LEARNER CORPUS RESEARCH: SOME RECENT TRENDS

Yukio TONO
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

ABSTRACT

This paper aims to describe some recent trends in learner corpus research. First, the results of a survey on more than 400 related studies on learner corpora will be reported in order to disclose the change of research interest and orientations. Then, as an example of such new perspectives, a new area of research based on the Common European Framework of Reference (CEFR) and corpus-based research to identify criterial features across the CEFR levels will be presented. Finally, a few case studies were reported to show how this new research programme could be meaningfully integrated into the mainstream of learner corpus research.

1. INTRODUCTION

The Centre of English Corpus Linguistics at University of Louvain, where the first computer learner corpus project was launched in the early 1990s, will celebrate its 20th anniversary next year. This means that we have a twenty-year history of learner corpus research (LCR) by now. When I started my project of compiling essay corpora by Japanese-speaking learners of English in the late 1980s, there were very few articles on learner corpora. Now, as will be seen in the next section, we have more than 400 research papers and books on this particular theme. The importance of learner corpora for the field of corpus linguistics and second language acquisition (or foreign language teaching for that matter) has now been duly acknowledged. The genuine question we might want to ask, therefore, is: What has been accomplished so far? In this paper, I will first make a brief survey of bibliography of learner corpora in order to see what has been investigated so far. Then I will discuss in more details some recent trends in LCR, especially focusing on research into criterial features in the CEFR-labelled corpora.

2. A SURVEY OF BIBLIOGRAPHY ON LEARNER CORPORA

A survey was conducted on 438 articles and books on learner corpora. This bibliography has been kept updated, covering major books of selected papers, conference proceedings, academic journals, monographs, and Ph.D. dissertations. It covers the publications till spring 2010. Table 1 shows the breakdowns.

Table 1. Breakdowns of the bibliography

Category	Number
Books; articles in books	179
Journal papers	86
Papers in conference proceedings	150
Dissertations (Ph.D.)	23
Total	438

I carried out a survey in the following way. All the titles were extracted and put into a text file, over which a textual analysis was performed to produce frequency statistics of terms and phrases used in the titles. We hoped that this would reveal some aspects of research in LCR. I will summarize the results of the frequency analysis as follows.

2.1 Corpora

There are many references to corpora used in the studies. As shown in Table 2, some papers use the terms "learner corpus/corpora" (87 times), which is contrasted with "native English/speakers/corpora" (55 times). Thus, a comparison is usually made between native and non-native corpora. Another common distinction is written vs. spoken. Written corpora seem to have been used 2-3 times more often than spoken corpora.

Table 2. Types of corpora

learner corpora	47	spoken	23	ICLE	28
learner corpus	40	written	22	NICT JLE	6
writing	53	speech	18	PELCRA	2
speaking	8	native	55		

2.2. Countries

There are many references to learners' countries as profile. Table 3 shows major countries that appear in the titles.

Table 3. Countries

Japanese	35	Spanish	7	Brazil	2
Chinese	16	Polish	7	British	2
Swedish	12	French	4	Bulgarian	2
Italian	10	American	3	Hong Kong	2
German	8	Belgium	2		

The fact that Japanese learners were mentioned very frequently in the titles shows that research using learner corpora is very active in Japan. The rest of the countries belong to Europe with a few exceptions (Chinese, Brazil, Hong Kong), which shows that most of those papers use the same corpora called the International Corpus of Learner English (ICLE). All the other papers not listed here may not refer to specific countries in the titles, but the patterns should not be very different from the findings here.

2.3. Research topics

2.3.1 Overall tendencies

Table 4 shows main research topics expressed in the titles. Most papers deal with specific aspects of learner language such as the ones in Table 6 and thus not many papers actually use as a title broader terms listed in Table 4. However, these terms should suffice to show that research topics in LCR are mainly concerned with second language acquisition, language pedagogy (teaching, data-driven learning, CALL), materials design (dictionaries) and evaluation. The terms such as "annotation/tagging" show that encoding linguistic information including learner's errors into a corpus is also an important issue.

Table 4. Research topics

acquisition	17	lexicography/dictionaries	16
application	15	annotation/tagging	7
teaching	15	data-driven learning	3
pedagogical	7	CALL	2
evaluation	5		

Learner's errors are very important elements in learner corpora, and thus should be treated separately here. Table 5 shows how various aspects of errors are mentioned in the titles:

Table 5. Aspects of errors

error		formal ~	2
~ analysis	2	interlanguage ~	2
~ annotation	2	phraseological ~	2
~ coding	1	grammatical ~	2
~ detection	1	referential ~	2
~ diagnosis	1	triggers for ~	2
~ identification	1	lexical ~	1
~ patterns	1	spelling ~	1
pragmatic ~	4	collocational ~	1

Besides various types of errors investigated in the articles, annotations or coding of errors are crucial issues. Also automatic detection or identification of errors is also a theme that attracts much attention these days.

Table 6 summarises specific linguistic features shown in the titles. General categories of topics indicate that grammar and lexis are the central issues. The word "syntactic" is not as popular as others, due to the background of learner corpus research rooted in the European tradition of descriptive and functional linguistics. This is also reflected in the other terms such as 'discourse,' 'pragmatics' and 'phraseology.'

Table 6. Linguistic features

General categories:					
grammar/grammatical	18		phraseology	10	
word	17		discourse	13	
vocabulary	8		pragmatic	4	
syntactic	4				
Specific categories:					
verb	17		formulae/ formulaic sequences	4	
causality/causal/causative		4	construction	4	
progressive		3	interlanguage ~		1
past tense forms		2	tough-movement ~		1
high-frequency verbs		1	causative ~		1
argument realization		1	support verb ~		1
lexical	16		pragmatic	4	
~ bundles		1	apology production		1
~ patterns/patterning		2	involvement		3
~ profiling		1	cohesive devices	3	
~ variation		2	tense	3	
adverb/adverbial	12		anaphora	2	
adverbial connectors		2	adjective intensification	2	
adverb placement		1	discourse marker	2	
time adverbials		1	epistemic modality	2	
stance		5	form-function mapping/split	2	
interlanguage	8		grammatical morpheme	2	
~grammar		2	idomaticity	2	
~construction		2	noun phrase development	2	
aspect	8		prefabs	2	
article	5		anticipatory 'it'	1	
indefinite article		1	auxiliary	1	
zero article		2	false friends	1	
collocation	5		inflectional vs. non-inflectional languages		1

Linguistic features dealt with in specific categories closely correspond with general categories. Grammar is always central, which is indicated by frequently mentioned key terms such as 'verb,' 'adverb/adverbial,'

'tense' and 'aspect' among others. Recently, however, the term 'construction' has been used more widely to denote constructions as meaningful building blocks for language acquisition. Focus on lexical knowledge is also clear, which is shown in terms such as 'lexical bundles,' 'lexical patterns,' 'collocation,' 'prefabs,' among others. Discourse and pragmatic knowledge has been studied, which is mainly shown in terms such as 'adverbials,' 'cohesive devices,' 'discourse marker' among others.

2.3.2 Recent trends

In order to capture recent trends in LCR, publications before and after the year 2001 were compared. The titles of 141 articles and books published before 2001 were compared against the rest (total n= 438). I used log-likelihood to extract significantly more frequent terms used in the publications before 2001. Table 7 shows the results:

Table 7. Keywords before and after 2001

Rank	freq	LL	Before 2001	Rank	freq	LL	After 2001
1	7	10.055	ICLE	1	30	6.773	Japanese
2	9	8.587	International	2	24	5.116	spoken
3	3	6.542	American	3	9	3.053	contrastive
4	3	6.542	compositions	4	22	3.035	use
5	3	6.542	exploiting	5	12	2.758	based
6	3	6.542	student	6	11	2.456	patterns
7	3	6.542	tagging	7	2	2.456	second
8	3	6.542	Uppsala	8	9	2.456	verb
9	3	6.542	USE	9	9	2.370	language
10	5	6.316	advanced	10	14	2.168	acquisition
11	6	5.725	lexical	11	7	1.887	proficiency
12	4	4.538	clauses	12	4	1.815	research

Table 7 shows that before 2001 the primary data source was the International Corpus of Learner English (ICLE), thus the keywords such as 'ICLE' or 'International' were part of them. Uppsala Student Corpus (USE) was also a unique essay corpus used back then, but they did not appear in the papers after 2001. After 2001, on the other hand, various learner corpora appeared in the field, including the Japanese EFL learners' corpora such as JEFLL or NICT JLE. Here again, the keyword 'Japanese' shows that the number of papers using JEFLL or NICT JLE increased dramatically after 2001. Another interesting contrast lies between written and spoken texts. Before 2001, most learner corpora were written, as is indicated by the keywords 'compositions,' which is overtaken by the term 'spoken' after 2001. More and more attention has been paid to spoken (oral) learner corpora.

After 2001, more emphasis has been on linking learner corpus findings to second language acquisition (shown in the keywords 'second' 'language' and 'acquisition') and the methodological terms such as 'contrastive' or 'comparative' only appear in the titles of papers published after 2001. Finally, while the papers before 2001 seemed to focus on lexical and clause/sentence structures, the publications after 2001 focused on 'patterns' of verb, lexis, and phraseology across different 'proficiency' levels.

2.4. Section summary

This brief survey of articles on learner corpora suggests that the research areas of learner corpora have been diversified in terms of corpus resources, approaches, and topics. While ICLE was almost the only source until 2000 (the only exceptions were HKUST corpus and Longman Learner's Corpus), various other types of learner corpora have been added. Major differences between ICLE and those corpora are modes (written vs. spoken) and proficiency levels (static vs. variable). More attention has been paid to identifying linguistic patterns (use/misuse) across different modes of learner output (spoken vs. written) as well as their proficiency levels. Among them, the most significant recent trend in LCR would be the integration of learner corpus-based approach into the research within the framework of the Common European Framework of Reference

(CEFR). In the following section, I will make a brief sketch of this approach and discuss the implications of previous findings in LCR to this area.

3. LCR MEETS CEFR: THE ENGLISH PROFILE

3.1. What is the CEFR?

The Common European Framework of Reference (CEFR) is a descriptive scheme that can be used to analyse L2 learners' needs, specify L2 learning goals, guide the development of L2 learning materials and activities, and provide orientation for the assessment of L2 learning outcomes (Little 2006: 167). The descriptive scheme has a vertical and a horizontal dimension. The vertical dimension uses 'can do' descriptors to define six levels of communicative proficiency in three bands (A1, A2 – BASIC USER; B1, B2 – INDEPENDENT USER; C1, C2 – PROFICIENT USER). There are also scales for LISTENING and READING (reception), SPOKEN PRODUCTION, WRITTEN PRODUCTION, SPOKEN INTERACTION and WRITTEN INTERACTION. The horizontal dimension is concerned with the learners' communicative language competences and strategies.

Since the CEFR is language-independent, each country in Europe is now developing a set of so-called Reference Level Descriptions (RLDs), a set of linguistic features that are criterial for each CEFR level. The French publisher of the CEFR has begun to produce a series of reference books each of which is devoted to a single proficiency level in French (Beacco et al. 2004, 2006). A project with German, Swiss, and Austrian funding has developed *Profile deutsch*, an interactive CD-ROM that presents the CEFR in German together with a functional-notional resource, functional and systematic treatments of German grammar, among others (Glaboniat et al. 2005). In the case of English, the English Profile (EP) is responsible for RLD development. One unique feature of the EP programme is that they aim to identify 'criterial features' of English for each CEFR level in a corpus-driven approach.

3.2. Criterial features and LCR

Salamoura and Saville (2009) defined a 'criterial feature' as follows:

> A 'criterial feature' is one whose use varies according to the level achieved and thus can serve as a basis for the estimation of a language learner's proficiency level. So far the various EP research strands have identified the following kinds of linguistic feature whose use or non-use, accuracy of use or frequency of use may be criterial: lexical/semantic, morpho-syntactic/syntactic, functional, notional, discourse, and pragmatic.
> (Salamoura and Saville 2009:34)

What is unique in their project is that they seek for criterial features by looking at learner corpora with the CEFR level classifications. Hawking and Buttery (2009), for example, have identified four types of feature that may be criterial for distinguishing one CEFR level from the others. Table 8 shows the classifications:

Table 8. Possible criterial feature types

Type of feature	Descriptions
Acquired/Learnt language features	Correct properties of English that are required at a certain L2 level and that generally persist at all higher levels. E.g. property P acquired at B2 may differentiate [B2, C1 and C2] from [A1, A2 and B1] and will be criterial for the former.
Developing language features	Incorrect properties or errors that occur at a certain level or levels, and with a characteristic frequency. Both the presence versus absence of the errors, and the characteristic frequency of error can be criterial for the given level or levels. E.g. error property P with a characteristic frequency F may be criterial for [B1 and B2].
Acquired/Native-like usage distributions of a correct feature	Positive usage distributions for a correct property of L2 that match the distribution of native speaking (i.e. L1) users of the L2. The positive usage distribution may be acquired at a certain level and will generally persist at all higher levels and be criterial for the relevant levels.

Developing/Non-native-like usage distributions of a correct feature	Negative usage distributions for a correct property of L2 that do not match the distribution of native speaking (i.e. L1) users of the L2. The negative usage distribution may occur at a certain level or levels with a characteristic frequency F and be criterial for the relevant level(s).

The EP researchers have done preliminary studies with regard to the criterial features, using the Cambridge Learner Corpus (CLC) (Williams 2007; Parodi 2008; Hendriks 2008; Hawkins and Buttery 2009; Filipovic 2009). The CLC currently comprises approximately 30 million words of written learner data, roughly half of which is coded for errors. It has been also parsed using the Robust Accurate Statistical Parser (RASP) (Briscoe, Carroll, and Watson 2006). As the reports showed, the CLC mainly covers A2 level and above, which is the reason why they started to build a new corpus called the Cambridge English Profile Corpus (CEPC), mainly focusing on lower-proficiency level students' writing and speech.

Considering the sheer size of the CLC with error annotations and the CEFR as a framework, this EP programme seems to create a new research paradigm in LCR. Those who are interested in using learner corpora in SLA research can relate their findings to the EP researchers' findings in terms of criterial features. Those who are involved in syllabus/materials design will find the RLDs for English very informative once those items are actually identified. Test developers will make full use of the results of the EP research for improving their test design and contents.

Since I have done a series of research to find criterial features across proficiency levels using the JEFLL Corpus, my goal is quite similar to what the EP researchers aim at, although I was not aware of the availability of the CEFR until recently. In the following sections, I will try to relate my previous findings to some of the principles and hypotheses proposed by the EP programme to see how much previous research can answer to some of the issues raised by the EP researchers.

4. LINKING PREVIOUS RESEARCH TO THE EP RESEARCH FINDINGS

4.1. Verification of the findings of the EP programme

As a researcher, it is important not to accept any research findings blindly. Since the influence of the EP programme will definitely increase in the future, it is necessary to have an objective method of verifying their claims and findings as a scientist. One big problem for this objectivity is the fact that the CLC is not publicly available. It is an in-house resource at the Cambridge ESOL and the Cambridge University Press for test and materials development, it is a shame that there is no way to check the findings against the actual corpus data.

There are mainly two ways of verifying the claims by the EP programme. One is to verify their findings against comparable learner corpora. The ICLE will not do the job, since it is comprised of only the university students' essays with the assumingly same proficiency levels. I will show how the JEFLL Corpus can be used to produce similar types of results with some implications. Another way is to examine previous LCR findings to see whether they have already confirmed some of the hypotheses proposed by the EP researchers. I will show an example of this case. Finally, I will bring up some methodological issues and future directions.

4.2. Verification I: new verb co-occurrence frames

One of the findings in the EP programme is the progression pattern of new verb co-occurrence frames. Williams (2007) has found that there is a clear progression in the data from A2 to B2 in the appearance of new verb co-occurrence frames. For the sake of brevity, I will list some of the major frames in Table 9. For the full list, see Hawkins and Buttery (2008). Their original hypothesis was that there is a progression from A2 to C2, but she has found no evidence for new verb co-occurrence frames at the C levels. It appears that these basic constructions of English have been acquired by B2. Hawkins suggested that they require a different kind, and a more subtle kind, of analysis in order to capture progress at the C levels (ibid: 12).

Another interesting finding by Williams is that the progression from A2 to B2 correlates with the frequencies of these co-occurrence frames in the British National Corpus (BNC). In other words, learners are first learning the more frequent frames used by English native speakers and then progressively less frequent

frames. Table 10 shows the average token frequencies for the verb co-occurrences found by Williams, and also their average frequency ranking, in a number of corpora including the BNC.

These findings seem to be solid research findings from the CLC, but I notice that there are some confounding results in my previous studies, thus we should be careful interpreting these findings. For an example, I will demonstrate the case in which my learner corpus data will show a slightly more complicated picture in this acquisition of new verb co-occurrence frames.

Table 9. New verb co-occurrence frames in different CEFR levels (based on Williams 2007)

Verb co-occurrence frames	Examples	CEFR level
NP-V	*He went*	A2
NP-V-PP	*They apologized [to him]*	A2
NP-V-NP-PP	*She added [the flowers] [to the bouquet]*	A2
NP-V-VPinfinitival (Subj Control)	*I wanted to play*	A2
NP-V-NP-NP	*She asked him [his name]*	B1
NP-V-VPinfin (Wh-move)	*He explained [how to do it]*	B1
NP-V-S (Wh-move)	*He asked [how she did it]*	B1
NP-V-P-S (whether = Wh-move)	*He thought about [whether he wanted to go]*	B1
NP-V-NP-AdjP (Obj Control)	*He painted [the car] red*	B2
NP-V-NP-as-NP (Obj Control)	*I sent him as [a messenger]*	B2
NP-V-NP-S	*He told [the audience] [that he was leaving]*	B2
NP-V-P-VPinfin (Wh-move)(Subj Control)	*He thought about [what to do]*	B2

Table 10. Frequencies for verb co-occurrence frames in English corpora (including BNC)

Average token frequencies in the BNC etc. for the verb co-occurrence frames appearing at each learner level

A2	B1	B2/C1/C2
1,041,634	38,174	27,615

Average frequency ranking in the BNC etc for the verb co-occurrence frames appearing at each learner level

A2	B1	B2/C1/C2
8.2	38.6	55.6

Firstly, I will show from my previous research findings that the acquisition of verb co-occurrence frames has some interactions with input frequencies as well as internal complexities of the given verbs. Tono (2002) compared the acquisition of verb subcategorization frames for ten major lexical verbs in English. I extracted all the instances of those verbs from the JEFLL Corpus, a corpus of Japanese EFL learners' written compositions, and classified the examples in terms of their subcategorization frame (SF) patterns and use vs. misuse. I performed log-linear analysis to see how explanatory variables such as L1 effects (i.e. degrees of proximity between L1 and L2 verb SF patterns), L2 internal effects (i.e. verb semantics and frequencies of SF patterns in native corpora) and L2 input effects (frequencies of SF patterns in English textbook corpora) affect the frequencies of use/misuse of SF patterns in L2 writings. The statistical analysis shows a complex picture; whilst there was a positive correlation between the input frequencies in textbook corpora and the frequencies of occurrences of the given SF patterns, there was no relationship between the textbook frequencies and the number of errors. The error frequencies, however, correlate positively with the mixed effects of L1 effects (similarities of SF patterns between L1 and L2) and verb semantics.

This result in Tono (2002) suggests that the findings by Williams (2007) in the CLC have to be interpreted carefully. Overall, Hawkins claims that more frequent properties in L2 are more easily acquired, in general; fewer errors, more of the relevant L2 properties learned and earlier acquisition and infrequency will have the reverse effects (ibid: 9). Williams' findings seem to be such a case. However, there are many grammatical items such as definite and indefinite articles, which are very high in frequency but very late in acquisition. Verb co-occurrence frames should also exhibit more complex patterns than they have found, when they take into account such factors as the behaviour of individual lexical verbs, the patterns of use vs. misuse in the co-occurrence patterns.

Secondly, let me verify some of the findings in Table 9 against my JEFLL data. I conducted a small study to extract the following verb co-occurrence frames from the JEFLL:

(a) NP-V-VPinfinitival (Wh-move): *He explained how to do it.*

(b) NP-V-S (Wh-move): *He asked how she did it.*

In order to extract the above frames from JEFLL, I searched for "V + *how*" patterns. Out of 85 instances in total, the verb '*know*' occurred 36 times. For this preliminary analysis, I chose only this verb and examined the use/misuse of the two frames (a) and (b) above. The results are shown in Figure 1:

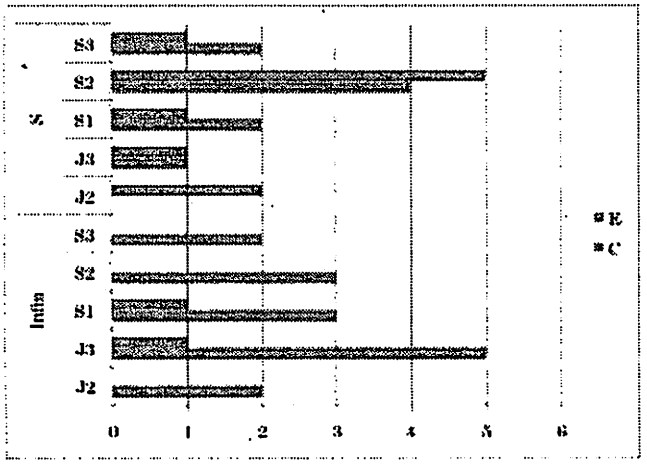

Figure 1. Frequencies and accuracy rates of "know + how to V" (Infin) vs. "know how + S" (S) (keys: E=error; C=correct; J2/3=junior high 2nd and 3rd year; S1/2/3=senior high 1st to 3rd year)

As shown in Figure 1, the NP-VP-Vinfinitival construction (i.e. '*know how to ...*') appeared earlier in junior high students' writings and accuracy rates are relatively high (88.2%). The NP-VP-S construction, on the other hand, appeared much later and with low accuracy rates (47.4%). Despite the fact that the frequencies of these constructions in JEFLL are relatively low, the error analysis in Figure 1 suggests that the criterial features should be identified using multidimensional perspectives. The fact that the two verb co-occurrence frames classified as the same B1 level actually show different performance results in my data is a good example that mere frequency analysis of constructions is not sufficient. We should take into account both correct and incorrect uses of those frames. Williams (2007) was cited by Salamoura and Saville (2009) as an example of acquired language features and native-like usage distributions, but it is not at all clear how she actually distinguished correct frames from erroneous frames in the parsed data. It would not be straightforward to parse the learner data and automatically extract frame frequencies. I doubt that she took into account only correct usage distributions.

Another implication is that we can use the results by Williams and my data for different purposes. For example, my data shows that two frames labeled as B1 did not seem to work in the same way. This could be useful information to further break down the B1 levels. As the Swiss research project suggested, the CEFR levels could be further subdivided up to nine levels and the B1 level can be subclasses of B1 and B1+ (Council of Europe 2001:31). Therefore, the overall level description could be done using the average frequencies of verb co-occurrence frames and further subdivision can be made using information on the accuracy rates for each construction.

4.3. Verification II: Lexical choice errors

In the descriptions of the UCLES-RCEAL funded research projects, Hawkins presented his project on hypothesis formation and testing using the CLC. In the report, he has identified a set of 20 lexical and grammatical areas that are promising initial candidates for criterial feature identification and transfer effects at different proficiency levels. I do not have a space to elaborate on those 20 hypotheses here, but I will take

the first hypothesis on lexical choice errors to show that my previous research using Japanese learner corpora has already confirmed some of his hypotheses.

He described lexical choice errors as follows and proposed three relevant hypotheses:

Lexical Choice Errors: Noun (N) and Verb (V)

RN	Replace noun	Have a good travel!
FFN	False friend noun	It was an interesting history
RV	Replace verb	I existed last weekend in London
FFV	False friend verb	I passed last weekend in London

Hypotheses:

(I) error rate will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors (quantify R and FF errors separately and together)

(II) items subject to error will correlate inversely with native speaker frequencies in the BNC, i.e. the more errors the lower the frequency of the N or V in the BNC

(III) error rates will vary with L1: e.g. genetically distant L1s (Chinese, Japanese, Korean) will exhibit more lexical choice errors than for L1s that are genetically close to English or lexifier languages or closely related to lexifier languages (German, French, Spanish respectively)

(UCLES-RCEAL project, p.3)

Some of his 20 hypotheses were tested against the analysis of the CLC in the same report, but many of them are still awaiting trial. When I first saw these hypotheses, I found some of them have been already investigated in the previous research initiated by my group. Abe and Tono (2005), for instance, investigated the transition of error patterns across proficiency levels in both spoken and written learner corpora. We found that noun lexical choice errors are criterial among other types of errors. Table 11 shows the summary of error types in relation to a spoken learner corpus called the NICT JLE Corpus:

Table 11. Errors that serve as criterial features in the NICT JLE Corpus

Verb agreement errors	[SST 2-3 & 4-6] > [SST 7-9]
Verb tense errors	[SST 2-3 & 4-6] > [SST 7-9]
Noun agreement errors	[SST 4-6 & 7-9] > [SST 2-3]
Noun lexical choice errors	[SST 7-9] > [SST 4-6] > [SST 2-3]

SST stands for the Standard Speaking Test used for this corpus construction. It is a 15-munite oral proficiency interview (OPI) similar to ACTFL-OPI. It has nine levels (1 – lowest and 9 - highest). Table 11 should be read as follows: verb agreement errors are criterial in the sense that it distinguishes SST levels 2-6 from the upper levels (7-9). The errors decrease at the SST level 7 and higher. Noun lexical choice errors are criterial in that it distinguishes the SST levels into three: [2-3], [4-6] and [7-9], where lexical choice errors will increase at higher levels and never go down.

The results apparently contradict with the hypothesis (I) proposed by Hawkins. Interestingly, however, he shows the preliminary results from the CLC later in the same report, which is remarkably similar to my previous findings (see Figure 2):

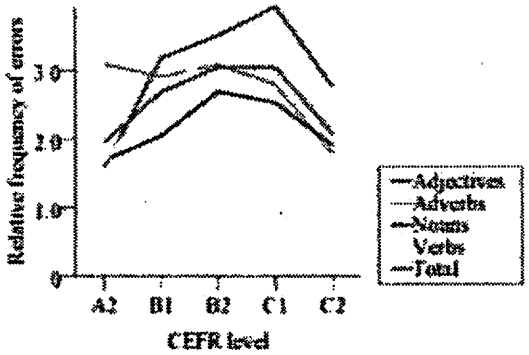

Figure 2. Relative frequency of lexical choice error rates for various parts of speech plotted against CEFR levels (source: the UCREL-RCEAL report)

Lexical choice errors tend to increase from A2 to B2 (even to C1 in the case of adjectives), and then gradually decrease toward C2. That means learners are prone to make lexical choice errors until they reach C1 levels, which is almost equivalent to SST 8-9 levels, thus two findings seem to match quite nicely. In this way, we could test some of Hawkins' hypotheses against previous findings and amend, reformulate hypotheses for further research.

This section has explored the possibility of using the EP programme as a framework for revisiting some of the findings in previous studies in order to identify criterial features for each CEFR level. Although JEFLL or NICT JLE is not classified for the CEFR levels, it seems very promising to integrate the approach taken by the EP programme and our on-going research agenda.

5. CONCLUSION

This paper has summarised research areas in LCR in the past 20 years. The survey has shown that the area of interest is shifting toward more comparative/contrastive research across different proficiency levels with possible learner variables in mind, such as L1 backgrounds or learning environment (e.g. L2 exposure), using multimodal corpora. One good example is the English Profile Programme, where they seek for criterial features for each CEFR level. I have shown that many previous findings can be related to this framework in such a way that the new hypotheses can be tested in light of previous findings or at least the findings using the CLC can be verified. It is worth noting that the CEPC will be 10 million words in size, which is supposed to supplement the CLC for mainly the A1 level data. From my experience, it is a very difficult task to collect A1 level learners' writing or speech to reach 10 million words. The success of the EP programme will depend on the success of data collection for A1 level. Here again, the Japanese learners' data might fill the important gap. Most of the data collected for JEFLL will be classified into A1 to B1 levels, which would be useful to compare against the CEPC or even supplement it. I am keen to see the progress in this area in the coming decade. Let us anticipate that some exciting things will lie ahead for us to further advance our understanding of L2 learners' acquisition processes and how teaching should intervene in that process.

ACKNOWLEDGEMENTS

I would like to thank Mariko Nomura, my former postgraduate student, for her dedicated work on the learner corpus bibliography. Special thanks should go to Dr Tony Green and Dr Masashi Negishi for their comments on my project. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (Grant No. 20242011).

REFERENCES

Abe, M. & Tono, Y. (2005). Variations in L2 spoken and written English: investigating patterns of grammatical errors across proficiency levels. *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, no. 1, ISSN 1747-9398. (Available online at http://www.corpus.bham.ac.uk/pclc/ index.shtml)

Beacco, J.-C., Bouquet, S., & Porquier, R. (Eds.). (2004). *Niveau B2 pour le français (utilisateur/apprenant indépendant): Textes et références* . Paris: Didier.

Beacco, J.-C., de Ferrari, M., & Lhote, G. (Eds.). (2006). *Niveau A1.1 pour le français: Référentiel et certification (DILF) pour les premiers acquis en français* . Paris: Didier.

Briscoe, E., Carroll, J. & Watson, R. (2006). The second release of the RASP System. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney, Australia, available online http://acl.ldc.upenn.edu/P/P06/P06–4020.pdf

Council of Europe (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Filipovic, L (2009). English Profile – Interim report. Internal Cambridge ESOL report, April 2009.

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H. & Wertenschlag, L. (2005). *Profile deutsch*. CD-ROM version 2.0 and accompanying manual. Berlin: Langenscheidt.

Hawkins, J. & Buttery, P. (2008). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme.

Hawkins, J A & Buttery, P (2009). Criterial features in learner corpora: Theory and illustrations. Paper presented at the English Profile Seminar, Cambridge, 5–6 February 2009.

Hendriks, H (2008). Presenting the English Profile Programme: In search of criterial features. *Research Notes* 33, 7–10, Cambridge: Cambridge ESOL.

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching* 39, 167-190.

Parodi, T (2008). L2 morpho-syntax and learner strategies. Paper presented at the Cambridge Institute for Language Research Seminar, Cambridge, UK, 8 December 2008.

Salamoura, A. & Saville, N. (2009). Criterial features of English across the CEFR levels: evidence from the English Profile Programme. *Research Notes* 37: 34-40.

Tono, Y. (2002). *The role of learner corpora in SLA research and foreign language teaching: The multiple comparison approach*. Unpublished Ph.D. Thesis. Lancaster University, UK.

UCREL-RCEAL Funded Research Projects. Retrieved online from www.tinopolisphp.com/ep/images/pdf/ ucles_rceal_projects.pdf (available 12/08/2010).

Williams, C (2007). A preliminary study into the verbal subcategorisation frame: Usage in the CLC. RCEAL, Cambridge University, UK, unpublished manuscript.