highlighted by a table of detailed statistics given by Gui Shichun (2002: 27), which shows the modal as the 107th in the CLEC frequency list, but within the first 60th in all the cited ECNS such as BROWN, LOB, WELL, AHI, LONDON, FROWN and BIRMINGHAM.

3.2.2 Acquisitional explanation

Sweetser (1990: 50) suggests that with English-speaking children, root meanings emerge earlier in language acquisition than epistemic ones. In other words, there is an acquisitional priority of root over epistemic meanings. After reviewing the studies and findings of many linguists, Papafragou (1998) concludes that in English, the use and mastery of *can* conveying ability and permission, *will* expressing intention, *must*, *have to*, *should* expressing obligation and necessity, appear much earlier than *may* and *might* with a possibility meaning (in this case we may include *could*, which often expresses possibility as well). She also notes that children's control over modal meanings of the same modal verb tends to extend from root to epistemic modality, and therefore in the case of *will*, its 'intention' aspect precedes its 'prediction' aspect in acquisition. Besides, studies of modal acquisition in other languages also reveal similar results. Ma and Bahetnisia (2002) conducted a survey among university students from ethnic groups in China's Xinjiang who are learning mandarin Chinese. They found for such students, learning to express tentative politeness is much more difficult and thus comes later than learning other modal meanings. As in English, tentative politeness is often conveyed by epistemic modals in Chinese. For learners of Chinese, communicating obligation and necessity is anterior to uncertainty and possibility in acquisition.

Papafragou also points out that epistemic modals in English are rather formal expressions and thus both emerge and are mastered at a later time. Among the deontic modals *ought to* lag behind all the others for its greater degree of formality.

The speculation here is, if there is an acquisitional anteriority of epistemic modals like *might* and *could* over root modals in English, then it is quite reasonable to predict that as an interim between L1 and L2, the interlanguage of Chinese learners should contain a larger number of root modals than epistemic ones. The corpus-based evidence above has confirmed this hypothesis and results from developmental corpora of Chinese learners of English may uncover further interesting findings.

3.3 Other explanations

In contrast to the heavy reliance of ST6 students on the auxiliaries, native speakers of English enjoy a wider

freedom in expressing modality. Linguistically they have more alternatives to communicate the same message. The standard type/token ratio in each corpus is a good proof. With ST6 it is 40.25, but 45.52 with FLOB. Naturally due to the smaller size of their vocabulary, ST6 learners are forced to restrict their choice to particular, perhaps more familiar modal verbs.

Another reason of modal verb overuse among the Chinese learners may be their inability to fully comprehend and thus manipulate modal nuances in English. For instance, although *might*, *could* and *would* are past forms of *may*, *can* and *will*, their major difference is not time but degrees of modalization. However, meanings in each of these pairs are represented by a single modal in Chinese like *keneng*, *neng* and *yao* (or perhaps *yuan*), the latter group making no semantic or modalized differences. We may consider this an effect of either L1 negative transfer or problematic areas in L2 itself.

4. Conclusion

Modality is an important component of a language. As an overt means of expressing modality, modal verbs contribute to the felicity of the utterance (Chen Guangwei, 2001). The choice and use of modals reflect the speaker's attitude and mentality and therefore determine the distance between the interlocutors and may affect the speaker's image. In cross-linguistic exchanges improper use of modal verbs can thus be an obstacle to fluent communication. Nevertheless, English modal verbs involve lots of variations and complications and therefore prove to be a 'problematic' area of the English grammar (Quirk, 1985: 220). Understandably Chinese learners at all levels tend to make more mistakes in this area. Statistics indicate that 105 out of all the errors in ST6 involve the use of modal verbs (Gui, 2003:711), which may partly explain both the communication failures that happen in conversations between Chinese students and native speakers of English, and the gap between the interlanguage of the Chinese learners and the target language of English.

References:

Bussmann, H. 1996: *Routledge Dictionary of Language and Linguistics*. Beijing: Foreign Language Teaching and Research Press.

Chen, Guanglei. 2001: *Morphology of Chinese*. Shanghai: Xuelin Publishing House.

Chen, Guangwei. 2001: Semantic Classification and Pragmatic Analysis of English Modals.*Journal of Guangxi Teachers College*, No. 4, Vol. 22.

Chen, Wangdao. 1978: *A Brief Introduction to Chinese Grammar*. Shanghai: Shangghai Education Press.

Ding, Shengshu. 1999: *Talks on Modern Chinese Grammar*. Beijing: The Commercial Press.

Gui, Shichun, Yang Huizhong. 2003: *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.

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

Holmes, J. 1988: Doubt and Certainty in ESL Textbooks. *Applied Linguistics*, No. 9, 21-44.

Jacobsson, Bengt. 1979: Modality and the Modals of Necessity: Must and Have to. *English Studies*, No. 3, Vol. 60, 296-312.

Lei, Zhimin. 2001: The Uncertainty in the Use of English Modal Verbs. *Journal of Hunan Agricultural University*, No. 2, Vol. 2.

Liang, Xiaobo. 2002: An Overview of Studies of Modality. *Journal of PLA University of Foreign Languages*, No.1, Vol. 25, 28-33.

Li, Ji'an. 1999: Modal Meanings and Modal Verb Meanings. *Foreign Languages*, No. 4, Vol. 122, 19-23.

Li, Jie, Zhong Jiaping. 2002: The Modal System of English and Its Functions. *Foreign Language Education*, No. 1, Vol. 23, 9-15.

Lu, Shuxiang. 1982: *An Outline of Chinese Grammar*. Beijing: The Commercial Press.

Ma, Cuiling, Bahetnisa. 2002: Study on Tone and Mood in Chinese Teaching. *Language and Translation*, No. 2, Vol. 70, 61-63.

Ma, Guanghui. 2002: Contrastive Analysis of Linguistic Features Between EFL and ENL Essays. *Foreign Language Teaching and Research*, No. 5, Vol. 34, 345-349.

Modern Chinese. 2000. Beijing: The Commercial Press.

Palmer, F. R. 1983: Semantic Explanation for the Syntax of the English Modals. In F. Heny et al (ed) *Linguistic Categories: Auxiliaries and Related Puzzles*, Vol 2, 29-43. Holland: D. Reidal Publishing Company.

Papafragou, A. 1998: The Acquisition of Modality: Implications for Theories of Semantic Representation. *Mind and Language*, No.3, Vol. 13, 370-399.

Qi, Huyang. 2002: Analysis of the Function of Modal Particles in Modal Category. *Journal of Chinese Language and Culture Nanjing Normal University*, No. 3, Sept., 141-152.

Qin, Yuxiang. 1994: The Meaning, Semantic Features and Tense of English Modal Verbs. *Foreign Languages*, No. 2, 37-44.

Quirk, R., Sidney Greenbaum et al. 1985: *A Comprehensive Grammar of the English Language*. London: Longman Group Ltd.

Searle, J. 1975: Indirect Speech Acts. In Cole, P. et al. (eds) *Syntac and Semantics*, Vol. 3. New York: Academic Press.

Siewierska, A. 1991: *Functional Grammar*. London: Routledge.

Sweetser, E. 1990: *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. London: Cambridge University Press.

Wang, Li. 1950: *Theories on Chinese Grammar*. Vol. 1. Beijing: The Commercial Press.

Yu, Zechao. 2002: Exploring Politeness in Written Argumentation: an Interlanguage Study of Modality. *Journal of Zhejiang Normal University*, No. 3, Vol. 27, 50-53.

Zhang, Jin, Chen Yunqing. 1981: *An Outline of Chinese-English Comparative Grammar*. Beijing: The Commercial Press.

Teach Science Students Collocation — Make it a practice

Wang Xiuwen Zheng Shutang Guo Hongjie
Shanghai Jiaotong University, 200030, China

Abstract: In recent years collocation has emerged as an important category of lexical patterning and it is fast becoming an established unit of description in language teaching courses and materials. Teachers should teach students collocation. Scientific students in particular should learn collocation and chunks, because technical language has a greater tendency (than creative writing, for example) to use fixed chunks. This paper makes an initiative analysis as to the content (particularly) and method relating to teaching science students collocation in class.

Key words: collocation, science students, corpus, frequency list

Introduction

We have *mistakenly* been teaching vocabulary in the form of isolated words in the past——partly because we did not have the benefit of evidence from corpus. Also, old exercises often asked students to replace one word with a near-synonym. Unfortunately when you look at corpus data, you find that the near-synonym has different collocates or grammatical/structural patterns. Meanwhile, dictionaries are forced by their format to give the impression that meaning is generated by individual words. Whereas it is very clear from corpus evidence that meaning comes from words in *context*. And, it is known to all that children learn chunks. They keep producing advanced phrases and sentences (not always exactly correct, but usually very close). Most modern EFL dictionaries and many recent bilingual and even native-speaker dictionaries are increasing their coverage of phrases and other chunks. Teachers should teach students collocation. Scientific students, in particular, should learn collocation and chunks, because technical language has a greater tendency (than creative writing, for example) to use fixed chunks as Bo Svensen (1993:49) puts it "...Technical language resources often emphasize terms, rather than other aspects". A major problem with printed reference books, especially in technical domains, is that technology changes very quickly, so the books soon become out-of-date. Fields like computing really require a new edition every year or so, whereas editions of general language dictionaries usually appear at much longer intervals. Teachers can help science students to identify and learn the chunks by looking at corpus data, which updated every now and then.

1. The Lexical Approach to the study of collocation

The lexical approach is based on the assumption that words receive their meaning from the words they co-occur with. These linguists, Firthians in particular, perceived collocations as a lexical phenomenon independent of grammar.

> "*You shall know a word by the company it keeps*" (Firth, 1957:12)
> "*...Lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, ether a scale or at least a cut-off point. It is this syntagmatic relation which is referred to as 'collocation'.*" (Halliday, 1976:75)

"Collocation is the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words in intervening". (Sinclair, 1991: 170)

2. Features of academic and scientific writing

Examining a typical academic text on architecture, we noticed that the style of writing involved long, complex sentences, containing many technical terms (e.g. *motif, relief, figure, plane, contour, composition*). Not only the words, but the phraseology and expressions (e.g. *represents a borrowing from...*) are typical. This kind of nominalization (cf. *borrows from*) is a common feature.

Halliday and Martin, in *Writing Science* (Falmer Press 1993), discuss the main features of academic writing:

1) Not just terminology, but `wording'... technical grammar deploys nominal groups and clauses in rhetorical structures to form arguments.
2) Verbs and adjectives are transformed into nouns, which allows "new" information from previous discourse to be reused as "given".
3) Extending the nominal group, using prepositional phrases, embedded clauses, and recursion. The noun group generates "objectivity" which allows reasoned argument.

According to Halliday and Martin, the characteristics of scientific English are:

1) interlocking definitions (e.g. *This distance is called the radius. The diameter of a circle is twice the radius.*)
2) technical taxonomies (superordination, composition; e.g. suffixes such as -berry, -fish, etc)
3) special expressions (technical grammar; e.g. *The process of finding the truth set is called "solving the open sentence over D".*)
4) lexical density (i.e. the number of lexical items or content words per clause)
5) syntactic ambiguity (e.g. *Lung cancer death rates are associated with/reflected in smoking*. Does this mean they are the cause or just the evidence? Does rate mean "number" or "speed"?)
6) grammatical metaphor (i.e. one class or structure is substituted for another, abandoning the traditional association of verbs with processes, nouns with participants, adjectives with qualities, adverbials with circumstances, conjunctions with process relations, and modals with assessment; e.g. *he departed > his departure.... unstable > instability*)
7) semantic discontinuity (i.e. scientific writing makes semantic leaps which readers are expected to follow)

3. Corpus evidence

Technical language has a greater tendency to use fixed chunks. By comparing a General Corpus (e.g. the Bank of English, which contains written and spoken texts from many different genres) with a Science/Technology Corpus such as JDEST (consisting of only journal articles, textbooks, and newspaper articles on science/technology). From the following comparison (in 4) we can see that: a) chunks like "by reference to" and "seek to provide" will be *more frequent* in a Business Corpus than in a General Corpus, but may also be common in corpora of other *formal* genres. b) chunks like "seek to provide a high level of current income exempt from federal income tax" will occur *very rarely* in a General Corpus, and will generally be found *prominently* only in a Business Corpus, not in any other corpora of formal genre. All this is the case in Business English, but also holds in Scientific English.

4. What to teach

Teachers can choose collocations from a frequency list taken from an appropriate corpus, which consists of the type of text that teachers want their students to read/produce. For example, if we are teaching "science and technology", presumably the students should be able to read journal articles, textbooks, newspaper articles, etc. So the corpus should consist of journal articles, textbooks, newspaper articles, etc. And teachers can use **Wordsmith Tools**, which is easy to deal with, to make frequency list. For example, from the 10-million-word Wolverhampton Business Corpus, the most frequent 2-word chunks are:

1	of the	6	by the
2	in the	7	on the
3	to the	8	and the
4	the fund	9	the company
5	for the	10	of a

These chunks would obviously be part of the students' general English teaching, and not especially significant for Business English. However, if you look at the 5-word chunk frequency list, several important phrases appear:

13	there can be no assurance	15	exempt from federal income tax
14	purchase at net asset value	19	the market value of

10-word chunks:

473 incorporated into this filing by reference to post-effective amendment no.
397 fund seeks to provide a high level of current income
384 high level of current income exempt from federal income tax
384 a high level of current income exempt from federal income
380 to provide a high level of current income exempt from
380 seeks to provide a high level of current income exempt

As we can see, there is much repetition and overlap in these chunks, so it is up to the experienced science teacher to extract the chunks that they wish to teach. For example, (in the above list) "incorporate something into something", "by reference to", "seek to provide", "a high level of income", "exempt from tax", etc. This can also be done by reference to the texts the students are studying: find the sentences that contain chunks similar to the ones found in the frequency list. Because they are clearly important to the whole *genre* (in this case Business English, but the principle is equally valid for Scientific English) beyond the actual text the student is looking at. Or we can start with the text, and see which chunks in the text are also frequent in the corpus, and teach those chunks. Otherwise, if we deal too specifically with the text only, we may emphasize features of the text which are *not* typical of the whole genre. Some degree of discretion must be left to the teacher at the beginning. If several teachers can use the same corpus and the same text, they can discuss which chunks the students found difficult, which examples they could understand easily, which method of presenting the chunks was most effective, etc.

For comparison, here are similar chunk frequency lists from a corpus of Junk Emails, which Ramesh Krishnamurthy has just created with a colleague at Wolverhampton University.

2-word chunks:

937	if you	621	you can
917	of the	594	you will
911	in the	564	on the
774	do not	528	to the

444 this be
443 for you
441 will be
435 be a
403 to you
396 you have
389 you be
365 be the
331 to the

I have separated the chunks which do *not* appear so frequently in the Business Corpus: "if you, do not, you can, you will", etc. Notice that these are more conversational phrases. Junk Emails imitate the spoken genre more than the Business Corpus.

The following is a direct comparison of 3-word chunks (NB the first two lists are lemmatised chunks: i.e. wordforms are reduced to their root form) in 3 corpora: junk emails, leaflets, and the British National Corpus. Personal pronouns are in bold, verbs in italics. We can see that the 3 corpora are quite different in many respects, but some chunks (underlined) are common in all genres.

Junk	Leaflets	BNC
254 to *be remove*	171 if you *be*	17398 one of the
226 you *do* not	114 one of the	9855 the end of
195 *be remove* from	111 *be able* to	9682 as well as
183 in the subject	102 there *be* no	8279 I *do* n't
166 if you *do*	92 part of the	8105 part of the
145 if you *be*	91 you *will* find	7819 there *is* a
139 on the internet	89 there be a	7479 some of the
130 you *will be*	86 the regional council	7478 out of the
115 the subject line	85 you *will be*	6602 a number of
114 *remove* in the	84 you *do* not	6592 end of the
110 if you *have*	77 if you *have*	6222 it *was* a
105 *would like* to	71 to *help* you	6060 there *is* no
102 *remove* from we	68 the number of	6020 the fact that
99 one of the	64 the end of	6008 there *was* a
91 you *want* to	61 it *be* a	5889 *be able* to
86 you *would* like	59 a number of	5645 to *be* a
85 *click* here to	58 *have to be*	5511 in order to
83 if you *would*	52 to *ensure* that	5478 it *is* not
82 *be able* to	51 you *want* to	5400 per cent of

Table 1

So the leaflets are more similar to BNC (general English)-they have 8 chunks in common. Junk Emails are very different——it has only 2 chunks in common with Leaflets and BNC.

Obviously, we have only looked at the 20 or so most frequent chunks in this comparison. One would do comparisons between the complete frequency lists before teaching scientific chunks.

Sinclair and Renouf (1988) agree that the most frequent words are not necessarily the most useful for learners. The statistical selection parameters for choosing words are: frequency, coverage and distribution (Yang Huizhong, 2002:23). Since the three parameters are not in linear relationship with each other, we should take all of them into consideration when considering which words should be taught.
The following table (Yang Huizhong, 2002:26) is the statistical characteristics of three types of words.

Functional words	sub-technical words	special terms
Very high frequency Very high distribution	sub-frequent words	very high frequency very low distribution

Table 2

So functional words and sub-technical words should be covered in teaching.

Also, when choosing words for teaching, we should consider the following factors:

- specialist' subjective judgement
- social criteria
- language teaching criteria
- linguistic criteria

5. How to teach

Like any vocabulary teaching, the teaching of collocation depends on the time available, the length of the course and the level of the students.

On the whole, researchers like Professor Sinclair (1991) have shown that *significant* collocations usually occur within four or five words. So one strategy for teaching would be to build up gradually, from two word chunks to three-word chunks, etc. The maximum span will vary from one word to another.

In a word, Collocation can be approached in a variety of ways. At the advanced level, McCarthy et al. (1985: 158) offer straightforward gap-filling; at the intermediate level, Redman and Ellis (1989) also have collocation activities.

Pearson and Johnson (1978) suggest that concepts are not randomly related but follow predictable lines. Word-association phenomena bear this out. Also the semantic field theory reflects the general linguistic tendency to move from an isolating, atomistic, discrete view to a holistic, systematic approach. So we can first make a *word-list* using Wordsmith tools, and then *key-key words* analysis by comparing the word-list with the text the students are studying, and finally make *associate* analysis by seeing which words can be semantically categorized together. For instance, in Chinese, if we want to find ***gang tie***, we usually first find *gang*, and then look for *tie*. However, if we are wanted to find the same category as ***gang tie***, we won't follow the practice above. We usually do the search basing on the semantic field. That is to say, we can give a *topic* and a *word-list* to the students, and then try to categorize the word, including collocation.

And in the case grammar introduced by Fillmore (1968), the most important cases in the Fillmore model are agent, instrumental, objective (later termed patient or goal), benefactive and the location of the action. According to Fillmore, when there is an action, there must be an *agent* doing the activity and there must be the *tools*, also *beneficial* and *suffer*. Hence, we can subjectively make *network associate* using Fillmore's case grammar theory for teaching students lexis, including collocation. For example, we can ask students to write a composition on the topic *job-hopping*, and give them the following figure for reference, through which teach students lexis, including collocation.

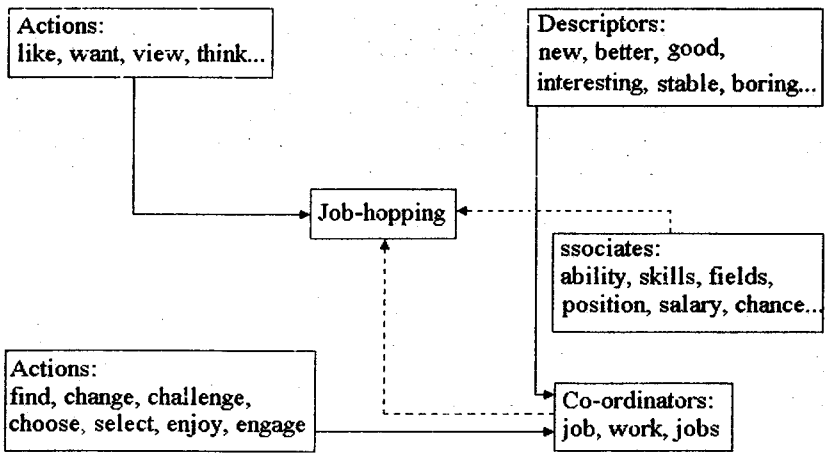


Figure 1

Summary

Teachers play a prominent role in helping learners identify collocations on class. This seems to go against the idea that encourages a student-centered, exploration approach to language learning. While Morgan Lewis (2000) thinks that although learners should take responsibility for their own learning, they should not be taking responsibility for choosing which language items are more linguistically useful. Swan (1996) also points out that vocabulary will not take care of itself. Students with time limit for learning will not learn **high priority lexis** if it is not deliberately selected and incorporated into classroom materials and activities. So it is teachers' job to provide students with the most common and useful collocations based on their professional understanding of both language and learning. Interestingly, after a period of teacher-dominated ——more accurately learner training, learners begin to notice more of this kind of language for themselves, thus becoming more autonomous in their approach. It can be maintained that this kind of teaching serves as a basis for students' own discovery and study of collocation.

Bibliography

Fillmore, C.J. (1968) The Case for Case. In E. Bach and R.T. Harms (eds) *Universals in linguistic theory*. New York.

Firth, J.R. (1957) Modes of Meaning, in F.R. Palmer (ed) *Papers in Linguistics 1934-51*, Oxford University Press, London

Halliday, M.A.K. (1976) *System and Function in Language*, ed. G. Kress, Oxford University Press, London

Halliday, M.A.K. and Martin, J.R. (1993) *Writing Science: Literacy and Discursive Power*, Falmer Press, London

Johns, T. *Microconcord*: http://web.bham.ac.uk/johnstf/

Lewis, Michael. (2001) *Teaching Collocation: Further Development in the Lexical Approach*. Commercial Colour Press plc, London.

McCarthy et al. (1985) *Proficiency Plus: Grammar, Lexis, Discourse*. Basil Blackwell, Oxford.

Redman and Ellis (1989) *A Way with Words. Book 2*. Cambridge University Press, Cambridge.

Scott, M. *Wordsmith Tools*: http://www.lexically.net/

Sinclair, J.M. (1991) *Corpus, Concordance, Collocation*, OUP, Oxford

Sinclair and Renouf (1988) '*A Lexical Syilabus for Language Learning*' in Carter and McCarthy 1988.

Svensen, B. (1993) *Practical Lexicography*, OUP, Oxford

Swan, M. (1996) "Language Teaching is Teaching Language" in Plenary address to IATEFL. *IATEFL Conference Report*. P34-38.

Yang Huizhong. (2002) *An Introduction to Corpus Linguistics*. Shanghai Foreign Language Education Press, Shanghai.

Using Learner Corpus Research in Teaching Writing

Ding Man
Dalian University of Technology

Abstract: This article aims to demonstrate how to use learner corpora research in EFL writing class from ordinary EFL teachers' point of view. This involves 3 stages: 1) Building of a learner corpus; 2) Analysis of the learner corpus by using software; 3) Materials development based on the learner corpus and a general NS corpus. The learner corpus that present study is based on, comprises 49 compositions written by Chinese first-year college students, whose major is science or engineering, and *WordSmith Tools* are used to obtain frequency list and concordance lines from the learner corpus. The analysis of the learner corpus starts from the frequency list of the learner corpus, to in-depth investigation of a frequent word *'in'*, which is classified into three groups. Detailed investigation of *'in'* in this study sheds some light on the understanding of learner English for usage of the search word. One striking feature of learners' use of *'in'* is that category 1 (prepositional phrases of *'in'*) is significantly more frequent than the other two categories (fixed expressions with *'in'* and language patterns with *'in'*) in the learner corpus and its percentage of occurrences accounts for 80% of all the occurrences of *'in'*. Analysis of prepositional phrases shows that the dominating uses of *'in'* by EFL learners are its indications of place and time, which might conform to most EFL teachers' intuition. However, the variety of learners' uses of *'in'* in this category might be beyond teachers' expectation. To some extent, the detailed findings of the search word in this research help to give us a general picture of the use of *'in'* of the learner corpus. As a result, the insights into learner English may help EFL teachers to develop materials, which take the learner corpus as the base and the Bank of English as the reference.

Key words: DDL, learner corpora, interlanguage

Introduction

For years EFL teachers of writing have been aware that problems exist in the traditional way of writing class, in which teachers spend a lot of time correcting learners' errors. It seems that the time-consuming job does not help the learners to improve their writing, but teachers have no other choices to change the situation. With the initial attempt to find a method to help EFL teachers respond to learners' writing in an efficient and effective way, this article, which is based on a mini learner corpus that consists of 49 compositions of first year college students, tries to explore the possibility of using classroom-based small corpora to complement writing class. This study does not focus on learner error, but would like to take learner English as the starting point to investigate its features and see how materials could be developed based on insights into learner English that we could gain from the study of learner corpora. Though classroom concordancing or data-driven learning (DDL) has been accepted as an innovative approach to EFL teaching for a decade, it seems that its application to the EFL writing class is limited. The present study attempts to combine learner corpora with DDL to see how well learner corpora and NS corpora could be integrated to promote DDL.

Literature Review on Learner Corpora

Leech (1998:xiv) defines learner corpus as 'a corpus, or computer textual database, of the language produced by foreign language learners'. According to Granger (ed.1998), the building of learner corpora began at the early 1990s. The three best-known learner corpora are the International Corpus of Learner English (ICLE), the longman Learners' Corpus (LLC), and the Hong Kong University of Science and Technology (HKUST) learner corpus.

In the investigation of learner corpora, the contrastive approach, which includes both comparison of native language (NL) and interlanguage (IL) and comparison of different interlanguages, is the dominating approach. Granger (ed.1998) claims that NL/IL comparison aims to uncover the features of non-nativeness of learner language. Papers in Granger report their research done on the comparison between the ICLE (which comprises argumentative essays produced by advaced learners) and the LOCNESS (Louvain Corpus of Native English Essays, which is a 300,000-word corpus of essays written by native university students). The main contribution of earlier investigation of learner corpora in lexis is the findings of overuse and underuse. By comparison between NNS (non-native speaker) English and NS (native speaker) English, researchers are able to find out the words that learners use significantly more often or less often than native speakers. These two groups of words are called overused and underused respectively.

Though the study of overuse and underuse of interlanguage is quite informative in some sense, the drawback to Granger's approach is that it assumes that learners have native-speaker norms as a target (Hunston, 2002). Overemphasis on overuse and underuse might frustrate learners when they are in the process of learning in that it takes time for them to digest the knowledge they have learned. In fact, learner corpora not only give us a chance to find learners' errors, but also give us a chance to understand the process of learning.

Therefore, learner corpora should not be treated as the source of learner errors only, but the starting point for us to exploit together with learners, for the purpose of knowing what learners have mastered. Information about learner English is likely to enable classroom teachers to find ways to complement effective materials that are appropriate to learners' level. In addition, learner corpora could stimulate students' interest to know their own language and how to improve it. Seidlhofer (2000:222) reports her successful experience of using learner corpora in her corpus linguistics class and she interprets the key to success as the fact that they have a secure "home base" through focusing on familiar, non-threatening texts, not decontextualized bits of language from "remote native corpora".

In order to introduce learner corpus research to EFL writing class, one solution to the methodology problem I suggest for EFL teachers is to get rid of the idea of comparison first. Since it is likely that EFL teachers build their own learner corpus and get access to a large general corpus, it might be worthwhile to encourage EFL teachers to work with these two corpora, though they might not be comparable in many respects. For the building of a learner corpus, EFL learners could send their writing to their teachers by email or save it in floppy disk for teachers' use (for details see Barnbrook 1996:28-41). Even the most difficult method, which needs some people to type the learners' writing, is within teachers' ability. For the large general corpus that EFL teachers need, any general native speaker corpus could be valid, such as the British National Corpus (BNC), the Brown Corpus, the LOB and the Bank of English. It might be more suitable if one sub-corpus of these large corpora is used to be the source of the expert corpus.

Since a small learner corpus and a large expert corpus are not comparable in many aspects, comparisons between the two should be carried out in a limited way, and it is not appropriate to adopt the quantitative

and prepositional phrases (fixed expressions: e.g. *in a word, in addition;* prepositional phrases: e.g. *in England, in 21 century).* The results of the general classification are given in Table 4.

Table 4: The results of the classification of ***in*** of the DUT learner corpus

Category	Number of occurrences in the DUT learner corpus	Percentage of the total occurrences of ***in***
1. Prepositional phrases	220	80%
2. Fixed expressions	30	10%
3. Used after some verbs, nouns, and adjective in order to introduce more information (used in language patterns)	30	10%
Total	280	100%

One striking features of Table 4 is that category 1 (prepositional phrases) is significantly more frequent than the other two categories in the DUT learner corpus and its percentage of occurrences accounts for 80% of all the occurrences of ***in***. The reason for the high frequency of prepositional phrases might be the flexibility and L1 transferability of this category. By flexibility, I mean this category includes many usages of ***in*** that can be used in different semantic groups. For example, ***in*** can express 'within the limits, bounds, or area of' (e.g. *in the spring, in America*); it can also express 'at a situation or condition of' (e.g. *in debt, in love*). By L1 transferability, I mean some usages of ***in*** can be translated directly into Chinese, so it will be easy for the learners to use it freely. In order to have a detailed look of the learners' usages of ***in***, analysis of the three categories is carried out in the study. In the following, I will show my analysis of prepositional phrases.

2. Sub-categories of prepositional phrases

In order to have a detailed look of the uses in this category, the prepositional phrases are divided into 10 sub-categories (see the following Table).

Table 5: Sub-categories of prepositional phrases

Sub-categories	**Examples from DUTLC**	**Number of occurrences**
1. Indicating the place in which something happens (concrete, abstract place or metaphorical use).	e.g. *in America* *in the classroom* *in such a society* *in their eyes*	110
2. Indicating the time when something happens (in a period of time or a particular situation).	e.g. *in 1886* *in human history* *in our life*	47
3. Indicating that somebody is in something such as a play or a race, which means that s/he is one of the people who take part.	e.g. *in the school contest* *in the exam*	21
4. Indicating relation, reference, or respect.	e.g. *in most respects, in a way*	12
5. Indicating something is in a book, film, or picture you can read it or see it there.	e.g. *in the article, in the story* *in his works*	11
6. Indicating physical surrounding, circumstances.	e.g. *in the dark, in that case* *in bad circumstances*	8
7. Indicating state or condition.	e.g. *in panic, in trouble* *in poor health*	6
8. Indicating a general subject or field of activity.	e.g. *in a certain field* *in our national industry*	3
9. Indicating that you are wearing a piece of clothing.	e.g. *in dark color*	1
10. Used before relative pronoun.	e.g. *a world in which people...*	1
	Total	220

Table 5 shows that the dominating uses of ***in*** by learners are its indications of place and time, which accounts for 71%. This might conform to most EFL teachers' intuition. However, the variety of learners' uses of ***in*** might be beyond teachers' expectation, though the incidences of some sub-categories are much fewer than the first two dominating groups. Furthermore, some incidences of ***in*** demonstrate quite sophisticated language use (e.g. *in panic, in trouble*).

It is not difficult to interpret the high percentage of the first two sub-categories. They are quite transferable from Chinese to English. Though the first two sub-categories account for the majority of the prepositional phrases, it would not be advisable for EFL teachers to advise learners to use fewer prepositional phrases in that teachers might frustrate learners' creativity in the categories they have mastered quite well. In these two sub-categories, some learners not only have mastered the uses of indications of concrete and abstract places (e.g. *in America, in his society*), but also mastered metaphorical use (see Figure 1). Such examples could be more easier for co-learners to master.

Figure 1: Metaphorical use of '***in***' in DUTLC

```
·rate story for it teaches me what to do in the battle of survival.    (5)
At I want to. That is, there is no fair in the battle of survival. The strong pe
Ed and there are only benefit and money in their eyes. We shouldn't become this
```

Another feature of the first two sub-categories is that there are many typical frameworks, which consist of the same beginning and ending words. (e.g. *in the world, in a world, in this world* and *in a cold world*).one function of the *wordlist tool*, which is to show the 3-word cluster in the text, could help us find frequent 3-word clusters. However, the function is unable to reveal more-than-3-word patterns. So manual work is still needed in the search for the patterns with the same beginning and ending words.

Table 6: Frameworks with preposition ***in***

Frameworks	Number of Occurrences
In + ? + world	19
in + ? + exam(s)/examination	16
in + ? + society	11
in + ? + life	11
in + ? + future	7
in + ? + heart(s)	7
in + ? + time(s)	5
in + ? + body	4
in + ? + mind	4
Total:	84

Note: ? stands for one or more than one word.

Observation from Table 5 and Table 6, tells us that the frameworks plus the three 3-word clusters (*in my opinion, in the story* and *in that case*) contribute nearly half of the prepositional phrases (altogether there are 97 such kind of frameworks). Some frameworks appear quite frequently in most learners' writing, no matter what topic their writing concerns. Indeed, these frameworks are quite useful in writing, but another psychological factor might explain the high frequency of the frameworks. Manipulating something they are sure of gives the learners a sense of security and self-confidence, even though sometimes the frameworks are not necessary.

3. Learners' problems of the usages of prepositional phrases

Though the first two sub-categories of prepositional phrases are probably easier for Chinese learners, they sometimes tend to use ***in***, when ***on*** should be used partly because of L1 interference (see Figure 2). Chinese

learners tend to use the general rules of ***in*** and ***on*** to infer which one is correct in the context. They know that ***on*** refers to '*covering or forming part of a surface*' while ***in*** refers to that '*something is surrounded by something else*'. However, these rules are not applicable in some cases. Take *in campus* for an example, native speakers might also agree that ***in*** is logically more apt than ***on***, but conventionally native speakers use ***on campus***. Such kind of problem not only frustrates learners but also frustrates EFL teachers. Teachers are confident when they are able to use rules to explain language phenomena, whilst when teachers simply tell their students that native speakers tend to use it that way, they might experience difficult time with their students who cast doubt on the teachers' ability. When choosing the preposition before *internet*, Chinese learners might experience the tough analysis whether it is on the surface of internet or inside the internet. For problems like these, authentic language from NS corpora might do the job to give correct and convincing exposure.

Figure 2: Concordance lines of misused ***in*** (1)

ms. Nowadays, people should understand **in** campus we and they have to do this. W
Ked me to tell him everything happening **in** the campus. And I'm sorry that as you
and two of us are abroad. When we meet **in** the internet we always result the bea

We can also see that learners try to manipulate language based on complex hypotheses of language rules (see Figure 3). It is likely that the student who made the sentence have known sub-category 7 (*e.g. in panic, in trouble*. For details see Table 5) quite well. Even though '*in fatigue and hunger*' is not used in native speakers' English, the learner's attempt to create the target language is valuable. Teacher's attitude toward this kind of error is crucial in the learning process of EFL learners. If learners' active use of language could be discovered and guided in the right way, learners might be motivated to go on learning. Correction of such errors is to learners' benefits if they are abie to consult a NS corpus.

Figure 3 : A concordance line of Misused ***in*** (2)

between a man and a wolf. Both them were **in** fatigue and hunger. Who fell first wo

As a whole, from the evidence obtained from the DUT learner corpus, we can see that within the common usages of prepositional phrases, there are a great variety we may not be able to reach without the help of learner corpora. The analysis of prepositional phrases helps us gain insights into the learner language and with the help of an expert corpus teachers will be confident to assist learners in an effective way.

Materials Development

As has been discussed, EFL teachers need to build their own learner corpus and a general NS corpus (in the case of this study, the Bank of English is used). Besides a learner corpus and a general NS corpus, I also make use of two reference books: 1) Collins COBUILD Grammar Patterns 1: Verb (Francis et al. 1996); 2) Collins COBUILD Grammar Patterns 2: Nouns and Adjectives (Francis et al. 1998). These COBUILD series, based on the Bank of English, present the structure of English in an innovative, user-friendly way. Adequate use of COBUILD series is time-saving. Alternative use of COBUILD dictionaries and the Bank of English might enable EFL teachers to develop convincing materials efficiently.

There are three steps of materials development. The first step of materials development is to get concordance lines of a search word and classify the uses of the search word. (as discussed in previous chapter). Classification of a search word helps EFL teachers to arrange learner English in a systematic way. The second step is to consult COBUILD Grammar patterns series and the Bank of English. Three things could be done: 1). Check whether learner English correct or not; 2). Know to what extent learners have

mastered the different categories of the search word; 3). Select the language points that learners have not mastered. The third step is to design activities. When designing activities, EFL teachers should start with concordance lines from learner corpora, and draw the learners' attention to the language points of interests. It is to learners' benefits if concordance lines from a NS corpus are provided at the same time in activities. More authentic language exposure might help the learners be aware of the defect of their interlanguage, and in the long-run, acquire the correct way of using language.

Sample extracts

In this part, sample extracts from the learner corpus and the general corpus will be introduced to demonstrate the process of materials development.

Sample extracts can be used to help learners to extend their knowledge of the usages they are not familiar with. It starts from the good examples of a few learners (see Figure 4). Most probably good examples are not enough in quantity and variety, so still concordance lines from NS corpora are necessary (see Figure 5-7). In fact, more examples of this pattern (***V + in + n***) could be obtained by consulting Collins COBUILD Grammar Patterns 1: Verb. During complementation of new materials, one factor should be paid attention to. EFL teachers should follow the rule of learning, which is from easy to difficult. For instance, in table 7, the first three groups that I choose are groups that include examples in the DUT learner corpus (see underlined verbs in table 7). Such kind of materials facilitate learning on one hand, on the other hand it gives certain degree of flexibility to individual learner.

Figure 4: Concordance lines of the pattern (***V + in + n***) from the DUTLC

-Sony. As a young collage student major **in** electronic engineering, we are all fa
Eacher had forced the girl to take part **in** the competition regardless of disease
D courage to experience it again. I put **in** a lot of energys and perspiration on
R or own feelings. Which I think result **in** the crisis of marriage. What's more,
eel really sorry that he didn't succeed **in** the college entrance examination. He

Figure 5: Concordance lines of ***put in*** From the Bank of English

the inside of a large baking dish. **Put in** the potatoes first, then the
silver plate to. This cutlery can be **put in** a dishwasher, but never wash
dish. <p> 2 Chop meat into cubes, **put in** the dish and cover with the
You have to learn how to do it, then **put in** lots of practice." <h> CHIC AT
Ask if there was any whisky for him to **put in** his tea. He'd tell unsuitable
we're only doing alright because we **put in** the hours. <p> We got asked the
From £400 in 1989). <p> <c> PHOTO </c> **Put in** context, the `Level 5" penalty of
Hounds quickly found in Birch Wood and **put in** a lot of work between there and
Technically a rest day, the team still **put in** a couple of hours on the bike, to

Figure 6: Concordance lines of ***result in*** From the Bank of English

In classrooms would automatically **result in** their professional growth. In
Involved. Screening processes that **result in** just the right person for the
Of the viewer coherent patterns that **result in** the creation of new aesthetic
Year in Britain, domestic accidents **result in** 5000 deaths (equalling road
maybe so, but why does it have to **result in** prose like this? `We played
of racism, sexism and violence **result in** punishment for offending
On Jewish people in 1990 (rightly) **result in** a series of BBC programmes on
Out by each heartbeat, which may **result in** lower blood pressure. <p> 2)
weight loss is too drastic, it can **result in** a loss of vital lean tissue, a

Figure 7: Concordance lines of ***succeed in*** From the Bank of English

In Liar's Poker. If they **succeed in** acquiring Canary Wharf,
An outside chance that Zuccotti may **succeed in** transforming the billionaire
Another subject), and they rarely **succeed in** covering everything. Their
who dare to try to have it all, or **succeed in** a male world. <p> Nowhere is
and Jacques Cousteau types may **succeed in** peeling and slicing onions in a
Expertise and inspiration needed to **succeed in** this very competitive field.
create a positive environment and **succeed in** it," she says in her book: When
The racing industry and, if they do **succeed in** improving racing as an

Table 7: Four meaning groups of the pattern ***V + in + n***

Meaning group	Examples
The **'participate'** group	*join, aid, help, assist, participate*
The **'succeed'** group	*succeed, excel, fail*
The **'lecture'** group	*major, graduate, lecture, train, specialize*
The **'persist'** group	*persist, persevere*

The example that I have demonstrated in the above, is only one way to develop teaching materials by using a learner corpus and a general corpus. Actually, the use of these two corpora can also help learners to correct errors and be aware of the style of discourse. When using concordance lines from a learner corpus and a NS corpus, two points needs consideration (For detailed information on how to use concordances in the classroom, see Tribble & Jones 1990): 1) At the initial stage of using concordancing in the classroom it is advisable for EFL teachers not to give vast quantities of information in that it runs the risks of frustrating the learners at the beginning. 2) When designing activities based on extracts from corpora, EFL teachers can use their own way (which may be familiar to the students) to explain how to work with the extracts from corpora.

Conclusion

Based on a small learner corpus, this study has demonstrated small-scale research on how to use classroom-based learner corpora to complement the teaching materials in EFL writing class with a NS corpus as the reference. With a small learner corpus (the DUT learner corpus) and a large NS corpus (the Bank of English) as the base, the present study is qualitative rather than quantitative in nature.The present study takes learners' own output as the starting point and classifies learners' English in a systematic way with which we usually treat the target language. There are some qualitative generalizations about Chinese intermediate EFL learners, which may be useful information for the group of learner being researched.

Indeed, the aim of classification is not to calculate how many errors in the learner corpora, but to try to draw a detailed picture of learner English. Information about learner English profits both teachers and learners. From the teachers' perspective, they might gain a clear base from which they could complement effective materials to help learners. From the learners' perspective, knowing the features of their own interlanguage is exciting and may help them be aware of their weak points and merits in the target language.

However, using learner corpora itself is not the end of the work. As Seidlhofer (2000:223) says, working with learner corpora will always 'include the consultation of L1 corpora and descriptions based upon them'. According to Tomlinson's (1998:5-22) summary of the basic principles of SLA in materials development, there are 3 advantages of activities based on both learner corpora and L1 corpora. First, the materials are relevant and useful to learners in that they are based on evidence of learner English. Second, the materials facilitate and require learners' self-investment in that concordance lines from learner corpora may stimulate learners' interest and encourage them to make discoveries for themselves. Third, the materials provide language points that learners are ready to acquire.

References

Collins Cobuild English Dictionary for Advanced Learners (3rd edition). 2001. Glasgow: HarpersCollins Publishers.

Fox, G. 1998. Using corpus data in the classroom. In Tomlinson, B. (ed.), 25-43.

Francis, G., Hunston S. and Manning E. 1996. *Collins COBUILD Grammar Patterns 1: Verbs* London: HarperCollins.

Francis, G., Hunston S. and Manning E. 1998. *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives.* London: HarperCollins.

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

Granger, S. and Tribble, C. 1998. Learner corpus data in the foreign language classroom: form-focus instruction and data-driven learning. In Granger, S. (ed.), 199-209.

Hunston, S and Francis, G. 1999. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam: Benjamins.

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: CUP.

Leech, G. 1998. Preface. In Granger, S. (ed.), xiv-xx.

Lewis, M. 1993. *The Lexical Approach: The State of ELT and a way forward.* Hove: Language Teaching Publications.

Scott, M. 1999. (Version 3.0), *WordSmith Tools.* Oxford: OUP.

Seidlhofer, B. 2000. Operationalising intertextuality: using learner corpora for learning. In Burnard, L and McEnery, T. (eds) *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang, 207-223.

Tomlinson, B. (ed.) 1998. *Materials Development in Language Teaching.* Cambridge: CUP.

A Collection of Abstracts

Reframing the Object of Teaching and Learning: the impact of corpus evidence on language teachers

Amy B.M. Tsui
The University of Hong Kong

It is now widely accepted that corpus linguistics has a unique contribution to make in language teaching. By presenting students with corpus evidence, teachers can engage them in a "bottom-up" process of linguistic enquiry and help them to formulate and test their hypotheses about the target language. More language teachers have begun to introduce corpus evidence in their classrooms and more papers have been published in this area. Surprisingly, the impact of corpus evidence on language teachers' own language awareness seems to have received much less attention and far fewer papers have been published in this area.

In this presentation, I shall discuss an analysis of around 2000 grammar questions sent by primary and secondary English teachers in Hong Kong to a website *TeleNex* over a period of nine years. It focuses on areas in which corpus data are found to be extremely useful in addressing teachers' questions. One example is agreement or concord which has the highest number of teacher questions. Agreement is an area that is taught very early on in the ESL and EFL curriculum. ESL and EFL learners are given a simple and straightforward rule that when the subject is singular, the main verb must be singular. The various "deviations" from the rule, and the seemingly conflicting patterns that teachers have found in authentic linguistic materials, however, puzzled them. They have difficulties figuring out any regular patterns. Synonyms is another area in which teachers frequently ask questions. They either have problems differentiating the synonyms themselves or find it difficult to explain to the students the ways in which the so-called synonyms are different, though they feel that these words are not entirely synonymous. This paper identifies the sources of difficulty for teachers and discusses the ways in which these questions can be addressed by corpus evidence. It argues that because a "corpus-driven" approach has challenged many of the long-standing assumptions about language and has brought about a qualitative change in linguistic descriptions, it is essential that language teachers are engaged in interrogating corpus evidence, and in reflecting and reframing their understanding of the target language. This in turn should lead to a reframing of the object of ESL or EFL teaching and learning.

Spoken Corpora – Are Only Native Speakers Interesting?

Anna Mauranen
University of Tampere, Finland

In this presentation I discuss the use of spoken corpora in relation to foreign language use. I suggest that English as lingua franca (ELF) offers an important basis for modelling a target for foreign language learners. ELF is also fascinating as data for more theoretical research, in helping us understand mechanisms of complex language contact, successful foreign language learning, as well as current developments in the English language itself.

Spoken corpora are attracting increasing interest among corpus scholars, despite the still prevalent overrepresentation of writing. Clearly, spoken language is fundamental to language theory, even though most models for linguistic description have been based on written language.

These days speaking is also becoming more and more important in international contacts, which means a growing demand for applying linguistic knowledge to the spoken language. If we look at applied linguistics and the teaching of English, we quickly notice that while spoken language is prioritised in syllabuses and in textbooks for beginning and intermediate levels, it tends to diminish in proportion as we move up the scale of proficiency. More importantly, dialogues in textbooks are usually not based on authentic speech, or analyses of spoken data. This is not surprising, of course, if linguistic models derive from writing in the first place.

Corpus data should be ideally suited for offering a wealth of authentic data for language learning and teaching purposes; it should also inform the compilation and assessment of language testing. Corpora have indeed made their way to English language teaching, but again this is overwhelmingly true of written corpora. It is in practice hard to get hold of spoken corpora, and they are also very laborious and expensive to compile. Those few spoken corpora we do have available may appear as parts of general corpora, and may not be easily separable from the written parts, as is the case with the BNC, for example.

Moreover, the corpora we have currently available are exclusively native-speaker based. The exclusion of the non-native speaker persists despite the fact that English is spoken more widely as an international language, a lingua franca, among people from all over the world than between native speakers.

I want to raise a question in respect of the validity of prioritising the native speaker. Is the native speaker really the best model for a NNS to emulate? This might be defended in the case of more nationally based languages, but English as today's global lingua franca is a special case. Language teaching has a tradition of setting up the native speaker as the ideal model for every learner. The fact that nobody ever reaches this goal, by definition unattainable, has not seemed to bother the profession much. Learner-centred models of language teaching have tended to adopt more lenient attitudes towards learner errors than was customary before the communicative approach. Sadly, however, the "soft" approaches have not presented linguistic principles for distinguishing permissible from non-permissible deviations. The assessment is left to intuition, and the implicit target remains the native. The introduction of corpus data in language teaching has meant that the native speaker's usage can be modelled more realistically than the idealised and intuitively depicted speaker of old, but this has largely meant changes in the written mode only, and has kept the native in his place.

There are of course also learner corpora. The best known of these is the ICLE corpus, which is written, like most others following in its wake. The way learner corpora are used, though, supports the dichotomy of the native speaker vs. the learner. Usually they are used for pointing out where learners go wrong, that is, what they have not yet mastered in a native-like fashion. Now, by advocating English as lingua franca as an alternative model, I am not suggesting that we should turn the tables and take learner corpora as the target. I want to draw a clear distinction between learners and lingua franca users. In other words, to see the familiar term EFL as clearly different from ELF. Despite the similarity of the acronyms, the conceptual backgrounds are far apart. English as a foreign language, EFL, is a pedagogical term, which is concerned with progress in acquisition, its measurement and support. Its perspective is normative and it views learners pedagogically. However, when we move out of the classroom to using English in the real world of business, professional

discourse, etc., it ceases to make sense to regard speakers as learners. Speakers who use English successfully for their needs of international communication are not learners attempting to improve their proficiency, but users in their own right who also inevitably influence and change the norms of the language. English as a lingua franca, ELF, thus refers to a type of English use, and is not a pedagogical term.

ELF is theoretically interesting in that it offers a glimpse into the ways in which the language is currently changing outside the native use, in a unique mixture of source and target language contacts. These changes take place primarily in speech, because writing has almost always undergone many editing cycles before it reaches its readers.

If we put together these facts – the fundamental role of English as the global lingua franca, and the primacy of speech in language change, linguistic theory and practical needs, we can ask why there are no large electronic databases of spoken English as lingua franca. To start filling this gap, I shall describe a specialised corpus, which we are currently compiling at the University of Tampere. I argue that ELF corpus data offers a good basis for modelling what works in successful foreign language communication; it requires us to reconsider our received notions of relevant linguistic descriptions.
I shall illustrate ELF use from the academic speech corpus at Tampere, the ELFA corpus (English as Lingua Franca in the Academia), and compare it to the MICASE corpus (Michigan Corpus of Spoken Academic English). We can see that the patterning in ELF data is highly similar to comparable NS data in terms of discourse structuring: ideas are formulated within turns by similar means of for example repetition and paraphrase, and participants interact to co-construct meanings and propositions. The language consists of sequences which are mixtures of relatively fixed schematic elements and more variable elements, that is, expressions which are partially formulaic while also being productive. I have called these adjustable complex expressions (ACEs). The individual ACEs that speakers use are not identical in NS and ELF data, but the principles appear similar.

Parallel Corpora and Language Teaching

Wolfgang Teubert
University of Birmingham, England

Teaching English as a foreign language has, until very recently, taken for granted that the vocabulary has to be introduced in form of new single words. Embedded in a context the new word is unambiguous. But as a lexical item within the dictionary the word turns out to be ambiguous and/or fuzzy. The information provided by the bilingual dictionary is not sufficient to have a language learner to choose the appropriate translation equivalent.

An analysis of mistakes commonly made by language learners shows that most lexical errors result from a failure to deal with the ambiguities of single words. Before corpus linguistics words were considered to be basic units of meaning. There was a very limited set of idioms belonging to the cultural heritage; and they were taught regardless of the (in-) frequency in which they occurred (eg *it's raining cats and dogs* which is

actually quite rare). Since then corpus research has shown that compounds, multi-word units, collocations and set phrases account for the majority of units of meaning of any text. They haven't been listed in dictionaries because there was no way to detect them, and lexicographers could not detect them because language users do not have developed an 'intuition' for them.

In order to improve foreign language learning we have to introduce the notion of the unit of meaning into language teaching. Language teaching materials will have to contain corpus evidence, to empower the students to recognise units of meaning themselves. Units of meaning are words embedded in a context of other words; within that context they are unambiguous.

Units of meaning are often determined by a syntactic pattern (e.g. Adj+Noun, e.g. closing *remarks*, or Noun+Noun, e.g. *bullet format*, or Verb+Noun, e.g. *invite proposals*) or just by frequency. In translation, these units are translated as a whole. Together with their translation equivalents, they can be detected in parallel corpora of source language texts with their target language translations and re-used for composing a text in (or preparing a translation into) the target language.

To learn a foreign language properly means to know how they differ from one's native language. Parallel corpora represent a much better interface between the source and the target language. Working with a parallel corpus will enable the students to work out for themselves how the language they learn differs from the language they grew up with. It will draw their attention to different grammatical structures, to different vocabulary registers, to different ways to express content far better than any grammar book, any dictionary, any language activator could do it. I will give examples in my presentation how students can successfully use parallel corpora to find target language equivalents that are not provided by bilingual dictionaries, not even if they are used in combination with monolingual dictionaries.

Use of Verbs in Teacher Talk: a study of comparison between local English teachers and native English teachers in Hong Kong primary schools

He An E
Department of English, The Hong Kong Institute of Education

The presentation reports the preliminary findings of a study which attempts to identify features of teacher talk in the classrooms of local English teachers (LET) and native English teachers (NET) in Hong Kong primary schools. Specific attention is focused on the use of lexical verbs in teacher talk. Eighteen LETs and ten NETs participated in the study. By using WordSmith tools, a lexical-verb wordlist was generated from the data collected in the classrooms. Using the framework of Halliday (1994), the first ten most frequently used lexical verbs were then classified in terms of material, mental, and verbal categories. The study indicates a tendency for LETs to use more verbal-type verbs and for NETs to use more mental-type verbs. There is also a tendency for LETs and NETs to use the same verb in a different way in terms of the meaning conveyed, collocations and variations. The difference identified may reveal the different roles LETs and NETs have played in the classroom. It may also be related to the issue of language proficiency of the teachers. The findings have born some pedagogical implications for teacher education courses.

The Web as a Generalized Corpus

Shouxun YANG
Foreign Language Teaching and Research Press
Beijing Foreign Studies University

In the paper we propose to treat the World Wide Web as a generalized corpus, which can be used in a number of ways to complement specifically constructed corpora. Some limitations in specifically constructed corpora are discussed, where we demonstrate the utility of the web as a corpus in a wide range of applications. The most noteworthy limiting factor in traditional corpora is that they are almost fixed once they have been constructed. On the other hand, the web is dynamic, with new resources being added and some old resources gone each day. We can construct various virtual corpora out of the web. The distributed corpora are no more collections of web links, which avoids the headache of licensing. Links can be automatically checked, updated, and, if necessary, deleted. Or we can directly take the web as a huge, heterogeneous corpus and search the web with well-established search engines with advanced search techniques. There are a few related works in this direction, but we have a wide coverage than previous works and emphasize its applications in the context of China.

Keywords: World Wide Web virtual corpus multi-view

A Corpus-Based Analysis of Adverbial Connectors in the Chinese EFL Learners' Written English

Deng Fei
South China Agricultural University, Guangzhou, 510642

Cohesion and coherence are essential features of a well-developed written text. A text, no matter in whatever language, has to be both cohesive and coherent so that the concepts and relationships expressed should be relevant to each other, thus enabling the reader to make plausible inferences about the underlying meaning. In English written discourse, adverbial connectors allow writers to show readers how different parts of the text are interrelated and how they should be interpreted. They aim not only to help readers with text comprehension, but also enable readers to interact with writers and to encounter as little difficulty as possible in capturing writers' intention while reading, and direct writers to the dimension along which a text production goes. If used appropriately, adverbial connectors have a positive effect on the clarity and comprehensibility of discourse. In this thesis, we attempt to undertake a descriptive research on the adverbial connectors in the essays written by the Chinese non-English major college EFL learners in CET-4 to explore the following questions:

1) What are the differences in the use of English adverbial connectors between Chinese EFL learners and the native English writers in terms of general tendency, semantic types, functional categories, clause positions and the top ten most frequently used items?

2) What are the causes that underlie the Chinese non-English major college EFL learners' use of adverbial connectors? Is there any evidence of mother tongue influence?
3) What are the patterned developmental changes in the use of adverbial connectors by Chinese learners? Is there any relationship between the Chinese non-English major EFL learners' use of adverbial connectors and the quality of writing?

In order to answer these research questions, we adopt the methodology of corpus-based analysis and CIA (contrastive interlanguage analysis) and use three corpora (i.e., a learner corpus: cltst3, a subcorpus of the Chinese Learner English Corpus which was constructed by Gui Shichun and Yang Huizhong, a native speaker corpus of English: NSC and a native speaker corpus of Chinese: L1 Chinese) as the sources of data. First, we set up a native speaker corpus of English (NSC) and a native Chinese corpus (L1 Chinese) which are comparable to the cltst3 corpus. Then we determine and identify the search items. Next with the help of search tools, MicroCorncord and Abproject, all the search items are searched and statistical findings were collected for various analytical purposes. Finally, the statistical findings are analyzed and discussed. Quantitative and qualitative studies are conducted and SPSS (10.0) are used to test the relationship between the Chinese learners' use of adverbial connectors and the quality of their English writing.

The major findings of the study are summarized as follows:

1) The Chinese non-English major college EFL learners have displayed a great tendency to overuse adverbial connectors in their essays. Despite the general tendency of overuse, they tend to use less corroborative adverbial connectors which serve interpersonal function and formal contrastive/concessive adverbial connectors like *however, yet*. Chinese learners have displayed a stronger preference for initial position and a weaker preference for medial position in the use of the English adverbial connectors compared with English native writers.
2) The Chinese learners' use of some adverbial connectors might be explained as a result of L1 transfer, the learners' lack of stylistic awareness, the influence of classroom instruction, overgeneralization and the learners' lack of audience awareness.
3) There is significant correlation between the learners' use of adverbial connectors and their writing quality. The Chinese learners with higher linguistic proficiency perform better in the use of adverbial connectors in terms of stylistic awareness, L1 transfer and audience awareness in their writings than those with lower linguistic proficiency.

The study is intended not only to shed light on the Chinese learners' use of adverbial connectors to help raise both teachers' and learners' awareness of the value of adverbial connectors in English learning, but also to establish a corpus-based approach that can be extended to the studies of other linguistic phenomena.

Analysis of the Misuse of Tenses in the English Composition of Chinese College Students

Duan Manfu
Inner Mongolia University

Chinese Students' English has its own characteristics. It is of significance for English teaching and research to analyze and study these characteristics. The paper uses the corpus approach to study the misuse of tenses in Chinese College students' CET-4 compositions of the sub-corpus of "Non-major College Students" in the Chinese Learner English Corpus (CLEC). Nine types of misuse are identified, and analysis is made about the causes of each type. It is found that the misuse of tenses in Chinese college students' English compositions is mainly in the simple present tense and the simple past tense; the wrong collocation of tenses is mostly in the kind of complex sentence with the main clause and subordinate clause both indicating the future action and the subordinate clause using the simple present tense; as for the two special words "could" and "would" with both declarative and subjunctive usages, Chinese students often get confused and misuse them; students have vague understanding of the time concept indicated by the different tenses. Based on the analysis, the author tentatively provide some interpretation for these mistakes: many problems exist in the teaching of grammar in Chinese English teaching, and the teaching of tenses is not systematic and the focus of teaching and tests is not in accordance with the actual uses of tenses; students are affected by interlingual transfer and intralingual transfer when making English sentences. At last, some suggestions are offered by the author for the teaching of tenses: teachers should combine systematic teaching of tenses with the teaching of tenses in actual contexts and students should try to write English compositions using different tenses.

Key words: Chinese learners, English writing, corpus, tenses

Bring A Corpus Within Everybody's Reach

Chuncan Feng
Ningbo Institute of Technology, Zhejiang University

This article gives a detailed introduction of ECBSS, E-Corpus Building & Searching System, in five respects: its background, history, current functions, functions to come and application. The value, importance and necessity of a language teacher, researcher or learner having an English corpus at hand are quite self-evident. There ARE several well-known English corpora abroad, but the few such corpora are not readily available to us for teaching, learning or researching purposes, either because of their limited access or because of their inaccessibility. Their on-line corpus service is only a shop window. Now ECBSS can help build a corpus of our own. It has a built-in powerful search engine that fully supports user-defined search or analysis conditions and a built-in powerful concordancer that gives instant statistics on collocation, frequency and so on, and a detailed sorted report with frequency statistics. Being more than a walking dictionary, ECBSS is helping us gaining expertise insights into the actual use of English.

Key words: Corpus, Corpora, Concordancer, Corpus building

A Corpus-based Study on the Uses of "Drunk" and "Drunken"

Wu Li-ying
Zhejiang Wanli University

In recent years, with the development of technology and the widespread use of the Internet, no linguists, dictionary compilers and language teachers can choose to ignore the boom of corpora, which has been moving from the margins to the center of both language research and language teaching. The goal to the present paper, which comprises of three parts, is just to recommend the use of corpora in language teaching and learning. The first part of the paper points out that synonymy and collocation tend to be among the most common obstacles in language study, and language learners usually turn to dictionaries for help when encountering such problems. But the findings from three authoritative dictionaries are somewhat confusing and discouraging concerning the uses of "drunk" and "drunken". The second part of the paper solves the problem by conducting a corpus-based study. And the last part of the paper mainly discusses the prospects of applying corpora to language teaching and learning.

Key words: corpus, dictionary, foreign language teaching and learning

A Corpus-based Analysis of the Common Use of Make

Hu Haizhu
Henan Normal University

Based on the corpus BROWN, several senses of the verb *make* have been identified in the paper. The concordance result shows that: 1) The 'core' meaning of the word is the most frequently used sense; 2) The frequency of the second sense comes closely next to that of the first one; 3) All the collocations of *make* as idioms are upward. The first common sense of *make* is found to be used in quite similar and simple patterns. The basic phrase pattern used is VO. The collocates as objects are variable and unexpected. Yet there are still some typical collocates as objects and some significant determiners of these objects. Concordance result of *made* in this sense shows that passive voice is used quite often together with a prepositional phrase (59%). The second sense of *make "to (cause to) become (to do) or appear"* is used in different patterns. All of them can be summarized as the phrase pattern of VO(C). The result shows that personal pronouns take a large percentage in the significant collocates in object position, reflecting a sense of personal relativity in the use of the verb in this sense. But the first person pronouns take little part, suggesting the sense of speaker irrelativity. And most of the significant collocates as complements have more or less sensory meaning, especially verbs. There are similarities for the use of *make* in these two common senses. The study of the most common use of *make* will greatly help our EFL teaching and learning. But there are some limitations in our research work. The corpus BROWN is comparatively small in size and old in language. And the study is to be furthered to make the result more systematic and scientific.

Key words: corpus-based, senses, patterns, collocates

CORPUS AND FOREIGN LANGUAGE TEACHER

Huang Xiaoying
Xi'an University of Electronics

Corpus linguistics is one of the hot issues in abroad linguistics and language teaching fields in recent years, giving objective description to language on the basis of corpus, and studying language performance not language competence.

Because the share of corpus is difficult, only a small community grasps most of the corpus, and the majority outsider only hear of those but not see its form. Many foreign language teachers do think corpus is for the researchers to touch and it is far from us. Aiming at such kind of idea, this paper gives a comparison between Collins cobuild and BNC(British National Corpus), and makes a conclusion that Collins cobuild will be the most suitable one for English teacher. For Chinese English teachers, cobuild is easy to access through google web or yahoo website. Generally English teachers should grasp second foreign language, while cobuild provides French Spanish and other languages corpus to improve teachers' second foreign language ability. And this paper describes how to take use of cobuild.

Key words: corpus linguistics, Cobuild text recourse share

A Corpus-based Study of the Difference of *As To* and *As For*: Application of Corpus in the Distinguishing of Synonymous Expressions in English

Kong Guang
Foreign Language University of PLA, Luoyang, Henan, 471003

Distinguishing of synonymous expressions has always been the weakness of language teaching which is difficult to solve. Studies of cognitive psychology have shown that to second language learners, synonymous expressions corresponding to one concept in Chinese are the most difficult to learn. *As for* and *As to* are hard to be distinguished in meaning and in use when used as prepositions. This paper intends to explore the different usages of the two phrases on the basis of corpus, trying to reveal some limitations of the classical grammatical descriptions. Further, Chinese learners' actual use of the two phrases is examined on the basis of CLEC in the expectation of giving some suggestions for the use of corpus in distinguishing synonymous expressions in second language teaching.

Key words: corpora, foreign language teaching, distinguishing of synonymous expressions

On the Use of *and* and *but* in Chinese Postgraduates' Academic Writing[1]

Liu Guobing
Henan Normal University

Abstract:
Cohesion is an important issue of text linguistics and discourse analysis. One of the ways to make our writing semantic coherent and consistent is to use conjunctive elements. Here is a contrastive study on the use of *and* and *but* in linguistic papers written by Chinese postgraduates and native speakers of English. We found: (1) Chinese postgraduates tend to use more conjuncts than the native writers, especially two sentence-initial conjuncts *and* and *but* are significantly overused. (2) Compared with native writers, Chinese postgraduates are inclined to use more *and* than *but*. (3) Chinese postgraduates not only overuse *and* and *but*, but also misuse them. To some extent, this phenomenon shows that Chinese postgraduates lack a full understanding of style appropriateness and semantic properties of *and* and *but*.

Key words: *and, but*, Chinese postgraduates, overuse

Chinese students' Acquisition of English Subordinate Clauses — A Corpus-Based Study

Lin Dehua
PLA Foreign Languages University, LuoYang

The paper is intended to study Chinese students' acquisition of English subordinate clauses in CLEC(Chinese Learner Corpus). The research is conducted from the perspective of corpus linguistics. The concordance software used in the research is ConcApp6.0. The findings of the research are: ①Chinese students tend to make three categories of errors: Errors of connectives (such as conjunctions) introducing clauses, errors in clauses, and errors in terms of the relation between a subordinate clause and a main clause. The errors of the first two categories are more frequent. All these errors are probably due to negative transfer of Chinese, transfer within English and the inadequate acquisition of laws concerning subordinate clauses.②The errors, to some extent, are regular in distribution. Students in different grades make both similar and different errors. All these are due to non-integrity, poor system and instability of interlanguage, and the teaching of English in China.

Key words: learners, English subordinate clause, negative Chinese transfer English transfer

---

*I acknowledge my heart-felt gratitude to Dr. Li Whengzhong who gave me insightful instructions and warm encouragement. He also stimulated and guided my thinking over the period of writing this paper.

A Comparative Study of the Collocates of Health

Lou Baocui
Henan Normal University

To expand one's word knowledge is always a focus of language learners and teachers. The development of vocabulary is a matter not only of getting acquainted with new words but also of deepening one's knowledge of familiar words. The depth of word knowledge is a much neglected area of vocabulary teaching and research. By applying the corpus approach, this paper takes *health* as an example to examine the patterns of Chinese college learners' usage of *health*. It is concluded that: (1) Chinese learners overuse modifiers before *health*, including adjectives, possessive forms and forms of noun+'s. (2) While using *health*, Chinese learners underuse the pattern of its collocation with nouns. (3) There are some misuse concerning the collocation of *health*, e.g. have / want / need / keep (a) good health, body health etc. (4) In order to expand their word knowledge, language learners tend to focus on the vocabulary size rather than on the depth of the word knowledge. Both language teachers and learners should realize the importance of usage patterns and collocations of the seemingly acquired words as well as the vocabulary size.

Key words: corpus, collocation, overuse, underuse

A Corpus-Based Study on Acquisition of English Conjunctions by Chinese Learners

Su Bing
PLA Foreign Languages University

Based on Halliday's cohesion theory, this paper analyses the usage of English conjunctions in CLEC and Brown corpus and has a further investigation on Chinese ESL learners of different levels. The findings of this study show that:

1) The results of the Chi-square show that there is significant difference between Chinese ESL learners and English native speakers. The Chinese ESL learners are intended to use more conjunctions in their compositions than the native speakers do.
2) Compared with the native speakers, the Chinese ESL learners over-use the conjunctions indicating the 'spatial' and 'causal' relations, but under-use the conjunctions indicating the 'addition', 'manner' and 'topic' relations.
3) Comparing the students of different levels, we find that the senior high school students use the fewest conjunctions for text cohesion. The college non-English majors use the conjunctions most frequently in the three groups. However, the English-majors use the conjunctions less frequently than the non-English majors.

Key Words: Chinese ESL learners, native speakers, cohesion, conjunctions

A Corpus-based Study of Negotiated Interactions in the EFL Classrooms of Chinese Middle Schools

谭伟民

广东外语艺术职业学院外语系

Negotiated interaction refers to the modification and restructuring of interaction when miscommunication or breakdown occurs in a conversation between second language learners and their interlocutors. Negotiated interaction is believed to play an important role in second language learning because it can bring about those conditions claimed to be beneficial for language acquisition. Negotiated interaction is of great significance in EFL classrooms, the reason being that it is not only a strategy to save communication breakdowns in classroom interactions, but also a strategy to generate comprehensible input for learners, bring out comprehensible output from learners and draw learners' attention to language forms. In this thesis, the author attempts to undertake a descriptive research of the negotiated interactions in the EFL classrooms of Chinese middle schools through a corpus-based approach.

The present study is a corpus-based analysis of the negotiated interactions in forty EFL classes in Chinese middle schools, which were chosen from an existing corpus of EFL classroom interactions compiled by the School of Foreign Studies, South China Normal University. First, a pilot study was carried out to discover the search words for the negotiated interactions in the research data, set up a set of taggers for their retrieval and prepare an analytical framework for the analysis of them. Then with the help of search tools, all the negotiated interactions in the research data were searched and tagged according to four categorization models. Next statistical findings were collected for various analytical purposes, including the general description of the negotiated interactions in the research data and comparative studies of the negotiated interactions in different types of English classes. Finally, the statistical findings were analyzed and discussed. The main findings of the present study include: (1) Although there is a general lack of negotiated interactions in the EFL classrooms of Chinese middle schools, it was found that the number of negotiated interactions in fine-quality classes is three times as that in ordinary classes. (2) 94% of the negotiated interactions in the research data are initiated by teachers. (3) 67% of all the negotiated interactions are about meaning communication and only 37% are about language forms. (4) 55% of all the negotiated interactions are of one-layer structure and 45% of them contain recursive multi-layer structures. (5) Among the modification devices used in negotiated interactions, clarification request occurs with the highest frequency and comprehension check with the lowest frequency.

Based on the findings from the research data, the present thesis also discusses what factors are conducive to promoting negotiated interactions in EFL classrooms and why there is an overall lack of such interactions in the EFL classrooms of Chinese middle schools.

This study is intended not only to shed light on the negotiated interactions in EFL classrooms of Chinese middle schools to help raise both teachers' and learners' awareness of the value of negotiated interaction in English learning, but also to establish a corpus-based approach that can be extended to further study of negotiated interactions in the corpora of EFL classroom interactions of other settings.

Keywords: negotiated interaction, corpus, modification device, data retrieval

"However": Do Chinese Students Really Know How to Use This Word? -- A Learner-Corpus Based Study

Wang, Jianxin
Beijing University of Posts and Telecom.

Listed by the Collins COBUILD English Dictionary as one of the 700 most frequently used words in English, "however " is an important word for the Chinese students to fully master in their English writing. Interestingly, when this word is used to express concession or contrast and translated into Chinese, its Chinese equivalent is generally put at the beginning of the translated sentence, whatever its original position in English. Based on this observation, we made such a hypothesis: The Chinese students may have problems using "however" in their English writing, due to the influence and interference of their mother tongue. Specifically, they may overuse it at the beginning of the sentence.

This study aims to check whether this hypothesis can be confirmed by analyzing how this word is used in the CLEC, a one million token written English Corpus produced by Chinese students. In comparison, the Brown and the LOB are used as the control corpora. The findings of this study, some of which are surprising and pedagogically useful, have more than confirmed our hypothesis.

Key words: "however" study, CLEC-based word study, learner corpus based study

On Corpus-based Language Teaching

Yan Canxun
PLA University of Foreign Languages

This paper tries to explain why and how to use corpora to assist language teaching in college. Corpora tend to promote a learner-centered approach to language teaching, and substantial data encourage a discovery learning. A computer, some corpora and some corpus analysis tools will enable a teacher to start a corpus-based language teaching. Three basic approaches of corpus analysis are introduced in this paper, which are frequency analysis, concordance analysis and keyword analysis. The paper also gives some suggestions on corpus-based teaching planning and corpus-based class activities. As it is only an introduction to corpus-based language teaching, it mainly focuses on the basic knowledge and skills.

Key Words: corpus-based, language teaching, concordance, class activities

A corpus-based study of the determiner collocation in English: the case of the "the central determiner + the postdeterminer" collocation

WANG Pu
Jiaozuo Institute of Technology

Remarkable achievements have been attained in the study of the English determiner, but most studies focus either on the generalization of the determiner theory or on the behavior of a single determiner and pay comparatively less attention to the actual collocation between determiners. This paper, through corpus investigation, reports the findings of a study of determiner collocation in fictions. It focuses on the behavioral patterns of the "central determiners + postdeterminers" collocation. On the basis of the data obtained, some determiners are re-classified. It is argued that it is better to regard "every", "each" and "either" that are generally believed to be central determiners as postdeterminers. A comparison of the above collocations in American English and British English leads to the conclusion that there is almost no differences between them. Taking "ten occurrences per million words" as a cut-off point, high frequency combinations are found to be: "a / the / this / my / his + last"; "the / that + other"; "the / that / these / those / my / his / her+ cardinal numerals"; "a / the / that / my / his / her / our / their / 's + ordinal numerals"; "the + few / many / next". Finally, postdeterminers that occur most frequently in collocation with central determiners are found to be "ordinal numerals", "cardinal numerals", "last", "other" and "next", the total frequencies of which are at least 405 times per million words.

Key Words: English, determiner, collocation, corpus

A Further Analysis on English Articles in LGSWE

YUE Fu-xin
Tianjin University of Commerce, Tianjin 300134,China

English articles, being minimum in English vocabulary are the most difficult to define their grammatical usage. Traditional grammar depicts roughly the outline of their usage. LGSWE based on a large corpus provides us new findings and fills the gap of the zero article usage in their distribution and frequency analysis. So LGSWE scientifically surpasses the traditional grammar.

Key words: zero article, corpus, distribution frequency

A Contrastive Approach to the Verb Errors in CLEC

Yang Dafu
Xi'an University of Foreign Studies

The Chinese Learner English Corpus (CLEC) consists of compositions and diaries written by high school students (ST2), non-English majors (ST3, ST4), English majors (ST5, ST6). Such a corpus is beneficial for comparing the features of written production by the Chinese learners at different levels and from different specialties. In CLEC, there are 72,532 tagged errors, among which 10,899 are verb errors (accounting for 15.03%). This ratio is the highest in the 7 types of errors tagged by parts of speech, and is twice as high as the number of noun errors (6,047 occurrences, accounting for 8.34%). This shows clearly that the command of verb usage is one of the most difficult aspects of English learning. It is therefore necessary to observe the general conditions under which such errors are made.

1) As far as the error frequencies at different levels are concerned, the general tendency is that when the English proficiency level rises, the frequency of error occurrences comes down (except ST2). What's more, there is no significant difference between the error occurrences of high school students (ST2, 17,748) and those of non-English majors (ST3: 18,869 occurrences; ST4, 16,873 occurrences) while the difference between non-English majors and English majors is significant (ST5: 10,616 occurrences; ST6: 8,426 occurrences).
2) As far as the frequency of verb errors is concerned, the highest is found in the compositions by the high school students (ST2: 3,222 occurrences), then comes the frequencies in the compositions by non-English majors (ST3: 2,347 occurrences; ST4: 2,734 occurrences). Verb errors made by English majors (ST5: 1,365 occurrences; ST6: 1,231 occurrences) are much lower than those made by non-English majors. Verb errors made by these 3 groups of students also follow the tendency of "higher proficiency, lower error occurrences".

The present paper analyzes verb errors made by the Chinese learners of English, and the analysis is based on CLEC. The point of departure is the problems in English learning viewed from the perspective of dissimilarities in the form of English and Chinese verbs. As far as this class of words is concerned, inflection is common in English but absent in Chinese. In English, there are also the finite and non-finite uses of verbs, which prove to be an obstacle in the systematic command by the Chinese learners because there is no variation of this kind in Chinese. About 1/3 of the tense errors in the Chinese Learner English Corpus (CLEC) are caused by the use of primary verb forms in the places where inflection or other non-finite forms are required. Besides, there are a certain amount of errors caused by learners' confusion of finite and non-finite forms. All this indicates that the use of a correct verb form is a noticeable difficulty for the Chinese learners.

The analysis of the verb errors in CLEC is based on the differences in the use of verbs in English and Chinese. English and Chinese verbs with similar meanings may share similar features in transitivity and colligation, but they can also be quite different in use when the context of verb use in one language is different from that in the other. Therefore, word for word interpretation of certain usage of Chinese equivalents in the process of English production can be a likely cause of errors in the use of English verbs. In the same way, the present paper analyzes the causes of errors in subject-predicate agreement and improper ellipsis, pointing out that the Chinese learners' knowledge of their mother tongue exerts its influence on the occurrence of such errors.

It is common to analyze the learner errors in their production of written English through the contrast of English and Chinese and then to explain them from the perspective of negative transfer. Such an approach is limited because causes of the learner errors vary from one learner to another. However, the findings of the present study reveal one important tendency in the commitment of learner errors, namely mother tongue influence, and shed some light on the improvement of learner competence in expressing themselves efficiently and accurately.

Collocation with Chinese Characteristics

Zhang Bin
PLA University of Foreign Languages, Luoyang, Henan 471003

Collocation is a hot research topic in corpus linguistics. Chinese learners' English collocation has its own characteristics and to know these characteristics is of vital importance to English language teaching and research. Through our corpus-based analysis, this paper tries to explore the Chinese learners' collocation from a new perspective and puts forward a new issue, that's collocation with Chinese characteristics.

We find that some collocations are not typical for native speakers, but in contrast, it has high frequency rate in CLEC corpus. Some researchers may conclude that Chinese learners overuse them due to the influence from their mother tongue. However, learners use a variety of such collocations which are different from inappropriate collocations, but at the same time show dissimilarities with native speakers' usage. We will name them **collocations with Chinese characteristics**. They are correct collocations bearing the influence of Chinese so they are typical collocations for Chinese learners.

In all, we should accept collocations with Chinese characteristics as a necessary part of learners' English. It is a special part existing in learners' interlanguage and is worth our studying.

Key words: corpus, collocation, Chinese characteristics

A Corpus-Based Testingon the Validity of Adverb-Placement Rules by TG grammar

Zhou Shijie Tan Wancheng
Dalian Maritime University, Liaoning, PRC

It is a commonplace that the availability of computer corpora permits a quick, easy and effective way of gathering data for both synchronic and diachronic research on a language. Based on the British National

Corpus, which provides the main source of the data, this paper aims at testing whether or not the placement of adverbs in everyday English, either spoken or written, complies with the rule(s) proposed by the Transformational-Generative (TG) grammarians.

As TG grammarians point out, syntactically, there are two different classes of adverbs: they are the sentence-adverbs such as ***certainly*** and VP-adverbs such as ***completely***.

This study is designed to extract a random sample of occurrences by SARA98 from BNC for ***certainly*** and ***completely*** respectively, convert the concordance lists into KWIC displays, annotate each sentence with the position information, and count the frequencies of each of the adverb positions.

Based on the frequency counts of the adverbs, this study comes to the conclusion that both S-adverbs and VP-adverbs can be found in the positions that the TG grammarians predict, and that the distributions of S-adverbs and VP-adverbs in sentences, according to the calculated Chi-Square value (86.006) versus the critical Chi-square value (11.07) at the 5 percent level and 5 degrees of freedom, show a significance of the differences.

Key Words: TG Grammar, adverbs, frequency, Chi-Square test

A Study of Discourse Features in L2 Oral Production: Corpus-based Discourse Analysis of Chinese Tertiary-level EFL Learners' Spoken Narratives

YU Hongliang
School of Foreign Studies, Nanjing University

Oral narration is said to be the basic linguistic requirement of human beings. Narrative Language is "an account of experience or events that are temporally sequenced and convey some meaning" (Engel, 1995). It can be employed to recapitulate past experiences; it can be the way people communicate memories; it can be embedded in conversation; and it can also be monologues.

This research is a study of discourse features in English majors' spoken narratives in the testing contest, based on the oral data in *The Spoken and Written English Corpus of Chinese EFL Learners* built up by Nanjing University. The oral part of the corpus includes the transcribed texts and audio files of the National Oral Test for English Majors (Band 4) (from 1999 to 2002), which consists of three tasks: 1) story retelling; 2) talk on a given topic; 3) dialogue. This study is directed towards Task 2, which is a monologue or a personal narrative in nature. The topic of the task is basically a narrative on a person or an event, in which speaker is prompted to tell about an actual experience.

This study attempts to answer the following questions: 1) What is the general pattern of discourse features, at the macro- and micro-level, of oral narratives in testing context? 2) How do patterns of discourse features vary across the different topics of the testing task? 3) How do patterns of discourse features vary with the test candidates' scores? 4) What can be said about these observed changes and differences?

Narrative analysis, in usual sense, is analysis of a chronologically told story, with a focus on how elements are sequenced and why some elements are evaluated differently from others. In this study, as the research questions indicate, there are two parts: 1) macrostructure analysis, which seeks to find out the general model of structural characteristics of features of learners' oral narrative discourse, or, more specifically, the content of a story (i.e. what a story is about) and the form used to tell a story (i.e. how a story is told); 2) micro-level features of narrative discourse, which include lexical repetition (as a function of topic or as an index of the semantic structure) and discourse markers.

This research is a corpus-based qualitative study by nature. The subjects are 60 test participants of the National Oral Test for English Majors (Band 4) in 2000 and 2001, of whom 30 are higher achievers and 30 are lower achievers. Both higher and lower achievers were chosen from 6 packages of tapes in the corpus. Their oral productions were all recorded while they were taking the test. In the corpus, these recordings were converted into audio files and were transcribed with a certain amount of tagging. So in this study the data consist of two components: audio files and their transcriptions. The instrument is the same task of different topics: 1) Talk on the most unforgettable birthday party you have ever had; 2) Talk on an unusual teacher you have met.

Possible contributions from this study may include: 1) This study has described the macrostructure of Chinese EFL learners' spoken narratives and the linguistic features at the discourse level; 2) This study is methodologically significant in the sense that the description of spoken discourse features is based on the EFL learners' spoken corpus; 3) This study is theoretically important in steering the research of oral English testing; and 4) This study is instructionally significant in the second language teaching and learning.

Parallel Corpora of Chinese and English and its Usage

Kefei Wang Saihong Li
Beijing Foreign Studies University

The ongoing 500 million PCCE(Parallel Corpus of Chinese and English), which is directed by Prof. Wang Kefei, consists of original texts and their translations (Chinese original to English translation and English originai to Chinese translation). It is going to be the biggest parallel corpus in China up to now. This project is co-conducted by the National Research Center of Foreign Language Education of Beijing Foreign Studies University and Beijing University. In this paper, I will introduce its construction, its sentence and paragraph alignment, the concordance software and its usage.

Key words: parallel corpus, alignment, concordance software, usage

Introduction

Corpus-based research grounds its theorizing in empirical observation rather than in appeals to linguistic intuition or expert knowledge. That is to say, corpus provides us empirical database for linguistics studies

(Graeme Kennedy, 1998). Parallel corpus, which contains the original language and its translation or vice verse, reveals what is general and what is language specific and is therefore important both for the understanding of language in general and for the study of the individual languages compared (Stig Johansson,1999). Since the first corpus BROWN & LOB were built in the 60' s, corpus construction has sprung out dramatically, esp. in the 80's. Large corpus, like COBUILD, BNC, ICE set a new standard in corpus design and compilation. However, parallel corpus, due to its technical and other reasons, appeared only in the 90th, the English Norwegian parallel corpus, Europarl parallel corpus, the English-Swedish parallel corpus, British Multiple parallel corpus, Portuguese-English parallel corpus are such cases in point. Britain, Norway and Sweden became the leader countries in this area. Parallel corpus and corpus-based linguistic studies in China is still in its infancy. Parallel corpus of Chinese and English and Chinese and Japanese, which is conducted by Prof. Xu Yiping in the Japanese Institute of BFSU(Parellel corpus of Chinese and Japanese is also one part of the corpora under the name of the project) is to match the parallel corpora in the world.

PCCE (The Parallel Corpus of Chinese and English) is intended as a general research tool, available beyond the present project for applied and theoretical linguistic research. PCCE, which is directed by Prof. Kefei Wang, co-conducted by National Research Center of Foreign Language Education of BFSU and Beijing University. The objective of the corpus is to create 500 million words, representative of modern Chinese and English in the 20th century as to establish the research platform for comparative studies of Chinese and English that can meet observational and descriptive adequacy, translation studies, second language teaching, statistical analysis such as frequency of occurrence, machine translation and compilation of bilingual dictionaries. (Wang Lidi & Wang Jianxin, 2000) Now there are about 300 million words has been successful aligned by paragraphs and sentences, 170 million words were tagged and parsed and thus, can be searched by using the concordance software.

Study and Design of Web Corpora Mining System

Zhang Xiaojun[1] *Zhang Linglan*[1] *Yang Daliang*[2] *Liu Jun*[2]

(Zhang Xiaojun Shanghai University of Electrical Power;

Zhang Linglan Shanghai Synruns Electric Co. Ltd)

Corpus plays more and more important role in the modern language studies. It is worth studying on corpora mining and corpus building for the linguists. The mining and collecting of corpora is possible for any linguist to build his own researching corpus for the rapid development of computer and Internet. This paper points out a web corpora mining system which uses XML techniques in web mining to set a personal web corpora mining. A system named LawsMiner focusing on laws corpora is applied.

Key words: web, corpora mining, LawsMiner, XML, corpus

A Corpus-Based Study of Discourse Marker Use in the Chinese EFL Learners' Spoken English

Lifei Wang
Nanjing University, China

Abstract: This paper reports a study on the use of discourse markers in oral English performance based on the spoken English corpus of the Chinese learners (SECCL). The native English corpus for comparison is the spoken component of the British National Corpus (BNC). The results of the study yielded four important findings. First, Chinese EFL learners and English speakers rely on different discourse markers in their respective speech production. Second, it was found that Chinese EFL learners tend to under-use discourse markers if compared with the native English speakers, which conforms to our impression of Chinese EFL learners' oral performance. Third, Chinese EFL learners tend to overuse only very few additive and emphatic discourse markers or fillers such as " and, but, very, I think." This is undoubtedly a reflection of their interlanguage development and L1 influence. Fourth, the positions of discourse markers in speech are identical for Chinese learners and native speakers, indicating little difficulty of the former in conforming to the target norm and little evidence of mother-tongue influence.

Key Words: corpus linguistics, discourse marker, spoken English, ELT

A Comparative Study on the Use of BIG, LARGE, GREAT in Native and Non-native Student Writing

Hu Chunyu

Guangdong University of Foreign Studies

Abstract: Correct use of synonyms belongs to one of the trickiest fields of English. This paper investigates EFL learner use of a group of seemingly synonymous words: *BIG*, *LARGE*, and *GREAT*. The major questions addressed are: do learners tend to over- or underuse these words? If so, what are the underlying reasons? To answer these questions, authentic learner data has been compared with native-speak data using computerized corpora and linguistic software tools to speed up the initial stage of the linguistic analysis. Results show that EFL learners, even at an advanced proficiency level, have great difficulty with a group of synonyms such as *BIG*, *LARGE*, and *GREAT*. They tended to universally overuse *BIG* and *GREAT*. Qualitative analysis demonstrated that some kinds of misuse might be L1-related while the major part of learner errors derived from their confused ideas of the three adjectives. In the conclusion, the pedagogical implications of the study are discussed and suggestions made for using concordance-based exercises as a way of raising learners' awareness of the complexity of common words.

Key words: synonym, ICLE, concordancing

An Investigation into the Washback Effect of the TEM4 Writing Item On the Second Year English Majors' Written English Competence

Hu Yuying

Hubei Normal University

Abstract: The present research investigates the washback effect of the coaching for the TEM4 writing item on the second year English majors' written English competence. The theoretical framework of the research is based on Bachman & Palmer's theory (1996) on the definition and operating mechanism of the washback effect of the use of test. Three approaches are employed in conducting the present research, such as questionnaires, statistical analysis, and corpus-based investigation. The main purpose of the research is to examine whether the coaching for the TEM4 writing item will have positive impact on the second year English majors' written English competence on 3 aspects: vocabulary, topic sentence and cohesive devices. Findings indicate that the washback effect works efficiently to bring about the students' improvement in their written English competence to some extent. They are as follows: 1) The frequency of words at higher bands (band 2-4) in data II is higher than that in data I; there are fewer spelling errors in data II than those in data I; the type/token ratios between data II and data I are significantly different. 2) The students tend to use more topic sentences in data II than they do in data I. 3) There are more cohesive anaphoric forms and conjunctions in data II than those in data I. The present research is a novel undertaking of using computer and corpus tools to investigate and study language phenomena in EFL teaching in China based on the authentic language data. It is sincerely hoped that the research methodology and the research findings will cast light on other EFL teaching research in China.

Key words: washback effect, written language competence, corpus investigation

Now the machine analysis of text is possible.

——J. R. Firth

The ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before.

——J. M. Sinclair

[Computerized corpus linguistics] ... defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject. The computer, as a uniquely powerful technological tool, has made this new kind of linguistics possible. So technology here (as for centuries in natural science) has taken a more important role than that of supporting and facilitating research: I see it as an essential means to a new kind of knowledge, and as an 'open sesame' to a new way of thinking about language.

——G. Leech

2003 上海语料库语言学国际会议述评①

李文中[1], 濮建忠[2], 卫乃兴[3]

(1. 河南师范大学外国语学院, 河南 新乡 453002; 2. 解放军外国语学院三系, 河南 洛阳 471003;
3. 上海交通大学外国语学院, 上海 200030)

摘 要: 2003 上海语料库语言学国际会议的议题包括: (1) 基于 CLEC 的中国英语学习者中介语分析; (2) 基于平行语料库的语言研究; (3) 基于 COLSEC 的中介语分析; (4) 英语语言与教学研究; (5) 英语变体研究; (6) 语料库技术研究。笔者认为, 中国的语料库语言学研究从一开始就与外语教学密切结合, 各校广泛合作、共享资源, 目前已取得长足的进步, 杨惠中教授等的工作在国际上具有很大的影响。语料库语言学作为一个学科, 今后需进一步构建理论, 注重语料库深度加工。中国的语料库语言学则应加强语料库技术开发, 更系统、全面地开展对中介语的描述和研究, 以期对外语教学做出更大的贡献。

关键词: 语料库语言学; 中国; 外语教学; 对比中介语分析
中图分类号: H0 文献标识码: A 文章编号: 1002-722X (2004) 01-0056-04

The 2003 International Conference on Corpus Linguistics at Shanghai

LI Wen-zhong[1], PU Jian-zhong[2], WEI Nai-xing[3]

(1. Faculty of Foreign Languages, Henan Normal University, Xinxiang, Henan Prov., 453002, China;
2. Department Three, PLA University of Foreign Languages, Luoyang, Henan Prov., 471003, China;
3. School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, 200030, China)

Abstract: The 2003 International Conference on Corpus Linguistics at Shanghai covered six areas in corpus linguistics: (1) CLEG-based analysis of Chinese learner English as interlanguage; (2) language research based on parallel corpora; (3) COLSEG-based interlanguage analysis; (4) studies on the English language and English teaching and learning; (5) corpus-based studies on English varieties; and (6) research on corpus technology. In the view of these authors, corpus linguistics research in China has made remarkable progress. The works of Professor Yang Huizhong and some other scholars have had worldwide influence. The research in China has had its focus on applied studies in foreign language teaching and learning, and has been pushed ahead through coordinated efforts and pooled resources of different universities and colleges. As an academic discipline, corpus linguistics still needs to be elaborated in theory and practice. Studies in China should lay greater emphasis on technology development, and on comprehensive and systematic analysis of the totality of interlanguage.

Key words: corpus linguistics; China; foreign language teaching and learning; Contrastive Interlanguage Analysis

1. 概述

2003 上海语料库语言学国际会议经过近两年的酝酿和筹备, 于 2003 年 10 月 25 日至 27 日在上海交通大学举行。本次会议由杨惠中教授为主要发起人, 以上海交通大学、广东外语外贸大学、上海外语教育出版社为主办单位, 出席会议正式代表 80 余人, 分别来自英国、意大利、芬兰、新加坡、日本, 以及中国大陆与香港地区各高校, 充分体现了国际性。会议邀请了国际著名语料库语言学家 John Sinclair 教授② (意大利 Tuscan Word Centre)、Wolfgang Tubert 教授 (英国伯明翰大学)、桂诗春教授 (广东外语外贸大学)、Anna Mauranen 教授

收稿日期: 2003-11-24

作者简介: 1. 李文中 (1963-), 男, 河南开封人, 河南师范大学外国语学院教授, 博士, 主要研究方向为语料库语言学和应用语言学; 2. 濮建忠 (1968-), 男, 浙江德清人, 解放军外国语学院副教授, 博士, 主要研究方向为语料库语言学和应用语言学; 3. 卫乃兴 (1957-), 男, 河南济源人, 上海交通大学外国语学院教授, 博士, 主要研究方向为语料库语言学、语言学及应用语言学。

(芬兰 Tempere 大学) 以及 Amy B. W. Tsui (徐碧美) 教授 (香港大学) 做主旨演讲，杨惠中教授致开幕词。在会上，先后有 38 人宣读了论文。会议闭幕前还安排了由 John Sinclair 和 Anna Mauranen 主持的讲习场。

本次会议的主要特点是：(1) 规模大，档次高。本次会议既有国际国内语料库语言学界的先驱参加，又吸引了一大批国内语料库语言学研究的中坚。与会代表大多是近十几年在该领域辛勤耕耘的中青年专家和学术带头人，其中博士和硕士生导师约占 1/3。80% 的代表具有硕士和博士学位。一部分代表为在读博士和硕士生。(2) 学术性强，与国际上的研究保持同步水平。在本次会议上，国际和国内语料库语言学学者在同一个平台上，进行双向、平等的交流。(3) 成果丰硕，特色突出。从会议交流的论文可以看出，我国的语料库语言学研究已超越了对西方理论的引介和评述阶段，国内学者的论文都针对中国外语教学实践或是基于自主开发的各类语料库的研究成果。

本次会议表明，我国的语料库语言学研究的队伍初显规模，潜力巨大。会议宣读的论文大多出自具有博士学位的中青年学者以及在读博士和硕士研究生之手，他们的研究显示了严谨的科学态度和研究规范，所取得的成果令人振奋。中国语料库语言学研究起点之高，发展速度之快，学术队伍之齐整，给与会的国外专家留下了深刻的印象。语料库语言学研究先驱 John Sinclair 对此感到兴奋不已。杨惠中教授在会议总结中指出，本次会议“总结和展示了中国语料库语言学近几十年的发展成果，不仅是一次成功的会议，还将成为我国语料库语言学研究史上的一次重大事件，并成为该领域研究的新起点”。

2. 研讨主题及贡献

本次会议的研讨主题可分为以下 6 个方面：

(1) 基于中国学习者语料库 (CLEC) 的中介语研究，占 49%。此类研究大多采用 CIA (Contrastive Interlanguage Analysis) 方法，基于学习者语料库与英语本族语语料库进行对比分析，目的在于描述中国英语学习者在语法、词汇、搭配等运用中的中介语特征，从而揭示对英语教学的意义。桂诗春对“中国英语学习者错误分析的认知模型”的研究最令人瞩目，他通过建立学习者语言运用生成机制的模型，认为学习者在“词语感知层”、“词语语法层”及“句法层”通过“词语实现” (lexicalization)、“句法实现” (syntacticalization) 和“词语再实现” (relexicalization) 生成序列产生语言输出，并试图解释学习者在以上 3 个层面上的错误成因及交互影响。该研究的意义在于为学习者语言的个案分析提供了深层的认知模型和整体理论框架。此外，文秋芳、丁言仁对专业英语学习者频度副词的分析和研究、严辰松对学习者连接词语运用的调查和分析以及其他类似研究，对我们深入系统了解学习者的典型困难，以及针对这些困难开展补偿式教学和辅导提供了可靠的依据。

(2) 基于平行语料库的语言研究，占 8.2%。平行语料库研究是近年来语料库语言学横向发展的新趋势。Wolfgang Tubert 提出，利用平行语料库进行母语和目的语对比，通过提供语境双语翻译促使学生学习文本中的“意义单位” (units of meaning)，可提高学生词汇学习的效率。值得注意的是，不少研究都同时关注词语组合或意义模块化对语言学习的重要意义，如 John Sinclair 和 Anna Mauranen 的“切块” (chunking)，濮建忠的“词块” (chunks)，李文中的“词丛”或“词簇” (word clusters)，卫乃兴的“搭配” (collocations)，邓耀臣的“搭配模式” (collocation patterns) 等。所有这些研究都认为，由多个词语组成并重复出现的片断具有显著的特点，从教学上看比孤立的词更值得重视和研究。此外，王克非等对“汉英平行语料库” (PCCE) 的报告、刘泽权利用平行语料库对虚构文本中报道性动词的翻译研究、杨牧陨等对基于平行语料库汉英词语翻译挖掘技术的研究以及李德俊对基于平行语料库的词典引用系统的研究等表明，我国在平行语料库研究方面已取得了初步成果。

(3) 基于中国口语语料库的中介语研究，占 11.4%。卫乃兴报告了 2001 年国家社科项目“大学学习者英语口语语料库” (COLSEC) 的建设情况，并根据该语料库分析了中国学生英语口语的语音错误特征、词块使用特征、话语结构模式和用于会话管理的语用策略。此外，何安平对学习者口语语料库中小品词进行了分析和调查，丁言仁、文秋芳对专业英语学生口语中套语 (formulaic sequences) 的运用做了描述和分析，王立非研究了学生口语中话语标记运用，梁茂成调查和分析了学生口语中的强势词。

(4) 英语语言与教学研究，占 21%。此类研究所占比重仅次于第一类研究。值得一提的是，John Sinclair 在本次会议上提出了构成语言能力的 4 种技能：1) 对口语或笔语文本进行切块的技能；2) 区分“离心” (exocentric) 和“向心” (endocentric)

结构的技能; 3) 识别和运用元语言 (language about language) 的技能; 4) 在各个层面上进行解释 (paraphrase) 的技能。这一全新的语言能力理念, 来自基于语料库的研究成果, 具有可靠的实证依据, 它对外语教学的意义值得进一步研究和论证。此外, Amy B. M. Tsui、He An E (何安娥)、濮建忠等结合外语教学实际对英语语言的个案分析表明, 语料库方法及研究成果对外语教学革新具有重要的意义, 势必影响外语教学的目的、内容以及教学方法各个环节。

(5) 英语变体研究, 占 4.7%。李文中对中国英语 (China English) 的前导研究 (pilot study) 以及 Anna Mauranen 对芬兰英语的研究提出了一个共同的问题: 即在当今英语日益全球化、各种非本族英语变体共存的形势下, 英语本族人的语言是否必须是惟一的标准? 任何一个英语学习者的目的不外乎是利用这一国际通用语 (ELF, English as a Lingua Franca〈Anna Mauranen 用词〉) 进行国际交流, 在保持自己文化身份 (cultural identity) 的同时, 增进对其他文化的理解和宽容, 而不是逐渐抛弃自己的文化身份, 彻底融入目的语文化中。

(6) 语料库技术研究, 占 4.7%。此类研究包括张宵军等的“网络语料库挖掘系统”以及“电子语料库建设与搜索系统”研究。比较起来, 此类研究在国内还比较薄弱, 其成果还难以形成共享的工具或产品。

3. 我国语料库语言学研究的基本特征

杨惠中教授在开幕式的发言中指出, 我国的语料库语言学研究从 20 世纪 80 年代中期第一个语料库 (上海交大科技英语语料库: JDEST) 起, 就与外语教学结下了不解之缘。由杨惠中主持建成的 JDEST 为我国大学英语教学大纲的制定和词表统计做出了积极的贡献。JDEST 是国际上第一代语料库, 在欧洲受到语料库语言学界广泛关注。杨惠中对语料库处理技术词汇和准技术词汇所提出的思想和原则为他在国际上赢得了学术声誉, 也深刻影响了以后的语料库研究。JDEST 为杨与欧洲语料库语言学界长期合作奠定了基础。

自上个世纪 80 年代以来, 国内已建成多个语料库, 如国际英语学习者语料库中国子语料库 (ICLE, 桂诗春)、中国学习者英语语料库 (CLEC, 桂诗春、杨惠中)。在建的语料库包括: 中国大学学习者英语口语语料库 (杨惠中)、中国专业英语学习者口语语料库 (文秋芳)、中国英语语料库 (CEC, 李文中)、中学英语口语语料库 (何安平) 等。这些语料库无一不与中国的外语教学紧密相连。这是因为, 一方面, 我国语料库语言学研究者本身就是英语教师, 其研究注定要密切关注中国外语教学的需求; 另一方面, 国际上已建成的语料库由于知识版权和其他因素, 大多不能为我国的研究者直接应用。(参见李文中, 2001) 我国的研究者如想开展真正意义上的研究, 只能依靠自己的力量开发出具有独立版权的语料库, 而不是依赖西方语料库资源。

我国语料库语言学研究的另一个主要特征是, 从一开始就具有横向合作、资源共享的良好态势。这是因为: (1) 语料库建设和开发需要大量的长期的人力物力投入, 如 COBUILD 语料库先后由百余位语言学家、统计学家、软件工程师及工作人员参加, 共花费了近十年时间。因此, 大型语料库项目不能只依靠单个学校或单位。(2) 中国的学术体制对语料库研究这种综合人文、社科、理工与计算机技术的边缘学科目前还不够重视, 在课题投入上仍按一般人文学科对待, 使得中国所有的语料库研究获得的资助严重不足。研究者需要自己动手, 不计报酬, 在极其艰苦的条件下开展研究, 并主动寻求横向联合, 集中资源, 勉力完成自己的项目。所幸这种广泛的合作也造就了一批语料库语言学研究者。

我国这方面研究的第三个特征是明确的应用取向和强烈的自主意识。多年来, 我国的外语研究大多停留在对西方现有理论的引介和诠释层面上, 很少开展基于本土实际的独立研究, 这使得中国的外语教师和研究者在国际相关领域长期以来少有发言权, 难以获得独立的身份和地位。语料库应用研究为中国的外语教师和研究者提供了向国际同行表达自己、展示并与他人分享成果的舞台。所以, 开展基于语料库的应用研究并与中国的外语教学实际紧密结合在一起, 既是中国语料库语言学学者自觉的选择, 也是中国外语研究发展的一种必然。基于语料库的外语教学研究为教师和学习者提供了一个全新的视野和平台, 显示出强大的生命力。

4. 存在的问题

本次语料库语言学国际会议也暴露了我国在这个研究领域的一些问题, 主要有: (1) 描述性研究呈现强势, 但缺乏解释和应用层面的进一步探索。任何研究既要求描述的完整性, 也要求解释的充分性。如果仅停留在描述层面, 而不能进一步进行理论解释和构建, 就会减弱研究的价值。(2) 对学习者语言的对比分析和个案研究只注重了非规范性特

征分析，缺乏全面系统的研究。值得指出的是，CLEC 的建立旨在对学习者语言的整体特征进行研究。学习者的语言运用并不仅仅只有错误，且其大部分非规范特征也不能简单地归结为错误，它们往往显示了学习者利用已获得的语言知识，为提高交际的有效性而采取的积极策略。（李文中，1999）此外，在对比分析中，以英语本族语为标准，必然导致本族语中心主义。在英语愈来愈全球化以及各种英语变体并存的今天，这种分析方法将逐渐失去其合理性。(3) 与语言研究相比，语料库技术开发研究相对薄弱。基于语料库的深度研究往往以先进的语料库处理技术为支撑。另外，构建语料库语言学系统理论仍是今后研究的一个重大课题。正如杨惠中在做大会总结时所言，进一步加强合作，重视技术开发和语料库的深加工，在语料库建设方面避免低层次重复，最大限度实现资源共享，同时注重技术方法培训，是我国语料库语言学研究今后发展中需要重点解决的问题。

注释：

① 本次会议主旨演讲和宣读论文篇目及全文请参见 http://www1.gzhtcm.edu.cn/bumen/yyxx/corpus（news and events）。相关图片报道请参见 http://home.henannu.edu.cn/fl/corpuspic.htm。

② John Sinclair 教授原为伯明翰大学语言学教授，是 COBUILD 的项目负责人，负责建成了世界上目前最大的语料库 Bank of English。他是 *Collins COBUILD English Dictionary* 等一系列词典、语法、教材的主编。

参考文献：

[1] 桂诗春. A Cognitive Model of Corpus-based Analysis of Chinese Learners' Errors of English [Z]. Keynote Speech at the 2003 International Conference on Corpus Linguistics at Shanghai. Shanghai: Shanghai Jiao Tong University. 2003.

[2] 李文中. 语料库与学习者语料库 [A]. 杨惠中. 语料库语言学导论 [C]. 上海：上海外语教育出版社. 2001.

[3] 李文中. *An Analysis of the Lexical Words & Word Combinations in the College Learners English Corpus* [D]. Unpublished PhD dissertation: Shanghai Jiao Tong University. 1999.

[4] 濮建忠. Noticing, Learning and Acquiring the Central Uses of Common English Words [Z]. Paper presented at the 2003 International Conference on Corpus Linguistics at Shanghai. Shanghai: Shanghai Jiao Tong University. 2003.

[5] 卫乃兴. A Preliminary Report on the COLSEC Project [Z]. Keynote Speech at the 2003 International Conference on Corpus Linguistics at Shanghai. Shanghai: Shanghai Jiao Tong University. 2003.

（责任编辑　严辰松）

语料库语言学发展趋势瞻望*

——2003语料库语言学国际会议综述

□甄凤超　张　霞

提要：语料库语言学研究经过40年的发展，不断成熟与完善。目前，语料库语言学的发展呈现出五大趋势：1)学习者语料库的建设与研究成为语料库语言学研究的重点之一；2)口语语料库的建设和相关话语特征研究不断加强；3)平行语料库在语言对比研究及翻译研究中的作用日益显著；4)语料库建设研究日益普遍化；5)语料库研究不断向纵深发展。这些趋势在"2003语料库语言学国际会议"上体现得更为明显。

关键词：语料库语言学；外语教学；国际会议；发展趋势

Abstract: In the past four decades, the field of corpus linguistics has been undergoing a revolutionary change and considerable achievements have been made in both corpus construction and corpus-based language studies and teaching. At the "2003 International Conference on Corpus Linguistics" held from October 25 to 27 at Shanghai Jiaotong University, five trends have been observed in the current development of corpus linguistics. They are: 1) learners' corpus construction and interlanguage study become one of the centrals of corpus linguistics; 2) spoken corpora attract more and more attention; 3) parallel corpora play a crucial role in comparative language studies and translation; 4) corpus construction becomes a popular practice among language teachers and researchers; 5) corpus-based studies develop both in scope and in depth.

Key words: corpus linguistics; EFL teaching; international conference; trends

中图分类号：H319　**文献标识码**：B　**文章编号**：1004-5112(2004)04-0074-04

自上世纪60年代初Francis和Kucera开始设计建设第一代大型电子语料库，即著名的布郎语料库(BROWN)至今，语料库语言学研究已经历了40年的发展历程。而中国语料库语言学研究也有近20个年头，最早可以追溯到上世纪80年代中期以上海交通大学杨惠中教授为首建成的JDEST学术英语语料库。在语料库语言学研究的发展历程中，各家学派著书立说，共同致力于这门学科的繁荣发展。目前，语料库语言学研究呈现出一些新的发展趋势，表现为如下几个方面：1)学习者语料库的建设和中介语的研究；2)口语语料库的建设和相关话语特征的研究；3)平行语料库的建设和研究；4)语料库建设的普遍性；5)语料库研究向纵深发展等。2003年10月25—27日在上海交通大学举行的、由来自英国、意大利、挪威、芬兰、新西兰、日本、新加坡和中国内地及香港地区等80余名语料库语言学研究者和专家参加的"2003语料库语言学国际会议"上，这些趋势体现得尤为突出。

首先，学习者语料库的建设与中介语的研究是今后语料库语言学研究的重点之一。上世纪末叶，学习者语料库的出现可谓异军突起，并很快成为当今语料库建设的一股新的力量(Granger 1998)。目前国外已建立的颇具影响的学习者语料库主要有：80年代末建立的Longman Learners' Corpus (LLC)，90年代中期建立的International Corpus of Learner English (ICLE)等。国内学习者语料库主要有：由广东外语外贸大学和上海交通大学共同建立的中国学习者英语语料库CLEC (Chinese Learners' English Corpus)，中国香港的Hong

* 感谢杨惠中教授在本文撰写过程中提供指导和修改意见。

Kong University of Science and Technology (HKUST) Learner Corpus 等(详情参照附录表一)。建立学习者语料库的目的是通过语料库方法深刻洞悉真实的学习者语言特征,最终服务于外语教学。事实上,中国的语料库建设从一开始就与外语教学密不可分,如最早建立的 JDEST 语料库,初衷即是为中国大学英语教学提供词汇以及技术词汇的应用信息,为大学英语教学大纲词表的确定提供可靠的量化依据。近年来 CLEC 以及其他学习者语料库的建立不仅为语言研究者提供了研究课题,而且为语言教学提供了有关学习者语言运用和典型困难的可靠信息。在本届会议上,桂诗春教授在会议主题发言"A Cognitive Model of Corpus-based Analysis of Chinese Learners' Errors of English"中,通过基于 CLEC 的错误分析,结合 MacWhinney 的语言习得竞争理论以及 Skehan 的语言学习认知法,在大量统计分析后建立了一个二语习得和错误分析的认知结构,对英语教学以及语言认知理论贡献良多。香港大学英语教师培训中心主任 Amy B. M. Tsui 教授在题为"Reframing the Object of Teaching and Learning: the Impact of Corpus Evidence on Language Teachers"的主题发言中,分类总结了香港地区英语教师在 9 年内通过 TeleNex 网站提出的 2000 个左右的英语语法问题,深入论述了语料库对英语教师语言意识(language awareness)的影响(会议论文集 2003)。另外,本次国际会议入选的 60 余篇论文中就有约占 65% 的论文涉及基于语料库的中介语研究和语料库在外语教学中应用的研究。另据中国期刊网的不完全统计,近年来在国内期刊杂志上发表的有关语料库的文章中有 35% 左右的论文直接涉及语料库在外语教学与学习中的应用。目前,建设学习者语料库的目的不再仅仅限于对学习者进行错误分析,服务于诊断式教学,更要扩展为对学习者的语言特征和语言发展进行全面系统的对比研究。

其次,口语语料库的建设和相关话语特征分析已成为该学科发展的一个方向。目前许多语言学家和教师认为口语比书面语更能揭示语言以及语言习得的本质,并通过语料库方法收集自然口语语料,进行口语话语特征分析。在本次国际会议上,芬兰 Tempere 大学国际通用语英语语料库 ELFA 项目负责人 Anna Mauranen 在主题发言"Spoken Corpora — Are Only Native Speakers Interesting?"中,详细论述了口语对语言研究以及语言教学的重要作用。目前国际上已建立的大型口语语料库主要有 The BNC Spoken Corpus, The London — Lund Corpus of Spoken English, CHILDES (Child Language Data Exchange System), ICE Spoken Texts 等。国内正在建设的汉语口语语料库主要有北京语言文化大学对外汉语研究中心的当代北京口语语料库,中国社会科学院的 Corpus of Situated Adolescent Speech 等;外语学习者英语口语语料库有上海交通大学的 COLSEC,南京大学的 SECCL,华南师范大学的 LINSEI-China 等(参照附录表一)。值得一提的是,这三个学习者口语语料库已统一了转写方案和标注规则,建成后将合而为一,成为中国最大的外语学习者口语语料库,库容量预计将达到 2,000,000词。这些语料库建成后将是国际上首批同类型的语料库,无论在语料库建设理论上,还是在技术方法、研究分析、应用开发上都属于开创性的工作。目前这些口语语料库的学术科研地位与教学应用价值已日益显著。如文秋芳在会议发言"A Corpus-based Analysis of the Use of Frequency Adverbs by Chinese University English Majors"中,基于 CEMC 和 SECCL,对英语专业大学生书面语以及口语中 TTFA (top twenty frequency adverbs)使用情况的不同,以及对中国英语专业大学生与英语母语使用者对 TTFA 使用情况的不同做了统计分析,试图探究他们的模式。卫乃兴在会议发言"Investigating Characteristics of Chinese Learners' English Speech"中,就正在建设中的 COLSEC 做了报告,并从学习者的典型发音错误、特定词块使用情况、学习者话语结构、话语模式以及常用语用策略等五个方面对当前基于 COLSEC 的研究做了详尽介绍(会议论文集 2003)。当然,目前口语语料库建设还存在一些问题,如口语语料库与书面语语料库发展不平衡,目前国内口语语料库的语料与自然交际语境的语料仍有较大不同等。

第三,平行语料库的建设和研究代表了当今基于语料库方法进行语言对比研究、翻译研究和外语教学研究的发展趋势。平行语料库作为语言对比分析与翻译研究的一项重要工具,对于促进语言对比研究和翻译研究,改进外语教学,提高翻译质量,改进双语词典的编纂,促进双语信息检索和机器翻译的开发都具有深远的意义。本次会议上,英国伯明翰大学语料库语言学中心主任 Wolfgang Teubert 在主题发言"Parallel Corpora and Language Teaching"中指出,平行语料库为学习者提供机会去自主发现母语与目的语在语法结构(grammatical structure)、词汇语域(vocabulary register)和意义表述(content expression)上的差别,从而达到学好外语的目的。

新加坡国立大学刘泽权提交的论文"A Corpus-Based Study of Reporting Verbs in Fictions: A Translational Perspective",通过建立两个平行语料库,即《红楼梦》原著及其英文翻译和《苔丝》原著及其汉译本,研究两种语言在使用"引出间接引语的动词"(reporting verbs)方面的区别(会议论文集2003)。世界上许多国家和地区已相继建立或正在建立各种双语甚至多语平行语料库,如90年代末建成的The English-Norwegian Parallel Corpus等。我国近年来也加大力度开发和建设平行语料库。目前主要有北京外国语大学中国外语教育研究中心正在建设的汉英平行语料库PCCE(Parallel Corpus of Chinese and English)等(参照附录表一)。建设平行语料库面临的瓶颈是语料库文本语言单位对应(alignment)的精确性,这主要是因为不同的语言在语序、句子结构和逻辑意义的表达方面都存在着明显的差异。

第四,语料库建设日益呈现出普遍性的特点。自上世纪80年代起,由于计算机科学的飞速发展以及计算机技术在语言研究领域中的迅速普及和应用,再加上人们逐渐意识到转换生成语法学派的片面性和局限性,越来越多的语言研究者跻身到语料库建设和研究的队伍中,使得语料库语言学研究首先在欧洲各国蓬勃发展,并有逐渐成为语言研究主流的趋势(Thomas 1996)。这期间相继出现了一些具有代表性的大型语料库,如COBUILD语料库、朗文语料库(the Longman Corpus Network)、英国国家语料库(the British National Corpus)、国际英语语料库(the International Corpus of English)等。而Sinclair在本届会议上指出,上世纪90年代末大型语料库建设的势头已缓,代之而起的是大批小型语料库的兴起。大量在线电子语料以及通过各种电子媒介发行的电子文本为语料库的建设提供了无尽的语料来源,使得建设各种小型语料库变得轻而易举。语言研究者可按照研究兴趣和方向自己建设形式多样的语料库。如解放军外国语学院的军事英语语料库(Corpus of Military Texts),河南师范大学在建的中国英语语料库(China English Corpus)等(参照附录表一)。

最后,语料库的研究不断向纵深发展。人们在借助语料库方法对语言系统以及人们对语言系统使用情况进行研究的同时,也加深了对语料库语言学本身的研究。语料库语言学不仅仅是一种语言研究方法,更是代表着一种新的哲学思维方式,深刻影响着人们对语言的认识和研究。

以上是对本届国际会议上呈现出的语料库语言学发展趋势的瞻望。杨惠中教授在开幕式致辞中高度概括了语料库建设和研究的现状:一是容量扩大,使得基于概率而非规则的研究更为可靠;二是纵深发展,从词和短语的研究上升到句子和篇章的层面;三是应用范围扩大,从早期的词典编纂、词频研究到如今的语音识别、信息检索和课堂教学。此外,杨教授也指出了目前语料库建设和研究中存在的一些普遍问题,如低层次语料库重复建设,软件开发不足,系统的理论研究欠缺以及语料库在外语教学中的应用尚欠发达等。这些问题都有待研究者共同努力解决。□

参考书目

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

[2] Sinclair J. *Corpus, Concordance, Collocation* [M]. Oxford: Oxford University Press, 1991.

[3] Thomas J and M Short. *Using Corpora for Language Research* [M]. Longman Group UK limited, 1996.

[4] Yang Huizhong and Wei Naixing (eds). *Conference Proceedings: 2003 International Conference on Corpus Linguistics*[C]. 2003.

作者单位:上海交通大学外国语学院,上海 闵行 200240

附录表一:国内语料库建设一览表

类　型	语料库名称及库容(型符数)	建设单位
英语学习者语料库(书面语及口语)	中国学习者英语语料库 CLEC(1,000,000)	广东外语外贸大学和上海交通大学
	大学英语学习者口语语料库 COLSEC(50,000)	上海交通大学
	香港科技大学学习者语料库 HKUST Learner Corpus	香港科技大学
	中国英语专业语料库 CEMC(1,480,000)	南京大学
	中国英语学习者口语语料库 SECCL(1,000,000)	南京大学
	国际外语学习者英语口语语料库中国部分 LINSEI-China(100,000)	华南师范大学
	硕士写作语料库 MWC(120,000)	华中科技大学
平行语料库	汉英平行语料库 PCCE	北京外国语大学
	南大—国关平行语料库	南京大学
	英汉文学作品语料库; 冯友兰《中国哲学史》汉英对照语料库; 李约瑟(Joself Needham)《中国科学技术史》英汉对照语料库	外语教学与研究出版社
	计算机专业的双语语料库; 柏拉图(Plato)哲学名著《理想国》的双语语料库	国家语言文字工作委员会语言文字应用研究所
	英汉双语语料库(15 万对)	中国科学院软件研究所
	英汉双语语料库:LDC 香港新闻英汉双语对齐语料 36294 段以及香港法律英汉双语对齐语料 31 万句子对,并从英汉双解词典中摘取例句 25000 个句子对	中国科学院自动化研究所
	英汉双语语段库(1,000,000),网上英汉语段电子词典及网上电子英汉搭配词典(10,000,000)	东北大学
	英汉双语语料库(40 - 50 万句子对)	哈尔滨工业大学
	双语语料库(5 万多对)	北京大学计算语言学研究所
	对比语料库 LIVAC(Linguistic variety in Chinese communities)	香港城市理工大学
	平衡语料库(Sinica Corpus);树图语料库(Sinica Treebank)	台湾
特殊英语语料库	中国英语(China English)语料库	河南师范大学
	军事英语语料库(Corpus of Military Texts)	解放军外语学院
	新视野大学英语教材语料库	上海交通大学
汉语语料库	汉语现代文学作品语料库(1979 年,527 万字)	武汉大学
	现代汉语语料库(1983 年,2,000 万字)	北京航天航空大学
	中学语文教材语料库(1983 年,106 万 8 千字)	北京师范大学
	现代汉语词频统计语料库(1983 年,182 万字)	北京语言学院
	国家级大型汉语均衡语料库(2,000 万字)	国家语言文字工作委员会
	《人民日报》语料库(2,700 万字)	北京大学计算语言学研究所
	大型中文语料库(5 亿字,10 分库)	北京语言文化大学
	现代汉语语料库(1 亿字)	清华大学
	汉语新闻语料库(1988 年,250 万字); 标准语料库(2000 年,70 万字)	山西大学
	生语料库(3,000 万字);《作家文摘》的标注语料库(100 万字)	上海师范大学
	现代自然口语语料库	中国社会科学院语言所
	旅游咨询口语对话语料库和旅馆预定口语对话语料库	中国科学院自动化所

(注:此表由卫乃兴、甄凤超、张霞提供,部分参考冯志伟《中国语料库研究的历史与现状》*Journal of Chinese Language and Computing*,11(2))

2003 语料库语言学国际会议纪要

上海交通大学、上海外语教育出版社和广东外语外贸大学联合举办的“2003 语料库语言学国际会议”于去年 10 月 25 日至 27 日在上海交通大学隆重举行。会议主题为:“语料库语言学与外语教学”,工作语言为英语。与会代表 80 余人,分别来自英国、意大利、挪威、芬兰、新西兰、日本、新加坡、中国香港和中国内地。

会议开幕式由杨惠中教授主持。上海交通大学副校长沈为平、上海交通大学外国语学院院长王同顺、上海外语教育出版社汪义群教授分别在大会致辞,并预祝会议圆满成功。杨惠中教授在开幕式致辞中高度概括了当今语料库语言学研究的三个维度:一是容量扩大,使得基于概率而非规则的研究更为可靠;二是纵深发展,从词和短语层面上升到句子和篇章层面的研究;三是应用范围扩大,从早期的词典编纂、词频研究发展到如今的语音识别、信息检索和课堂教学。杨惠中教授还总结了目前国内语料库研究的主要特征,如直接参与语料库建设研究的人日益增加、建成以及在建的语料库越来越多、英语学习者语料库以及口语语料库越来越受重视、语料库检索以及应用软件不断开发等。此外,杨教授也谈及语料库建设和研究中存在的一些问题,如低层次语料库重复建设、软件开发不足、系统的理论研究欠缺以及语料库在外语教学中的应用尚欠发达等。

会议特邀专家中,英国皇家学会会员、英国文化协会顾问、欧洲学术委员会委员 John Sinclair 教授首先作了主题发言。发言包括三个部分,一是回顾现代语料库语言学发展的 40 年历史,追溯语料库语言学与计算语言学以及自然语言处理的历史渊源;二是重申语料库语言学的哲学以及语言学理论基础;三是例释借助语料库工具研究自然语言事实的过程。广东外语外贸大学桂诗春教授在题为“A Cognitive Model of Corpus-based Analysis of Chinese Learners' Errors of English”的发言中,通过基于 CLEC 的错误分析,结合 MacWhinney 的语言习得竞争理论以及 Skehan 的语言学习认知法,在大量统计分析后建立了一个二语习得和错误分析的认知结构。芬兰大学国际语英语语料库 ELEA 项目负责人 Anna Mauranen 教授在主题发言“Spoken Corpora — Are Only Native Speakers Interesting?”中区分了 EFL (English as a foreign language) 和 ELF(English as a lingua franca),并强调了英语作为国际语的理论和现实意义。香港大学英语教师培训中心主任 Amy B.M. Tsui 教授在题为“Reframing the Object of Teaching and Learning: The Impact of Corpus Evidence on Language Teachers”的主题发言中,从语料库语言学研究应用于语言教学的角度出发,分类总结了香港地区英语教师九年内在 TeleNex 网站提出的 2000 个左右的英语语法问题,深入论述了语料库对英语教师语言意识(language awareness) 的影响。英国伯明翰大学语料库语言学中心主任 Wolfgang Teubert 教授在题为“Parallel Corpora and Language Teaching”的主题发言中,从语言教学和翻译的角度详细论述了平行语料库的重要性,并强调了“意义体”(unit of meaning) 在外语学习和翻译中的作用。

除上述特邀专家的主题发言外,本次语料库国际会议共收到国内外寄来论文百余篇,有 60 多篇论文入选,并有近 40 篇论文在会议上宣读。如(按先后顺序):河南师范大学李文中博士的“Word Cluster, Phrases, and Collocations in China's English News Articles”;南京大学文秋芳教授的“A Corpus-based Analysis of the Use of Frequency Adverbs by Chinese University English Majors”;上海交通大学卫乃兴博士的“Investigating Characteristics of Chinese Learners' English Speech”;华南师范大学何安平教授的“‘Small Words' in EFL Learners' Spoken Corpora”;香港教育学院何安娥的“Use of Verbs in Teacher Talk: A Study of Comparison between Local English Teachers and Native English Teachers in Hong Kong Primary Schools”等等。论文内容广泛涉及语料库建设、基于语料库的语言研究、学习者英语研究以及应用软件开发等诸方面。会议气氛热烈,成效显著,非常成功。

本届会议是国内首次语料库语言学国际会议。会议的胜利召开,不仅增加了国内外语料库语言学界同仁们之间的相互了解,而且推动了国内语料库语言学研究的进一步发展。此外,与会代表对上海交通大学各级领导对大会的支持和重视表示赞赏,对上海交通大学外国语学院出色的组织工作和热情周到的接待表示衷心感谢,并一致希望今后能多举行这类具有高学术水平的专业性会议。□ (甄凤超)